MASSACHUSETTS
Department of Elementary and Secondary Education

# 2023 Next-Generation MCAS and MCAS-Alt Technical Report

**September 2024**

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

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, 135 Santilli Highway, Everett, MA 02149 or 781-338-6105.



Massachusetts Department of Elementary and Secondary Education
135 Santilli Highway, Everett, MA 02149
Phone 781-338-3000 TTY: N.E.T. Relay 800-439-2370
www.doe.mass.edu

# Table of Contents

## List of Tables and Figures

# 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 (NCLB) of 2001 and the Every Student Succeeds Act (ESSA) of 2015.

The purpose of this *2023 Next-Generation MCAS and MCAS-Alt Technical Report* is to document the technical quality and characteristics of the 2023 next-generation MCAS English language arts (ELA), mathematics, grades 5 and 8 science and technology/engineering (STE), and high school biology and introductory physics tests, as well as of the 2023 MCAS-Alt, in order to present evidence of the reliability and the validity of test score interpretations, and to describe modifications made to the program in 2023.

Technical reports for previous testing years are available on the DESE website. The previous technical reports, as well as other documents referenced in this report, provide additional background information about the MCAS program, its development, and its 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), standard errors of measurement, reliability, 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 ELA, mathematics, and STE.

The MCAS is designed to do the following:

- Assess all students who are educated with Massachusetts public funds in designated grades, including students with disabilities and English learner (EL) students. (Historically, Massachusetts has had an annual state participation rate over 98% across all grades, subjects, and assessments [see section 3.3.3]).

- Measure student performance in relation to 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.
- Produce scaled scores and achievement levels that indicate students' readiness to engage in academic work at the next grade level, and to inform parents and students if they are not on track based on their test results.
- 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 by releasing test items each year—and by providing item descriptions, standards, and other related information for all test questions, whether released or unreleased.

## 1.2 Organization of This Report

This report provides detailed information regarding test design and development, scoring, and analysis and reporting of 2023 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:

- content descriptions of all tests
- an explanation of test administration
- an explanation of equating and scaling of tests
- statistical and psychometric summaries of the tests
  - item analyses
  - reliability evidence
  - validity evidence

In addition, the appendices contain detailed item-level and summary statistics related to each 2023 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 2023. 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 administered primarily via computer and aligned with the most recent Massachusetts curriculum frameworks. The 2020 MCAS administration was intended to be a continuation of this transition with the introduction of the next-generation high school biology and introductory physics tests. However, due to the COVID-19 pandemic, no new next-generation tests were administered in 2020 or 2021. Thus, the next-generation high school biology and introductory physics tests were first administered in 2022. The final administration of high school legacy chemistry and technology/engineering occurred in spring 2023, completing the plan to phase out legacy tests.

Table 1-1 shows which MCAS tests were administered at each grade level in spring 2023 and whether the tests were next-generation (NG) or legacy (L) assessments.

**Table 1-1. Spring 2023 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 | NG | L/NG* |

** Students must take a high school STE test by the end of grade 10. The legacy chemistry and technology/engineering tests were phased out after the 2023 administration. Additional information about the biology and introductory physics tests is available in Chapter 3.*

## 1.3.1 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 English language arts (ELA) and solving complex problems in mathematics and science and technology/engineering (STE)
- 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
- both computer-based and paper-based test administrations, with computer-based testing as the primary method

In 2023, all students in grades 3–8 and 10 took the next-generation assessments in ELA and mathematics; students in grades 5 and 8 took the next-generation assessments in STE. In addition, the next-generation high school biology and introductory physics tests were administered for the second time. Computer-based administration was required for all content areas at grades 3–8, for grade 10 ELA and mathematics, and for high school biology and introductory physics, but paper-based tests were available as a test accommodation at all grades.

## 1.3.2 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.
- **Spring 2020:** Due to the COVID pandemic, MCAS tests are not administered.
- **Spring 2021**: The fourth administration of next-generation MCAS tests in ELA and mathematics grades 3–8, using one-session tests and some remote administration. The second administration in ELA and mathematics grade 10 using full test forms and in STE grades 5 and 8 using one-session tests.
- **Spring 2022**: The return to full administration of next-generation MCAS tests in ELA and mathematics grades 3–8. The third administration of ELA and mathematics grade 10 and STE grades 5 and 8, plus the first administration of next-generation introductory physics and biology in grades 9 and 10.
- **Spring 2023**: The final administration of legacy chemistry and technology/engineering occurred in grade 10, completing the plan to phase out legacy tests.

# 1.4 Special Issues

Throughout 2023, the Department (DESE) continued to monitor the progressive recovery from COVID-19 and sought to understand more fully the cumulative impact of the pandemic on instruction. DESE's response to COVID-19 is documented in the 2021 MCAS Next-Generation Technical Report.

## 1.4.1 Return to Regular Administration

As recovery from the pandemic continued to progress in 2021–2022, DESE endeavored to return to regular administration of the MCAS, including administration of the full ELA, mathematics, and STE tests to all students. In 2022–2023, regular administration continued with no provisions for remote testing.

## 1.4.2 Change in ELA Essay Scoring

DESE changed the ELA essay scoring for grades 3–8 and removed the dependency between the two trait scores for idea development and conventions. Through spring 2022, students in grades 3–8 could only receive up to 1 point for Conventions if they obtained a 0 score for Idea Development. Beginning in spring 2023, students were able to receive up to 3 points for Conventions regardless of the score for Idea Development. The rule was applied to all 2023 field-test and operational essays in grades 3–8.

.

# Chapter 2. The State Assessment System: MCAS

## 2.1 Guiding Philosophy

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

Since the first administration of the MCAS tests in 1998, 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 25 years. The program continues to evolve. As described in section 1.3, Massachusetts is finalizing 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 2023 next-generation MCAS tests were aligned to the 2017 Massachusetts curriculum frameworks for English language arts (ELA) and mathematics and the 2016 Massachusetts curriculum frameworks for science and technology/engineering (STE).

All learning standards defined in the frameworks are addressed by and incorporated into local curriculum and instruction, whether they are assessed on MCAS or not.

## 2.3 Uses of MCAS Results

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 and high school biology and introductory physics are intended as follows:

1. To be used within the state's framework for district accountability and assistance, in accordance with state priorities and federal requirements.
2. To provide information to support program evaluation at the school and district levels.
3. To provide transparency into student performance through comprehensive reporting on the results of individual students, schools, districts, and the state.
4. To help determine ELA, mathematics, and STE competency (see Appendix A for modified competency determination information) 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.

# 2.4 Validity of MCAS and MCAS-Alt

Validity information for the MCAS and MCAS-Alt assessments is provided throughout this technical report, including 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<br>Appendices G and H | Scoring consistency, interrater agreement, and scoring accuracy |
| | 3.5<br>Appendices I and J | Classical item analyses |
| | 3.7<br>Appendix M | 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<br>Appendices B, C, and T | 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<br>Appendices K and L | Item response theory modeling; dimensionality; scaling; differential item functioning |
| **Consequential validity** | 3.8<br>Appendices L, N, and O | 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 also includes reliability and construct-related characteristics specific to the alternate 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<br>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 |
| **Cognitive processes** | 4.5.5<br>Appendices V and W | Distributions of score frequencies indicate that the tests elicit the expected range of cognitive processes for this population |
| **Precision over the full continuum** | 4.7.3<br>Appendix M | Measurement error calculated over respondent subgroups at each performance level indicate that the tests are sufficiently precise over the full performance continuum |
| **Validity based on other variables** | 4.10.5, 4.1.3, 4.2.1.1, and 4.6 | Resource Guides capturing the judgments of educators and experts about the curricular expectations |
| **Reliability and classical item analyses; subgroup statistics and scoring consistency** | 4.4, 4.7.4, and 4.8<br>Appendices H, N, R, and S | Procedures to ensure consistent scoring; interrater scoring statistics |
| | 4.5<br>Appendices I and J | Classical item statistics |
| | 4.7.1, 4.7.2, and 4.7.3<br>Appendix M | 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 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 ALDs provided in Appendix B 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 in grades 5, 8, and 9/10 are assessed in STE. In 2017, MCAS began the transition to tests 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 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 Test Design and Development

In 2023, the MCAS operational tests were administered to grades 3–8 and 10 in both ELA and mathematics and grades 5, 8, and 9/10 in STE. The tests were administered primarily on a computer with paper accommodations available. Legacy high school STE tests in chemistry and technology/engineering were administered on paper to a small population of grade 10+ students who had previously failed one of these tests. The 2023 school year was the last year these legacy tests were offered.

### 3.2.1 Test Specifications

#### 3.2.1.1 Criterion-Referenced Test

In 2023, the items used on 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 2023 MCAS tests were coded to the 2017 Massachusetts curriculum frameworks in ELA and mathematics and the 2016 Massachusetts curriculum framework for STE. All items were coded to at least one content standard and some were coded to more than one standard. For STE, 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 (SR) items** 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 limited cases in grade 10, 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 to receive 1 or 2 points.
  - **Two-point, technology-enhanced (TE) items** (computer only) use computer-based interactions 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 3 points and are used only on grades 3 and 4 tests. Students are expected to generate approximately one paragraph of text in response to a text-based 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 1 or 2 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 typically instructed to select a certain number of options. There are typically five to six 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 spot, and drag and drop that require the student to choose from a range of options presented. TE Items are machine-scored: students earn 1 point for a correct response and receive 0 points for an incorrect or blank response.
  - **Two-part items** have two parts (Part A and Part B) and are worth two points, each part independently being worth 1 point. They can be multiple-choice, multiple-select, TE, or a combination thereof. Items are machine-scored: students earn 1 point for each correct part and receive 0 points for an incorrect or blank response.

- **Short-answer (SA) items** (computer and paper) are worth 1 or 2 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 models or line plots 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 1-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 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 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 3 or 4 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. Student responses are hand-scored. 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. Students are expected to solve problems and generate one to two sentences in response to a prompt. Student responses are hand-scored. Students earn 4, 3, 2, 1, or 0 score points for these items.

## Science and Technology/Engineering (STE)

- **Selected-response (SR) items** (computer and paper) are worth 1 or 2 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 six 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 spot, 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.
    - **Two-part items** have two parts (Part A and Part B) and are worth 2 points, each part being worth 1 point. They can be multiple-choice, multiple-select, TE, or a

combination thereof. Items are machine-scored: students earn 1 point for each correct part and receive 0 points for an incorrect or blank response.

- **Constructed-response (CR) items** (computer and paper) typically 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, 3, or 4 score points. For 2-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. For 4-point CR items, students may earn 4, 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 to 20 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.

A computer-based test (CBT) common form and a paper-based test (PBT) common form were developed for grades 3–8 and 10 ELA and mathematics and for grades 5, 8, and 9/10 STE. To create the PBT common form, technology-enhanced items on the CBT form were revised and made into paper-based items, typically multiple-choice items. The PBT items tested the same content as the technology-enhanced items on the CBT.

### 3.2.2 ELA Test Specifications

#### 3.2.2.1 Standards

The 2023 MCAS grades 3–8 and 10 ELA tests, including all matrix items, were aligned to 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 (Standard 4)

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 are on a scale outlined in the passage specifications and are less at lower grades. Passage sets consisted of either a single passage or paired passages. Passages were selected from published works; no passages were specifically written for the MCAS tests.

Passages are categorized into one of two types:

1. **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.
2. **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; two sets were literary and two were informational. Across the forms, sets may be single, paired, or tripled selections.

The MCAS ELA test is designed to include a selection of passage sets with a balanced representation, considering gender, race and ethnicity, and socioeconomic status. Another important consideration is that passages be of interest to the tested age group.

Differences among the passages used at each grade level include the length of the passages (typically increases with increasing grade levels) and the degree of complexity (increasing sophistication in

language and concepts as the grade level increases). Test developers use a variety of readability measures 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 standalone 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 10%–15% of the items were technology-enhanced items.

#### Test Design by Grade

*Grades 3 and 4*

The common portion of each test at grades 3 and 4 included three passage sets. Two of the common passage sets included ten to twelve 1- or 2-point selected-response items plus one 7-point text-based essay item or one 3-point constructed-response item. The other common passage set included seven or eight 1- or 2-point selected-response items. 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. Two of the passage sets included eleven 1- or 2-point selected-response items and one 7-point text-based essay item and the other passage set included seven 1- or 2- point selected-response items. 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. Two of the passage sets included eleven 1- or 2-point selected-response items and one 8-point text-based essay item. The other common passage set included seven 1- or 2- point selected-response items. The test contained a total of 50 common points distributed across two testing sessions.

*Grade 10*

The common portion of each test at grade 10 included four passage sets. Three passage sets in the common portion included seven or eight 1- or 2-point selected-response items and two of those three sets included one 8-point text-based essay item. The fourth common passage set included five 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 included two passage sets. In grades 3 and 4, one matrix passage set included eight to ten 1- or 2-point selected-response items, and either two constructed-response items or one essay. The other matrix passage set included seven 1- or 2-point machine-scored items. In grades 5–8, one matrix passage set included seven to ten 1- or 2-point selected-response items and one essay item and the other matrix passage set included seven 1- or 2-point selected-response items.

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, the 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 | 90-120 | 240-270 |

*Common and Matrix Item Distribution*

The grades 3–8 and 10 ELA tests were administered to a large majority of students on the computer with relatively few students who were unable to use a computer taking the paper form as an accommodation. The paper form was derived from Form 1 of the CBT. Both forms had the same number of common and matrix points. Table 3-3 shows the distribution of common and matrix items in each 2023 ELA test, by grade level.

**Table 3-3. Distribution of ELA Common and Matrix Items by Grade and Item Type**

| Grade and Test | | | Items per Form | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Common | | | | Matrix | | | |
| Grade | Test | # of Forms | SR (1 pt.) | SR (2 pt.) | CR | ES | SR (1 pt.) | SR (2 pt.) | CR[1] | ES |
| **3** | ELA | 8 | 26 | 4 | 1 | 1 | 14 | 3 | 0-2 | 0-1 |
| **4** | ELA | 8 | 26 | 4 | 1 | 1 | 14 | 3 | 0-2 | 0-1 |
| **5** | ELA | 8 | 24 | 5 | 0 | 2 | 14 | 3 | 0 | 1 |
| **6** | ELA | 8 | 24 | 5 | 0 | 2 | 14 | 3 | 0 | 1 |
| **7** | ELA | 10 | 24 | 5 | 0 | 2 | 14 | 3 | 0 | 1 |
| **8** | ELA | 8 | 24 | 5 | 0 | 2 | 14 | 3 | 0 | 1 |
| **10** | ELA | 24[2] | 21 | 7 | 0 | 2 | 9 | 3 | 0 | 1 |

*[1] Each grade 3 and grade 4 matrix form contained either two constructed-response items or one essay item.*
*2 For grade 10, Cognia has included the two retest forms in this number. Retest forms do not include matrix items.*

### 3.2.2.5 ELA Blueprints

Table 3-4 shows the target and actual (in parentheses) percentages of common item points by reporting category. Reporting categories are based on the anchor standards in the 2017 Massachusetts curriculum framework for ELA. An in-depth look at the test blueprints is available in Appendix C.

**Table 3-4. Target (and Actual) Distribution of ELA Common Item Points by Reporting Category**

| Reporting Category | Percent of Points at Each Grade (+/-5%) | | | | | | |
|---|---|---|---|---|---|---|---|
| | 3 | 4 | 5 | 6 | 7 | 8 | 10 |
| Language | 25 (27) | 25 (27) | 25 (29) | 25 (20) | 25 (24) | 25 (22) | 25 (21) |
| Reading | 65 (64) | 65 (64) | 55 (54) | 55 (60) | 55 (56) | 55 (58) | 55 (59) |
| Writing | 10 (9) | 10 (9) | 20 (17) | 20 (20) | 20 (20) | 20 (20) | 20 (20) |
| Total | 100 | 100 | 100 | 100 | 100 | 100 | 100 |

### 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 their 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. Table 3-5 shows the range of score points and associated percentages targeted on the operational forms.

**Table 3-5. Targeted Percentage of Score Points by Cognitive Skill Level in English Language Arts**

| Grade | Cognitive Skill Level | Total Points | Percent of Score Points (+/-5%) | Score Points |
|---|---|---|---|---|
| 3–4 | I | 44 | 5% | 0–5 |
| | II | | 70% | 29–33 |
| | III | | 25% | 10–14 |
| 5 | I | 48 | 5% | 0–5 |
| | II | | 60% | 26–31 |
| | III | | 35% | 14–17 |
| 6–8 | I | 50 | 5% | 0–5 |
| | II | | 60% | 27–32 |
| | III | | 35% | 16–20 |
| 10 | I | 51 | 5% | 0–5 |
| | II | | 60% | 28–33 |
| | III | | 35% | 16–21 |

### 3.2.2.7 ELA Reference Materials

The use of authorized bilingual word-to-word dictionaries was allowed during 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 2023 MCAS grades 3–8 and 10 mathematics tests, including all field-test items, were aligned to the learning standards from the *2017 Massachusetts Curriculum Framework for Mathematics*. The 2017 standards are grouped by domains in grades 3–8 and conceptual categories in grade 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
- Conceptual Categories 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 2023 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 (TE) (SA or SR)** | 0 or 1<br>0, 1, or 2 | 3<br>4–8, 10 |
| **Two-part (SA or SR)** | 0, 1, or 2 | 4–8, 10 |
| **Short-answer (SA)** | 0 or 1 | 3–8, 10 |
| **Constructed-response (CR)** | 0, 1, 2, or 3<br>0, 1, 2, 3, or 4 | 3<br>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 and 8*

The common portion of the grades 7 and 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 eight 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 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 2023 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 |

Grades 3–8 and 10 mathematics tests were administered to a large majority of students on the computer with relatively few students taking the paper form as an accommodation. The paper form was derived from Form 1 of the CBT. Both forms had the same number of common and matrix points. Table 3-8 shows the distribution of common and matrix item types by grade level.

**Table 3-8. Distribution of Mathematics Common and Matrix Items by Grade and Item Type**

| Grade | # of Forms | Common | | | | Matrix | |
|---|---|---|---|---|---|---|---|
| | | SR/SA/TE | | CR | | SR/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** | 20 | 32 | 6 | 0 | 4 | 10 | 2 |

## 3.2.3.4 Mathematics Blueprints

Tables 3-9 through 3-12 show the target and actual percentages of common item points by reporting category. Reporting categories are based on the Massachusetts curriculum framework domains.

**Table 3-9. 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 (26) |
| Geometry | 10 (8) | 10 (9) | 10 (9) |
| Measurement and Data | 25 (25) | 20 (20) | 20 (20) |
| **Total** | **100** | **100** | **100** |

**Table 3-10. 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 (20) |
| Expressions and Equations | 30 (30) | 25 (24) |
| Geometry | 15 (15) | 15 (15) |
| Statistics and Probability | 15 (15) | 20 (20) |
| **Total** | **100** | **100** |

**Table 3-11. 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 (37) |
| Functions | 20 (20) |
| Geometry | 30 (30) |
| Statistics and Probability | 10 (13) |
| **Total** | **100** |

**Table 3-12. Target (and Actual) Distribution of Math Common Item Points by Reporting Category, Grade 10**

| Conceptual Category | % 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 apply common, routine procedures or algorithms (that may involve multiple steps) to solve a well-defined problem.

- **Level II (Analysis and Interpretation)**—Level II items require that the student engage in mathematical reasoning beyond simple recall, in a more flexible thought process, and in enhanced organization of thinking skills. These items require a student to make a decision about the approach needed, to represent or model a situation, or to use one or more non-routine procedures to solve a well-defined problem.
- **Level III (Judgment and Synthesis)**—Level III items require that the student perform more abstract reasoning, planning, and evidence-gathering. 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 Level III items at each grade. Table 3-13 shows the target score points and associated score point percentage by cognitive skill level.

**Table 3-13. Targeted Percent of Score Points by Cognitive Skill Level in Mathematics**

| Grade | Cognitive Skill Level | Total Points | Percent of Score Points | Score Points |
|---|---|---|---|---|
| 3 | I | 48 | 25–40% | 12–20 |
| | II | | 55–65% | 26–32 |
| | III | | 6–15% | 3–7 |
| 4–8 | I | 54 | 25–40% | 13–22 |
| | II | | 55–65% | 29–35 |
| | III | | 6–15% | 3–8 |
| 10 | I | 60 | 25–35% | 15–21 |
| | II | | 55–65% | 33–39 |
| | III | | 7–20% | 4–12 |

### 3.2.3.6 Mathematics Reference Materials

Rulers were provided to students in grades 3–8. Handheld 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 handheld 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 and a scientific 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, a TI-84 graphing calculator, and a Desmos graphing 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 (STE) Test Specifications

### 3.2.4.1 STE Standards and Practices

The next-generation STE MCAS tests for grades 5, 8, and 9/10 were aligned to the standards in the 2016 Massachusetts Science and Technology/Engineering Curriculum Framework. In addition, Instructional Guidelines were developed to help clarify some standards and can be found at www.doe.mass.edu/stem/ste/.

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 Pre-K–8 standards are grouped into the following four domains:

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

The grade 9/10 tests were based on the high school biology standards (biology test) and the high school introductory physics (introductory physics test) standards. The 2016 standards are grouped into four domains for biology:

- Molecules to Organisms
- Heredity
- Evolution
- Ecology

The 2016 standards are grouped into three domains for introductory physics:

- Motion, Forces, and Interactions
- Energy
- Waves

The *2016 Massachusetts Science and Technology/Engineering (STE) Curriculum Framework* can be found at https://www.doe.mass.edu/frameworks/scitech/2016-04.pdf.

In addition, the next-generation STE MCAS tests assessed the science and 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

### 3.2.4.2 STE Item Types

The grades 5, 8, and 9/10 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, 8, and 9/10 |
| Multiple-select (SR) | 0 or 1 | 5, 8, and 9/10 |
| Technology-enhanced (SR) | 0 or 1 | 5, 8, and 9/10 |
| Two-point (SR) | 0, 1, or 2 | 5, 8, and 9/10 |
| Constructed-response (CR) | 0, 1, 2, 3, or 4 | 5, 8, and 9/10 |

*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 tests included two common modules, which are groups of items based on a scenario/phenomenon. 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.

The common portion of the grade 9/10 tests included thirty-two 1-point selected-response items, five 2-point selected-response items, two 3-point constructed-response items, and three 4-point constructed-response items. The tests included two common modules, which are groups of items based on a scenario/phenomenon. Each module contained three to five 1-point selected-response items, zero to one 2-point selected-response items, and one 3-point constructed-response item. Each module was made up of a total of 8 points and module items in total made up 16 points of the test; discrete items made up 44 points of the test. The matrix portion included eleven to thirteen 1-point selected-response items, two to three 2-point selected-response items, one 3-point constructed response item, and one 4-point constructed-response item, for a total of 24 points. Each form contained a matrix module (field test) and discrete items. The test contained a total of 60 common points distributed across two testing sessions. Approximately 25% of the items were technology-enhanced items.

Table 3-15 shows the distribution of common and matrix points on the 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**

| 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 | 60–90 | 60–90 | 120–180 | 54 | 10 |
| **8** | 2 | 60–90 | 60–90 | 120–180 | 54 | 10 |
| **9/10** | 2 | 60-90 | 60-90 | 120–180 | 60 | 24 |

The STE tests were administered to a large majority of students on the computer with relatively few students taking the paper form as an accommodation. The paper form was derived from Form 1 of the

CBT. Both forms had the same number of common and matrix points. Table 3-16 shows the distribution of common and matrix item types by grade level.

**Table 3-16. Distribution of STE Common and Matrix Items by Grade and Item Type**

| Grade | # of Forms | Common | | | | | Matrix | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | SR1 (1 pt.) | SR2 (2 pt.) | CR2 (2 pt.) | CR3 (3 pt.) | CR4 (4 pt.) | SR1 (1 pt.) | SR2/CR2 (2 pt.) | CR3 (3 pt.) | CR4 (4 pt.) |
| **5** | 12 | 32 | 3 | 2 | 4 | 0 | 5 | 1 | 1 | 0 |
| **8** | 12 | 32 | 3 | 2 | 4 | 0 | 5 | 1 | 1 | 0 |
| **9/10** | 14/15* | 32–34 | 4-5 | 0 | 2 | 3 | 11–13 | 2–3 | 1 | 1 |

**Introductory physics was 14 and biology was 15, not including retest forms.*

### 3.2.4.4 STE Blueprints

Tables 3-17 through 3-19 show 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 3-17. 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 Science | 25 (26) | 25 (26) |
| Life Science | 25 (26) | 25 (26) |
| Physical Science | 25 (26) | 25 (24) |
| Technology/Engineering | 25 (22) | 25 (24) |
| **Total** | **100** | **100** |

**Table 3-18. Target (and Actual) Distribution of STE Common Item Points by Reporting Category, Grade 9/10 – Biology**

| Domain | % of Points |
|---|---|
| Molecules to Organisms | 35 (35) |
| Heredity | 25 (25) |
| Evolution | 20 (20) |
| Ecology | 20 (20) |
| **Total** | **100** |

**Table 3-19. Target (and Actual) Distribution of STE Common Item Points by Reporting Category, Grade 9/10 – Introductory Physics**

| Domain | % of Points at Each Grade (+/-5%) |
|---|---|
| Motion, Forces, and Interactions | 50 (50) |
| Energy | 30 (30) |
| Waves | 20 (20) |
| **Total** | **100** |

In addition to the content reporting categories, over 50% of the items were coded to an MCAS science and 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 Table 3-20.

**Table 3-20. STE Practices Assessed on MCAS**

| MCAS Practice Category | Science and Engineering Practices |
|---|---|
| **A. Investigations and Questioning** | Asking Questions and Defining Problems<br>Planning and Carrying Out Investigations |
| **B. Mathematics and Data** | Analyzing and Interpreting Data<br>Using Mathematics and Computational Thinking |
| **C. Evidence, Reasoning, and Modeling** | Developing and Using Models<br>Constructing Explanations and Designing Solutions<br>Engaging in Argument from Evidence<br>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 category 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 item that assessed a practice was coded to one of the three practice categories listed in Table 3-20.

### 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 3-21 is an example (grade 5 STE) of the descriptions of the cognitive skills used for the STE test items. Each STE test has its own cognitive skill description, which can be found at https://www.doe.mass.edu/mcas/tdd/cognitive_skills.html.

**Table 3-21. Grade 5 STE Cognitive Skill Descriptions**

| Cognitive Skill | Description |
|---|---|
| **Understanding/ Level 1** | **Students show an understanding of scientific and engineering concepts and skills by:**<br>• Ordering events or quantities for a simple phenomenon, such as ordering the age of rock layers.<br>• Completing a simple model, such as labeling some parts of the water cycle or adding an arrow to complete a model showing the path that light takes for an object to be seen.<br>• Identifying a scientific or engineering process, such as erosion or encoding in a given model or based on a description.<br>• Identifying or describing basic characteristics of an organism, substance, object, event, or environment such as the function of a plant's roots or that a desert receives only small amounts of rain.<br>• Interpreting information to determine a straightforward conclusion, such as where volcanoes occur on a map with plate boundaries.<br>• Determining the materials and tools needed for a basic investigation or to build a prototype, such as a ruler for measuring length or a thermometer for measuring temperature.<br>• Describing the purpose of a design feature for a given design solution, such as plastic being used because it is waterproof, or glass being used because it is see-through. |
| **Applying/ Level 2** | **Students apply their science and engineering knowledge and skills by:**<br>• Interpreting data from a graph or table to draw a conclusion, such as the amount of fresh water available for use by humans and other organisms.<br>• Interpreting a model to draw a conclusion, such as determining the flow of energy in a food web.<br>• Completing an unfamiliar or complex model, such as adding an arrow representing a force on an object to show the object is changing speed.<br>• Setting up a data table for an investigation, given certain criteria.<br>• Providing evidence that supports a claim about a scientific or engineering phenomenon, such as using masses of substances to support a claim that the amount of matter stays the same during a phase change of chemical reaction.<br>• Explaining a scientific or engineering concept when given an unfamiliar context, such as how water changes and moves through several steps of the water cycle for a certain watershed.<br>• Interpreting a diagram of a design solution to draw a straightforward conclusion, including using a ruler to determine if a design solution meets certain criteria.<br>• Determining what scientific question to ask given certain data and criteria.<br>• Determining which variables should be controlled in an investigation and those that may change, such as amount of water, sunlight, or air in a photosynthesis investigation.<br>• Writing a testable question that can be asked for an investigation. (CR items only) |
| **Analyzing/ Evaluating/ Level 3** | **Students analyze or evaluate data and information using their science and engineering knowledge and skills by:**<br>• Analyzing data from multiple sources or from a complex graph or table to draw a conclusion or develop an explanation, such as comparing weather or climate data from two or more locations.<br>• Drawing a conclusion from a complex model or multiple models using scientific or engineering knowledge, such as analyzing two life cycles and drawing conclusions about the two organisms.<br>• Evaluating two models or prototypes and explaining why one is better than the other. (CR items only)<br>• Revising a complex model to make it more accurate, such as determining an error in a food web and describing how to correct the error.<br>• Explaining how a design can be changed to address several criteria and constraints. (CR items only)<br>Note: Some items will reach this level due to students needing to construct an explanation in a constructed response (CR) based on an application of their knowledge. |

## 3.2.4.6 STE Reference Materials

Rulers were provided to students in all grades. Handheld 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 handheld 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 and in grade 9/10. Handheld calculators were given to students taking the paper-based tests.

A reference sheet was provided to students taking the introductory physics test. This sheet contains information, such as formulas, that students may need to answer certain items.

## 3.2.5 Item and Test Development Process

Table 3-22 provides a detailed view of the item and test development process, in chronological order.

**Table 3-22. 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 these passages are reviewed by the contractor's internal diversity, equity, and inclusion committee to minimize bias and sensitivity issue prior to the passages going to DESE. The passages are then presented 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. 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.<br>2. DESE test developers review and edit items prior to presenting the items to ADCs.<br>3. ADCs review items and make recommendations.<br>4. BSC reviews items and makes recommendations.<br>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. |
| **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 and edited by DESE test developers before presentation to the educator committees for review. Passage selection information can be found in section 3.2.2.3. DESE test developers evaluate new items for the following characteristics:

- **Alignment**: Are the items aligned to the standards?
- **Complexity**: Are the items at the appropriate level of complexity?
- **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?
- **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 work collaboratively to revise the proposed item sets in preparation for ADC review. DESE's initial review, subsequent revision, and following work by committees draws on long standing DESE guidance on standards alignment, appropriateness and quality, as well as newly revised guidance on cognitive complexity. This revised guidance defines three "skill levels" of cognitive complexity, articulates what each level means, and provides example items by grade level.

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

The ADCs and BSCs are each composed of approximately 10–12 Massachusetts educators from across the state (see Appendix D for lists of names). There is an ADC for each content area and grade (e.g., ELA grade 3), and one BSC. ADC and BSC members meet several times a year to review new passages and items and to review data from field-test items. Each ADC meeting is co-facilitated by DESE and Cognia's test developers. BSC meetings are facilitated by Cognia staff with one DESE test developer in attendance. ADC and BSC members review items using Pearson's online platform ABBI. Each participant enters their "vote" and recommendations, and the facilitators record the consensus of the committee. All ADC and BSC recommendations remain with each item. DESE 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. Committee members are also asked to consider whether a passage is well-known or comes from a book that is widely taught, since such a passage would likely provide an unfair advantage to those students who are familiar with it. 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/glossed 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. Committee members review and suggest edits to items for the following:

- content accuracy
- grade-level appropriate context and wording
- clearly written stem and question
- clear and accurate graphs, tables, and graphics
- correct answer(s) and scoring notes
- plausible but incorrect distractors
- alignment to correct standard(s)
- alignment to the correct practice (science only)
- alignment to correct cognitive skill
- appropriate use of technology-enhanced items

Members vote to accept, accept with edits (members may include suggested edits), or reject each item. If an item needs significant edits, it will be brought back to the ADC for review again. The meeting facilitators record the consensus/majority opinion of the group, including the suggested edits or reasoning for rejection.

*BSC Passage and Item Review*

After passages and items have been approved by the ADCs, they are also reviewed by a separate 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 is 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 uses a set of guiding questions, which provide the members with a list of considerations in their review of the passages and items for bias and sensitivity. 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 item content. 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 an institution of higher education in either a teaching or research position. Each expert reviewer has been approved by DESE. The External Content Experts recommend either accepting, accepting with edits, or rejecting an item, including their reasoning for edits, or rejecting an item. Expert reviewers' comments remain with each item.

## Status and Editing of Items

DESE test developers review the recommendations of the ADC, BSC, and expert reviewers and determine whether to revise or reject an item based on the suggested edits (in ELA, items are submitted for expert review after the field-test administration; reviewers' comments are considered for the next development cycles and items are rejected for use on the operational test if issues are found). 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; and
- 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 as described 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. We scored approximately 2,000 responses per field-tested constructed-response item or essay. 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 as available prior to committee review by the ADCs and BSCs. An item displaying statistics that indicate it did not perform as expected is closely reviewed and, if 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
- 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 be 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 were 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.
- **"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 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

#### MCAS Accessibility Features and Accommodations

MCAS is accessible to students with and without disabilities through the universal design of the testing platform and test items, the 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 documented disability either in an individualized education program (IEP) or in a 504 plan. English learners may also be eligible for selected special test forms and accommodations. MCAS 2023 operational tests and retests were available in the following special editions for eligible students:

- **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 and the grade 9/10 STE computer-based tests are 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 and grade 9/10 STE tests are intended for Spanish-speaking EL students who have been in the United States 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 E includes all accommodations and special edition test forms and 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 2023 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 in 2023 with staggered start dates, as shown in Table 3-23.

**Table 3-23. Test Administration Schedule—ELA and Mathematics Grades 3–8 & 10, STE 5 & 8, and High School STE**

| 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) |
|---|---|---|---|---|---|
| February Biology and Introductory Physics | December 5–December 16 | January 25 | February 1–2<br>(Last day for makeup testing: February 7) | February 7 | February 8 |
| March Retests (ELA and Mathematics) | January 30–February 7 | March 1 | ELA: March 8–9<br>Math: March 14–15<br>(Last day for makeup testing: March 20) | March 20 | March 21 |
| Grades 3–8 ELA | January 23–February 3 | March 13 | March 27–April 28 | Deadline to complete PCPA: May 30<br>Deadline to update accommodations and mark CBT complete: May 1 | May 2 |
| Grades 3–8 Mathematics | January 23–February 3 | March 13 | April 24–May 26 | May 30 | May 31 |
| Grades 5 & 8 Science and Tech/Eng. | January 23–February 3 | March 13 | April 25–May 26 | May 30 | May 31 |
| Grade 10 ELA | January 30–February 10 | March 14 | Session 1: March 28<br>Session 2: March 29<br>(Last day for makeup testing: April 6) | Deadline to complete PCPA: May 25<br>Deadline to update accommodations and mark CBT complete: April 6 | April 10 |
| Grade 10 Mathematics | January 30–February 10 | March 14 | Session 1: May 16<br>Session 2: May 17<br>(Last day for makeup testing: May 25) | May 25 | May 26 |
| HS Science and Tech/Eng. | April 14–May 2 | May 23 | NG Biology and Introductory Physics:<br>Session 1: June 6<br>Session 2: June 7<br>(Last day for makeup testing: June 14)<br><br>Legacy Chemistry & Tech/Eng.:<br>Session 1: June 6<br>Session 2: June 7<br>(Last day for makeup testing: June 14) | June 14 | June 15 |
| November 2022 Retest | September 19–30 | November 2 | ELA: November 9–10<br>Math: November 15–16<br>(Last day of makeup testing for all tests: November 21) | November 21 | November 22 |

## 3.3.2 Security Requirements

Principals were responsible for ensuring that all test administrators complied with the requirements and instructions contained in the *Principal's Administration Manual*. 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 patrolling the internet, websites, blogs, discussion forums, video archives, social media, document archives, brain dumps, auction sites, and media outlets
- identifying and verifying threats to MCAS test security and notified DESE and Cognia, as required
- working systematically through the steps necessary to have infringing content removed if a threat was verified
- providing summary reporting that included overall and specific threat analysis

DESE receives reports of testing irregularities from schools throughout the various test administrations. For serious irregularities, test results are invalidated, and any cases of educator misconduct are investigated and may be referred to DESE's legal office for a licensure investigation.

DESE performs data analysis on all spring MCAS results and flags schools as outliers if they fall outside certain defined parameters. In cases where the validity of results is called into question, DESE may place student results under review until an investigation can be conducted.

DESE also conducts school monitoring during MCAS testing, going into the field and observing schools' test administrations. Observations are conducted with a checklist and any deviations from the testing protocols are noted for correction.

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 2023 Principal's Administration Manual* (PAM), the *Spring 2023 Test Administrator's Manual for Computer-Based Testing* (CBT TAM), and the *Spring 2023 Test Administrator's Manual for Paper-Based Testing* (PBT TAM).

## 3.3.3 Participation Requirements

In spring 2023, students educated with Massachusetts public funds were required by state and federal laws to participate in MCAS testing. The 1993 Massachusetts Education Reform Act, state law M. G. L. Chapter 69, section 1I, 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 full-time at all publicly funded K–12 schools including
    - district schools
    - charter schools
    - publicly run innovation schools
    - Commonwealth of Massachusetts Virtual Schools
    - 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 at public expense
- students in military families enrolled in public schools
- 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) in grades 3–8 and 10

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 test administration. For a summary of participation rates, see the 2023 MCAS Participation Report on DESE's School and District Profiles website.

### 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, and who were required to participate in the ACCESS for ELLs test
- 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

In 2023, DESE maintained regular administration of the MCAS to all students. No provision was made for remote testing.

It was the principal's responsibility to coordinate the school's 2023 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 2023 testing can be found in Appendix F [note that the information presented in Appendix F is based on all test takers, and counts are broken out by all students, EL students, and students with IEP/Plan 504]; for a list of test accommodations, see Appendix E.)
- 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
- 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, email, and live chat 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 Responses

Scoring of the 2023 MCAS tests was conducted by both Cognia and Pearson.

Scoring 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 as follows:

- 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 in-person benchmarking meetings for field-test items.
- For operational ELA items that needed additional benchmarked responses, 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.

Table 3-24 shows the breakdown of how scoring work was divided between Cognia and Pearson.

**Table 3-24. Breakdown of Scoring Work**

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

For computer-based tests, images for field-test constructed response and essay items were loaded into iScore, Cognia's secure scoring platform. For operational constructed-response and essay items, images were uploaded into the ePEN scoring platform.

For paper-based tests, Cognia scanned each MCAS student answer booklet. Images for field-test constructed-response and essay items were loaded into iScore. Images for operational constructed response and essay items were transferred via FTP site to Pearson for uploading into the ePEN scoring platform. A set of quality-control procedures was enacted for scanning paper test forms. These procedures are provided in Appendix G 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.

## 3.4.2 Benchmarking Meetings

Samples of student responses of 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 and content leads. Each benchmarking meeting was 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 with justifications to a given set of student responses;
- approve anchor and practice sets of responses that are used to train scorers; and
- score additional papers that may be used for qualification sets.

## 3.4.3 Short-Answer Items

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

Prior to machine-scoring of selected-response items, DESE 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.

## 3.4.4 Scoring of Constructed-Response and Essay Items

### 3.4.4.1 Scoring Plan and Staff

The following scoring plan summarizes the approach to the scoring for all grades and content areas:

- All scoring was conducted by applying a virtual/synchronous scoring model maintaining the same quality control measures that were applied in a center-based, regional scoring environment.
- Prior to the start of scoring, scorers attended connectivity sessions to support their readiness for virtual/synchronous scoring and to answer any technology-related questions.
- Scorers evaluated student work on a fixed daily schedule under constant supervision of leadership.
- Training and all interaction between leadership and scorers occurred live via Zoom (Cognia) or Teams (Pearson) and/or via pre-recorded training module or a recording of live training.
- Breakout rooms were used to facilitate scorer training and individualized coaching.
- DESE had remote access to the scoring systems and Zoom/Teams links were provided to observe training sessions and scoring.
- Scorers worked in a non-public setting and were required to be on camera during training, scoring, and any one-on-one or group coaching sessions.
- A post-scoring survey was sent out to all MCAS scoring associates to elicit feedback on their scoring experience. The results were shared with DESE.

The following staff members were involved with scoring the 2023 MCAS responses:

- Cognia Staff
    - The *Scoring Director for Content and Quality* provided guidance, direction, and leadership to MCAS scoring.
    - The *Scoring Operations Managers* provided guidance and oversight of all operational and logistical matters related to scoring.
    - The *Scoring Project Manager* was responsible for the communication and coordination of MCAS scoring between Cognia and Pearson, and between Cognia and DESE.
    - *Scoring Content Specialists* facilitated all benchmarking meetings to ensure consistency of content area benchmarking and field-test scoring across all grade levels. They also handled all aspects for scoring of grade 10 ELA and mathematics, and grades 5, 8, and HS STE. Scoring content specialists prepared training materials for all operational scoring of ELA and mathematics 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 both scorers and Scoring Team Leaders, and for ensuring quality targets for their assigned items.
    - *Scoring Team Leaders* provided support and direction to scorers on quality, accuracy, and timely completion of scoring.
- Pearson Staff
    - The *Scoring Portfolio Manager* was responsible for the coordination, management, and oversight of MCAS scoring for Pearson.
    - The *Scoring Project Manager* oversaw communication and coordination of MCAS scoring between Pearson and Cognia.
    - *Scoring Content Specialists* ensured consistency of content area scoring across all grade levels. 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 both 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 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, potential scorers who 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 were eliminated from the applicant pool. A summary of scorers' backgrounds is provided in Table 3-25.

**Table 3-25. Summary of Scorer and Scoring Leadership Backgrounds (Operational Scoring)**

| Education | Cognia Scorers | | Cognia Leadership | |
|---|---|---|---|---|
| | Number | Percent | Number | Percent |
| Master's degree/doctorate | 236 | 42% | 31 | 37% |
| Bachelor's degree | 316 | 57% | 53 | 63% |
| Associate's degree/more than 48 college credits | 8 | 1% | -- | -- |
| Less than 48 college credits | -- | -- | -- | -- |
| TOTAL | 560 | 100% | 84 | 100% |
| **Teaching Experience** | | | | |
| 11 years or more | - | -- | -- | -- |
| 6–10 years | -- | -- | -- | -- |
| 3–5 years | 323 | 58% | 47 | 56% |
| 1–2 years | -- | -- | -- | -- |
| Less than a year | -- | -- | -- | -- |
| I have no teaching experience | 237 | 42% | 37 | 44% |
| **Scoring Experience** | | | | |
| 3+ years of experience | 88 | 16% | 72 | 86% |
| 1–3 years of experience | 472 | 84% | 12 | 14% |
| No previous experience as scorer/first season | -- | -- | -- | -- |
| **Education** | **Pearson Scorers** | | **Pearson Leadership** | |
| | **Number** | **Percent** | **Number** | **Percent** |
| Master's degree/doctorate | 793 | 30% | 54 | 33% |
| Bachelor's degree | 1,853 | 70% | 110 | 67% |
| Associate's degree/more than 48 college credits | -- | -- | -- | -- |
| Less than 48 college credits | -- | -- | -- | -- |
| TOTAL | 2,646 | 100% | 164 | 100% |
| **Teaching Experience** | | | | |
| 11 years or more | 358 | 19.3% | 21 | 19% |
| 6–10 years | 265 | 14.3% | 9 | 8% |
| 3–5 years | 361 | 19.5% | 18 | 16% |
| 1–2 years | 308 | 16.6% | 15 | 14% |
| Less than a year | 156 | 8.4% | 12 | 11% |
| I have no teaching experience | 450 | 24.3% | 41 | 37% |
| **Scoring Experience** | | | | |
| 3+ years of experience | **578** | **29%** | **39** | **35%** |
| 1–3 years of experience | **1,392** | **71%** | **71** | **65%** |
| No previous experience as scorer/first season | -- | -- | -- | -- |

### 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 were 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 content, grade level, or item scored.

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

- live group training via Zoom/Teams
- recording of live group training
- pre-recorded interactive modules

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 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** were DESE-approved sets consisting of two or three sample responses at each score point. Each response represented a typical response, rather than an unusual or uncommon one; it was solid and had a true score, meaning that this response had a precise score. Anchor sets were used to exemplify each score point.
- **Practice sets** may have included 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 were used to refine the scorers' understanding of how to apply the scoring rules across a wide range of responses.
- **Qualifying sets** consisted of 10 responses that were clear, typical examples of each of the possible score points. Qualifying sets were used to determine whether scorers could 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[1] (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

*1 "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 are either the same or off by only one point.*

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 Hand-Scoring of Constructed Response and Essay Items

Hand-scoring by human scorers was conducted on all field-test items in grades 3–8 and high school and on all operational items in science and mathematics across all grades and for ELA high school. In addition to human scoring, for 10 essay items in ELA in grades 3–8, 10% double-blind scoring (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 at a rate of 10%. All high-school operational scoring received 100% double-blind human scoring. 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.

The 2023 MCAS tests included constructed-response items and essays that were scored by hand. Hand-scored items included the following:

- 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 the following:

- 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).

### 3.4.4.6 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 for high school and the first score was used for grades 3–8 as score-of-record. 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 3-29 and 3-30 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 compared scores and advised or counseled 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 assigned scores that did not match. The examples are based on a 0–4-point high school (HS) item.

**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]** | | | |
| 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] At Grades 3–8: If double-blind scores are adjacent (only 1 point different), the first score is the final score. At Grade HS: If double-blind scores are adjacent, 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. 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.

IEA measures the content and quality of responses by determining the features 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 the following:

- number of grammar errors
- types of grammar errors
- variety of words
- maturity of vocabulary
- variety of sentence types
- coherence of the response
- similarity of the response to other responses and/or source materials

All 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.*

IEA was used operationally for the third consecutive 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 1,300 responses per prompt and then evaluated using approximately 640 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 standard metrics shown in Table 3-28.

**Table 3-27. N Counts by Prompt**

| Grade | Prompt | Training Set Size | Evaluation Set Size |
|---|---|---|---|
| 3 | EL909882556 | 1263 | 631 |
| 4 | EL007459900 | 1305 | 652 |
| 5 | EL030400392 | 1283 | 639 |
| 5 | EL624182427 | 1195 | 598 |
| 6 | EL007051004 | 1251 | 624 |
| 6 | EL807016586 | 1312 | 656 |
| 7 | EL006653237 | 1275 | 637 |
| 7 | EL713375305 | 1186 | 594 |
| 8 | EL007062902 | 1278 | 640 |
| 8 | EL007253494 | 1268 | 636 |

**Table 3-28. Standard Metrics for Evaluating Automated Scoring[2]**

| 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\| |

Ten prompts met the required performance criteria and 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.

Table 3-29 shows a comparison of IEA to human scoring on the validity papers, by exact score point (validity papers are student responses with known scores interspersed among the other student responses; these papers are used to check scoring accuracy). As shown below, IEA scoring accuracy on these validity papers is similar to or slightly higher than the human scoring accuracy at all score points. IEA accuracy tends to be higher than human accuracy at the highest score point, as seen in the Idea Development agreement statistics for grades 3–8.

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

| Grade | UIN | Trait | Validity | N | Exact Agreement | Exact Agreement by Score Point | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | 0 | 1 | 2 | 3 | 4 | 5 |
| 3 | EL909882556 | Idea Development | IEA | 136 | 90% | 79% | 97% | 89% | 83% | 100% | |
| | | | Human | | 84% | 91% | 93% | 79% | 65% | 77% | |
| | | Conventions | IEA | | 87% | 100% | 91% | 74% | 88% | | |
| | | | Human | | 87% | 97% | 92% | 76% | 79% | | |
| 4 | EL007459900 | Idea Development | IEA | 79 | 92% | 100% | 97% | 93% | 71% | 75% | |
| | | | Human | | 89% | 71% | 95% | 83% | 62% | 55% | |
| | | Conventions | IEA | | 98% | 50% | 100% | 100% | 91% | | |
| | | | Human | | 91% | 50% | 96% | 84% | 87% | | |
| 5 | EL030400392 | Idea Development | IEA | 104 | 78% | 84% | 80% | 72% | 67% | 94% | |
| | | | Human | | 78% | 91% | 87% | 63% | 58% | 46% | |
| | | Conventions | IEA | | 83% | 91% | 50% | 85% | 91% | | |
| | | | Human | | 80% | 87% | 80% | 64% | 78% | | |
| | EL624182427 | Idea Development | IEA | 43 | 70% | 100% | 71% | 61% | 75% | 67% | |
| | | | Human | | 76% | 80% | 84% | 72% | 72% | 37% | |
| | | Conventions | IEA | | 77% | 83% | 85% | 61% | 100% | | |
| | | | Human | | 77% | 75% | 81% | 74% | 74% | | |
| 6 | EL007051004 | Idea Development | IEA | 110 | 86% | 86% | 85% | 86% | 92% | 63% | 0% |
| | | | Human | | 75% | 84% | 84% | 72% | 49% | 35% | 0% |
| | | Conventions | IEA | | 91% | 100% | 90% | 90% | 88% | | |
| | | | Human | | 75% | 78% | 76% | 75% | 66% | | |
| | EL807016586 | Idea Development | IEA | 55 | 87% | 100% | 75% | 100% | 50% | 67% | 100% |
| | | | Human | | 89% | 97% | 92% | 74% | 65% | 48% | 71% |
| | | Conventions | IEA | | 95% | 100% | 82% | 80% | 100% | | |
| | | | Human | | 90% | 96% | 87% | 64% | 84% | | |
| 7 | EL006653237 | Idea Development | IEA | 130 | 85% | 95% | 96% | 85% | 93% | 31% | 100% |
| | | | Human | | 85% | 93% | 91% | 88% | 77% | 51% | 49% |
| | | Conventions | IEA | | 91% | 100% | 80% | 91% | 93% | | |
| | | | Human | | 88% | 97% | 85% | 86% | 85% | | |
| | EL713375305 | Idea Development | IEA | 76 | 92% | 100% | 100% | 87% | 92% | 90% | 85% |
| | | | Human | | 85% | 99% | 91% | 79% | 81% | 70% | 72% |
| | | Conventions | IEA | | 96% | 90% | 94% | 93% | 100% | | |
| | | | Human | | 90% | 90% | 89% | 79% | 96% | | |
| 8 | EL007062902 | Idea Development | IEA | 132 | 96% | 100% | 90% | 96% | 95% | 96% | 100% |
| | | | Human | | 78% | 95% | 86% | 75% | 72% | 50% | 67% |
| | | Conventions | IEA | | 97% | 100% | 91% | 96% | 98% | | |
| | | | Human | | 88% | 95% | 85% | 78% | 91% | | |
| | EL007253494 | Idea Development | IEA | 114 | 87% | 100% | 100% | 76% | 89% | 71% | 0% |
| | | | Human | | 74% | 85% | 84% | 73% | 69% | 52% | 0% |
| | | Conventions | IEA | | 85% | 94% | 83% | 85% | 83% | | |
| | | | Human | | 79% | 91% | 80% | 73% | 79% | | |

[2] *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.*

### 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)
- 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 their own score to the response and then compare that 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. Double-blind scoring served as an indicator for agreement of scoring between two scorers. 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 or consistency. As described above, double-blind scoring was the primary process used to monitor the consistency of the hand-scoring of students' constructed responses. Ten percent of responses to 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.

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, and the correlation between the first two sets of scores. The percentages of exact and adjacent scores will approach 100%; sums less than 100 denote that some proportion of third-score resolutions took place. This same information is provided at the item level in Appendix H. 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 (Cohen, 1968):

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

where

$$O = \sum_{i=1}^{n} \sum_{j=1}^{n} \left[1 - \frac{|i - j|}{n - 1}\right] a_{ij}$$

$$E = \sum_{i=1}^{n} \sum_{j=1}^{n} \left[1 - \frac{|i - j|}{n - 1}\right] p_i q_j$$

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 $q_j$ 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 | Items | Number of | | Percentage | | Correlation | LW Kappa |
|---|---|---|---|---|---|---|---|---|
| | | | Score Categories | Included Scores | Exact | Adjacent | | |
| ELA | 3 | 2 | 4 | 12,367 | 73.83 | 25.19 | 0.80 | 0.71 |
| | | 1 | 5 | 6,156 | 71.52 | 26.98 | 0.78 | 0.67 |
| | 4 | 2 | 4 | 12,694 | 72.04 | 27.17 | 0.72 | 0.62 |
| | | 1 | 5 | 6,268 | 71.94 | 27.46 | 0.76 | 0.65 |
| | 5 | 2 | 4 | 12,742 | 66.70 | 31.89 | 0.78 | 0.65 |
| | | 2 | 5 | 12,742 | 64.82 | 32.97 | 0.79 | 0.64 |
| | 6 | 2 | 4 | 13,015 | 68.97 | 30.15 | 0.84 | 0.72 |
| | | 2 | 6 | 13,015 | 64.42 | 33.54 | 0.85 | 0.71 |
| | 7 | 2 | 6 | 12,905 | 70.21 | 27.91 | 0.91 | 0.76 |
| | | 2 | 4 | 12,905 | 72.86 | 26.52 | 0.87 | 0.78 |
| | 8 | 2 | 6 | 13,259 | 64.10 | 33.03 | 0.86 | 0.75 |
| | | 2 | 4 | 13,259 | 74.34 | 24.78 | 0.86 | 0.72 |
| | 10 | 2 | 6 | 136,766 | 66.35 | 32.52 | 0.85 | 0.75 |
| | | 2 | 4 | 136,766 | 78.73 | 20.63 | 0.84 | 0.71 |
| Mathematics | 3 | 5 | 4 | 25,733 | 90.51 | 9.28 | 0.96 | 0.92 |
| | 4 | 4 | 5 | 26,220 | 79.89 | 18.82 | 0.92 | 0.84 |
| | 5 | 5 | 5 | 26,335 | 81.71 | 16.82 | 0.93 | 0.86 |
| | 6 | 4 | 5 | 26,516 | 87.84 | 11.04 | 0.96 | 0.91 |
| | 7 | 4 | 5 | 26,305 | 86.61 | 12.58 | 0.96 | 0.91 |
| | 8 | 4 | 5 | 26,854 | 79.09 | 19.28 | 0.94 | 0.84 |
| | 10 | 8 | 5 | 281,475 | 85.52 | 13.47 | 0.95 | 0.89 |
| STE | 5 | 2 | 3 | 19,359 | 70.34 | 26.93 | 0.70 | 0.62 |
| | | 4 | 4 | 32,995 | 75.04 | 22.51 | 0.83 | 0.69 |
| | 8 | 2 | 3 | 20,012 | 75.97 | 23.23 | 0.78 | 0.67 |
| | | 5 | 4 | 19,562 | 71.39 | 26.19 | 0.83 | 0.72 |
| Biology | HS | 6 | 5 | 187,833 | 76.07 | 21.10 | 0.90 | 0.76 |
| | | 4 | 4 | 63,645 | 74.35 | 24.18 | 0.86 | 0.80 |
| Introductory Physics | HS | 4 | 4 | 30,094 | 69.45 | 27.22 | 0.81 | 0.67 |
| | | 6 | 5 | 45,070 | 69.15 | 27.49 | 0.84 | 0.72 |

[1]*Caution should be used when interpreting the sums of exact and adjacent percentages for ELA items. This is because resolutions are done by response in ELA, and it is entirely possible that only one trait (either idea development or conventions) on a writing response has a non-adjacent score. For instance, if the idea development score for a response were non-adjacent, the response would also receive a third score for conventions, even if it initially received an exact or 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 point, Table 3-31 summarizes the proportion of exact agreement by score points at the test level. Item-level results are also included in Appendix H. 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, with the remaining item response scores provided by human scorers.

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

| Content Area | Grade | Number of Score Categories | Number of Included Scores | Exact | Score Points 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|---|---|---|---|
| ELA | 3 | 4 | 12,367 | 73.83 | 84.70 | 74.60 | 64.40 | 44.80 | | |
| | | 5 | 6,156 | 71.52 | 74.70 | 78.20 | 62.70 | 48.90 | 39.70 | |
| | 4 | 4 | 12,694 | 72.04 | 63.55 | 76.90 | 68.75 | 51.90 | | |
| | | 5 | 6,268 | 71.94 | 77.00 | 75.00 | 68.60 | 42.80 | 19.60 | |
| | 5 | 4 | 12,742 | 66.70 | 71.15 | 68.30 | 60.65 | 66.05 | | |
| | | 5 | 12,742 | 64.82 | 71.20 | 66.60 | 64.70 | 55.15 | 42.90 | |
| | 6 | 4 | 13,015 | 68.97 | 78.50 | 62.75 | 62.00 | 82.20 | | |
| | | 6 | 13,015 | 64.42 | 75.85 | 61.10 | 64.20 | 61.30 | 49.30 | 61.65 |
| | 7 | 4 | 12,905 | 72.86 | 83.65 | 63.15 | 66.30 | 81.45 | | |
| | | 6 | 12,905 | 70.21 | 83.90 | 66.55 | 69.00 | 65.70 | 49.45 | 47.60 |
| | 8 | 4 | 13,259 | 74.34 | 77.25 | 68.10 | 69.95 | 81.25 | | |
| | | 6 | 13,259 | 64.10 | 74.50 | 67.00 | 65.30 | 57.65 | 58.50 | 43.05 |
| | 10 | 4 | 136,766 | 78.73 | 64.65 | 71.25 | 65.10 | 88.40 | | |
| | | 6 | 136,766 | 66.35 | 65.70 | 75.25 | 66.55 | 63.55 | 63.80 | 18.00 |
| Mathematics | 3 | 4 | 25,733 | 90.51 | 94.80 | 86.30 | 87.30 | 91.38 | | |
| | 4 | 5 | 26,220 | 79.89 | 87.10 | 76.00 | 77.68 | 73.15 | 83.48 | |
| | 5 | 5 | 26,335 | 81.71 | 89.20 | 82.65 | 78.55 | 76.72 | 82.28 | |
| | 6 | 5 | 26,516 | 87.84 | 92.58 | 85.48 | 80.30 | 77.53 | 92.30 | |
| | 7 | 5 | 26,305 | 86.61 | 92.58 | 81.55 | 78.38 | 80.50 | 90.73 | |
| | 8 | 5 | 26,854 | 79.09 | 93.20 | 77.30 | 66.10 | 67.33 | 83.18 | |
| | 10 | 5 | 281,475 | 85.52 | 93.10 | 78.08 | 76.20 | 77.60 | 89.58 | |
| Science | 5 | 3 | 19,359 | 70.34 | 76.25 | 64.20 | 72.60 | | | |
| | | 4 | 32,995 | 75.04 | 81.18 | 68.65 | 61.15 | 58.93 | | |
| | 8 | 3 | 20,012 | 75.97 | 79.75 | 66.65 | 76.95 | | | |
| | | 4 | 19,562 | 71.39 | 83.00 | 65.50 | 64.37 | 63.30 | | |
| Biology | HS | 4 | 63,645 | 74.35 | 84.80 | 69.60 | 71.10 | 71.00 | | |
| | | 5 | 187,833 | 76.07 | 92.73 | 70.63 | 60.50 | 57.40 | 67.13 | |
| Introductory Physics | HS | 4 | 30,094 | 69.45 | 79.40 | 60.30 | 59.95 | 64.15 | | |
| | | 5 | 45,070 | 69.15 | 82.10 | 68.03 | 60.13 | 55.07 | 65.30 | |

As described in section 3.4.4.8, validity responses were used to monitor the scoring accuracy. Table 3-32 provides a summary of these “validity” statistics. These statistics denote accuracy in scoring; they provide an average of the human and IEA agreement with the validity responses (e.g., agreement with the true scores for each essay). Item-level results are also included in Appendix H.

**Table 3-32. Summary of Validity Statistics[1]**

| Subject | Grade | Number of Score Categories[2] | Number of Validity Responses[3] | Exact Agreement | Agreement by Score Point | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | 0 | 1 | 2 | 3 | 4 | 5 |
| ELA | 3 | 4 (SR) | 3,584 | 82.5% | 93.3% | 83.1% | 78.5% | 31.3% | | |
| | | 4 (Conv) | 3,282 | 86.6% | 97.1% | 91.7% | 75.5% | 78.8% | | |
| | | 5 (ID) | 3,282 | 84.5% | 91.1% | 92.6% | 79.3% | 64.7% | 77.0% | |
| | 4 | 4 (SR) | 3,765 | 83.9% | 84.9% | 83.9% | 81.9% | 84.7% | | |
| | | 4 (Conv) | 3,214 | 90.5% | 50.3% | 95.6% | 83.8% | 86.6% | | |
| | | 5 (ID) | 3,214 | 88.7% | 71.2% | 95.4% | 83.4% | 62.2% | 55.1% | |
| | 5 | 4 (Conv) | 6,464 | 78.7% | 84.9% | 80.7% | 71.3% | 76.3% | | |
| | | 5 (ID) | 6,464 | 77.2% | 90.1% | 84.7% | 69.5% | 63.3% | 42.9% | |
| | 6 | 4 (Conv) | 6,712 | 82.5% | 89.0% | 81.7% | 74.2% | 75.9% | | |
| | | 6 (ID) | 6,712 | 82.1% | 91.6% | 88.2% | 72.4% | 55.3% | 41.9% | 70.9% |
| | 7 | 4 (Conv) | 6,598 | 88.7% | 93.8% | 86.9% | 84.0% | 91.6% | | |
| | | 6 (ID) | 6,598 | 85.1% | 96.2% | 91.1% | 85.6% | 78.4% | 63.4% | 69.1% |
| | 8 | 4 (Conv) | 7,023 | 83.5% | 93.6% | 82.9% | 74.5% | 84.9% | | |
| | | 6 (ID) | 7,023 | 76.1% | 90.3% | 84.9% | 73.7% | 70.0% | 51.2% | 66.5% |
| Mathematics | 3 | 4 | 6,901 | 94.3% | 95.6% | 93.2% | 92.5% | 96.6% | | |
| | 4 | 5 | 7,208 | 91.4% | 91.6% | 91.3% | 91.1% | 87.5% | 94.0% | |
| | 5 | 5 | 7,100 | 94.8% | 96.1% | 94.0% | 94.0% | 95.5% | 94.5% | |
| | 6 | 5 | 7,107 | 94.9% | 97.0% | 92.8% | 95.2% | 93.2% | 97.1% | |
| | 7 | 5 | 7,150 | 94.1% | 98.2% | 93.2% | 90.8% | 91.9% | 95.9% | |
| | 8 | 5 | 7,327 | 93.2% | 98.4% | 92.3% | 91.7% | 88.6% | 93.7% | |

*[1]Includes all operational and equating items for ELA and mathematics.*
*[2]SR= Short response; Conv= Conventions; ID=Idea Development*
*[3] This column displays the number of validity reads (how many times all the responses were scored against validity papers) that occurred, not the number of validity papers used.*

# 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 2023. 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 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 (approximately 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 2023 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, in general, 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 for the CBT items administered in school is presented in Table 3-33. 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. Note that the information presented in this section and associated appendices are based only on first-time test takers who are not first-year EL students.

**Table 3-33. 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 | 32 | 0.61 | 0.13 | 0.52 | 0.10 |
| | | SR | 26 | 0.63 | 0.11 | 0.49 | 0.08 |
| | | CR | 5 | 0.58 | 0.16 | 0.62 | 0.07 |
| | | ES | 1 | 0.31 | | 0.73 | |
| | 4 | All | 32 | 0.62 | 0.13 | 0.47 | 0.10 |
| | | SR | 26 | 0.64 | 0.13 | 0.44 | 0.08 |
| | | CR | 5 | 0.56 | 0.09 | 0.56 | 0.09 |
| | | ES | 1 | 0.35 | | 0.69 | |
| | 5 | All | 31 | 0.65 | 0.15 | 0.48 | 0.13 |
| | | SR | 24 | 0.67 | 0.14 | 0.43 | 0.09 |
| | | CR | 5 | 0.69 | 0.10 | 0.61 | 0.06 |
| | | ES | 2 | 0.39 | 0.09 | 0.74 | 0.01 |
| | 6 | All | 31 | 0.61 | 0.12 | 0.46 | 0.12 |
| | | SR | 24 | 0.63 | 0.11 | 0.43 | 0.07 |
| | | CR | 5 | 0.58 | 0.11 | 0.52 | 0.12 |
| | | ES | 2 | 0.39 | 0.01 | 0.80 | 0.00 |
| | 7 | All | 31 | 0.60 | 0.12 | 0.47 | 0.11 |
| | | SR | 24 | 0.63 | 0.11 | 0.43 | 0.05 |
| | | CR | 5 | 0.55 | 0.12 | 0.53 | 0.07 |
| | | ES | 2 | 0.41 | 0.04 | 0.79 | 0.01 |
| | 8 | All | 31 | 0.65 | 0.10 | 0.46 | 0.13 |
| | | SR | 24 | 0.67 | 0.10 | 0.43 | 0.10 |
| | | CR | 5 | 0.65 | 0.08 | 0.48 | 0.09 |
| | | ES | 2 | 0.48 | 0.04 | 0.81 | 0.02 |
| | 10 | All | 30 | 0.71 | 0.10 | 0.50 | 0.12 |
| | | SR | 21 | 0.74 | 0.09 | 0.45 | 0.08 |
| | | CR | 7 | 0.64 | 0.09 | 0.55 | 0.04 |
| | | ES | 2 | 0.61 | 0.02 | 0.81 | 0.01 |
| Mathematics | 3 | All | 40 | 0.58 | 0.12 | 0.56 | 0.11 |
| | | SR | 16 | 0.57 | 0.11 | 0.49 | 0.09 |
| | | CR | 24 | 0.58 | 0.13 | 0.60 | 0.11 |
| | 4 | All | 40 | 0.59 | 0.12 | 0.55 | 0.11 |
| | | SR | 20 | 0.61 | 0.13 | 0.48 | 0.09 |
| | | CR | 20 | 0.57 | 0.11 | 0.62 | 0.08 |
| | 5 | All | 40 | 0.53 | 0.12 | 0.52 | 0.12 |
| | | SR | 18 | 0.52 | 0.13 | 0.45 | 0.09 |
| | | CR | 22 | 0.54 | 0.10 | 0.58 | 0.11 |
| | 6 | All | 40 | 0.52 | 0.12 | 0.54 | 0.14 |
| | | SR | 16 | 0.55 | 0.13 | 0.44 | 0.11 |
| | | CR | 24 | 0.50 | 0.11 | 0.60 | 0.12 |
| | 7 | All | 40 | 0.44 | 0.11 | 0.55 | 0.14 |
| | | SR | 17 | 0.46 | 0.10 | 0.44 | 0.11 |
| | | CR | 23 | 0.43 | 0.11 | 0.63 | 0.11 |
| | 8 | All | 40 | 0.51 | 0.12 | 0.54 | 0.14 |
| | | SR | 16 | 0.56 | 0.11 | 0.45 | 0.09 |
| | | CR | 24 | 0.48 | 0.12 | 0.59 | 0.14 |
| | 10 | All | 42 | 0.53 | 0.13 | 0.56 | 0.13 |
| | | SR | 22 | 0.60 | 0.11 | 0.51 | 0.09 |
| | | CR | 20 | 0.45 | 0.11 | 0.62 | 0.14 |
| STE | 5 | All | 41 | 0.58 | 0.16 | 0.50 | 0.08 |
| | | SR | 20 | 0.63 | 0.15 | 0.46 | 0.08 |
| | | CR | 21 | 0.54 | 0.16 | 0.53 | 0.08 |
| | 8 | All | 41 | 0.51 | 0.12 | 0.50 | 0.12 |
| | | SR | 26 | 0.53 | 0.13 | 0.46 | 0.09 |
| | | CR | 15 | 0.47 | 0.10 | 0.56 | 0.14 |
| Biology | HS | All | 42 | 0.56 | 0.13 | 0.51 | 0.12 |
| | | SR | 27 | 0.59 | 0.11 | 0.47 | 0.08 |
| | | CR | 15 | 0.51 | 0.15 | 0.59 | 0.15 |
| Introductory Physics | HS | All | 42 | 0.58 | 0.13 | 0.50 | 0.14 |
| | | SR | 21 | 0.60 | 0.12 | 0.44 | 0.10 |
| | | CR | 21 | 0.56 | 0.13 | 0.56 | 0.15 |

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 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 I. 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 none with difficulty below 0.20 and one item with discrimination below 0.20. Often, items with relatively lower discrimination are kept in the operational forms to ensure content blueprint coverage. Item-level score point distributions are provided for constructed-response items in Appendix J; 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 2023 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 2023 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. Note that the information presented in this section and the associated appendix is based only on first-time test takers who are not first-year EL 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 2023 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 2023 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 EL/FEL compared with EL/FEL[3]
- not Low Income compared with Low Income
- white compared with African American/Black
- white compared with Hispanic or Latino
- without disabilities compared to with disabilities

After the 2023 spring administration, DIF analyses were conducted again as a post-hoc quality check based on the operational data. The tables in Appendix K present the number of items classified as either "low" or "high" DIF, in total and by group favored. Following Dorans and Holland's recommendation, items with DIF indices between -0.10 and -0.05 and between 0.05 and 0.10 were categorized as "low" DIF, and values outside the -0.10 to 0.10 range were categorized as "high" DIF. Very few items exhibited high DIF in the operational data, suggesting that the bias and sensitivity review after the field-testing effectively ruled out items displaying large DIF for the MCAS 2023 spring tests.

## 3.5.3 Dimensionality Analysis

Because tests are constructed with multiple content area subcategories and their associated knowledge and skills, the potential exists for the invocation of multiple dimensions beyond the common primary dimension. Generally, the subcategories are highly correlated with each other; therefore, a primary dimension typically explains the majority of variance in test scores. The presence of one dominant primary dimension is the primary psychometric assumption to support the use of the unidimensional item response theory (IRT) models that are used for calibrating and scaling the 2023 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 2023 administrations. A total of 18 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

[3] *EL=English learner / FEL=former English learner*

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 non-clustered 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 2023. 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 to estimate the effect size of the multidimensionality. Note that the information presented in this section is based only on first-time test takers who are not first-year EL students and who took non-accommodated online forms.

### 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 most of the data sets (over 24,000 in 17 out of 18 tests), DIMTEST would be expected to be sensitive to even quite small violations of unidimensionality.

### 3.5.3.2 DETECT Analyses

Next, DETECT was used to estimate the effect size for the violations of local independence for the 2017 to 2023 tests. Table 3-34 displays the multidimensionality effect-size estimates from DETECT.

**Table 3-34. Multidimensionality Effect Sizes by Grade and Content Area**

| Content Area | Grade | Multidimensionality Effect Size | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 2017 | 2018 | 2019 | 2021* | | 2022 | 2023 |
| | | | | | Session 1 | Session 2 | | |
| ELA | 3 | 0.25 | 0.17 | 0.27 | 0.24 | 0.27 | 0.20 | 0.16 |
| | 4 | 0.30 | 0.35 | 0.29 | 0.34 | 0.25 | 0.25 | 0.20 |
| | 5 | 0.35 | 0.28 | 0.34 | 0.44 | 0.26 | 0.23 | 0.23 |
| | 6 | 0.38 | 0.26 | 0.42 | 0.44 | 0.37 | 0.33 | 0.30 |
| | 7 | 0.34 | 0.34 | 0.49 | 0.51 | 0.26 | 0.35 | 0.41 |
| | 8 | 0.38 | 0.35 | 0.47 | 0.32 | 0.20 | 0.31 | 0.34 |
| | 10 | 0.20 | 0.24 | 0.26 | 0.34 | -- | 0.28 | 0.32 |
| | **Average** | **0.33** | **0.29** | **0.36** | **0.38** | **0.27** | **0.28** | **0.28** |
| Mathematics | 3 | 0.20 | 0.17 | 0.20 | 0.23 | 0.18 | 0.21 | 0.19 |
| | 4 | 0.19 | 0.22 | 0.10 | 0.12 | 0.20 | 0.16 | 0.11 |
| | 5 | 0.19 | 0.15 | 0.15 | 0.26 | 0.22 | 0.18 | 0.10 |
| | 6 | 0.21 | 0.13 | 0.21 | 0.21 | 0.21 | 0.14 | 0.17 |
| | 7 | 0.13 | 0.14 | 0.15 | 0.34 | 0.14 | 0.16 | 0.15 |
| | 8 | 0.11 | 0.15 | 0.13 | 0.19 | 0.25 | 0.19 | 0.16 |
| | 10 | 0.12 | 0.09 | 0.09 | 0.11 | -- | 0.17 | 0.13 |
| | **Average** | **0.17** | **0.16** | **0.15** | **0.21** | **0.20** | **0.17** | **0.14** |
| STE | 5 | 0.08 | 0.11 | 0.08 | 0.22 | 0.18 | 0.09 | 0.13 |
| | 8 | 0.08 | 0.13 | 0.08 | 0.19 | 0.18 | 0.13 | 0.14 |
| **Biology**** | HS | -- | -- | -- | -- | -- | 0.10 | 0.10 |
| **Introductory Physics**** | HS | -- | -- | -- | -- | -- | 0.10 | 0.10 |
| | **Average** | **0.08** | **0.12** | **0.08** | **0.21** | **0.18** | **0.10** | **0.12** |

** In 2021, two sessions in each test were randomly spiraled among students, and each session was analyzed as a separate form except grade 10 ELA and math. Because each session had a different content blueprint than the entire test, caution should be taken when comparing the 2021 DETECT effect size results to any other year's results.*
*** Because 2022 was the first year of the next-generation tests for high school biology and introductory physics, no dimensionality analysis was conducted for these tests in prior years.*

The DETECT values indicate very weak (DETECT < 0.2) multidimensionality for all the 2023 mathematics and STE test forms, which are consistent with previous years' results. The 2023 high school biology and introductory physics tests also show very weak multidimensionality (DETECT < 0.2). The 2023 ELA tests mostly show very weak (DETECT < 0.2) or weak multidimensionality (0.2 < DETECT < 0.4; with larger DETECT effect size indicating stronger multidimensionality), though the ELA Grade 7 shows moderate multidimensionality (DETECT = 0.41).

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 with stronger multidimensionality, every form had one set of clusters dominated by selected-response items and another set of clusters dominated by essay items. These results give solid evidence that the essays form a distinct cluster from the selected-response items.

On the mathematics and STE test forms, there was less clear evidence of consistent separation of selected-response and constructed-response items. This lack of evidence is consistent with the weaker multidimensionality exhibited by those subjects historically.

In summary, for the 2023 dimensionality analyses, the violations of local independence, as evidenced by the DETECT effect sizes, were either weak or very weak in mathematics and STE test forms and were weak 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 and STE.

# 3.6 MCAS IRT Linking and Scaling

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

- 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 function curves (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
- 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 2023 next-generation MCAS tests on the same scale as the previous year. An equating report (Appendix L), which provided complete documentation of the quality-control procedures and results, was reviewed by DESE and approved prior to production of the *Spring 2023 MCAS Tests Parent/Guardian Reports*. Note that the information presented in this section and associated appendices are based only on first-time test takers who are not first-year EL students and who took non-accommodated online forms.

## 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 ($\theta$), and the probability [$P(\theta)$] of getting a dichotomous item correct or of getting a particular score on a polytomous item (Hambleton, Swaminathan, & Rogers, 1991; Hambleton & Swaminathan, 1985). In IRT, it is assumed that all items are independent measures of the same construct (i.e., of the same $\theta$). Another way to think of $\theta$ is as a mathematical representation of the latent trait of interest. Several common IRT models are used to specify the relationship between $\theta$ and $P(\theta)$ (Hambleton & van der Linden, 1997; Hambleton & Swaminathan, 1985). The process of determining the mathematical relationship between $\theta$ and $P(\theta)$ is called item calibration. After items are calibrated, they are defined by a set of parameters that specify a nonlinear, monotonically increasing relationship between $\theta$ and $P(\theta)$. Once the item parameters are known, an estimate of $\theta$ for each student can be calculated. This estimate, $\hat{\theta}$, is considered 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 facilitates equating to an IRT-based item pool (Kolen & Brennan, 2014).

For the 2023 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 $\theta$. 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,
$\theta$ is the student ability,
$\alpha$ 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 $\theta$, 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,
$\theta$ 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 L give the IRT item parameters and associated standard errors of all operational scoring items on the 2023 MCAS tests. Appendix L contains graphs of the TCCs and TIFs, which are defined below.

TCCs display the expected (average) raw score associated with each $\theta_j$ value typically 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 $\theta_j$ 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, $\theta_j$ runs from -4 to 4), and
$E(X|\theta_j)$ is the expected raw score for a student of ability $\theta_j$.

The expected raw score monotonically increases with $\theta_j$, 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 $\theta_j$. 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 $\theta_j$ is approximately equal to the inverse of the square root of the statistical information at $\theta_j$ (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 $\theta$ 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 most 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 3-35. The calibration went smoothly and converged in all subjects/grades.

**Table 3-35. Number of Cycles Required for Convergence**

| Content Area | Grade | Initial Cycles | FCIP Cycles |
|---|---|---|---|
| ELA | 3 | 30 | 8 |
| | 4 | 27 | 7 |
| | 5 | 43 | 8 |
| | 6 | 26 | 12 |
| | 7 | 21 | 9 |
| | 8 | 119 | 12 |
| | 10 | 43 | 11 |
| Mathematics | 3 | 66 | -- |
| | 4 | 71 | -- |
| | 5 | 50 | -- |
| | 6 | 43 | -- |
| | 7 | 80 | -- |
| | 8 | 33 | -- |
| | 10 | 55 | -- |
| STE | 5 | 50 | -- |
| | 8 | 47 | -- |
| Biology | HS | 23 | -- |
| Introductory Physics | HS | 66 | -- |

## 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 2023 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.

For mathematics and STE, the item parameter estimates for 2023 test forms were placed on the reference scale by using the Stocking-Lord method (SL; Stocking & Lord, 1983). However, a two-step equating approach was taken for ELA because of the finding in the 2022 dimensionality analyses that the dimensionality structure of the test displayed evidence of having changed from 2019. More detail on the 2022 dimensionality analyses can be found in the *2022 Next-Generation MCAS and MCAS-Alt Technical Report*. The 2023 equating followed the same procedure as established in 2022. The first step involved applying the SL method for all items except the essay items; thus, isolating any dimensionality variability in the essay items from the estimation of the equating relationship across years. Then, the essay items were brought onto the scale established in the first step by applying the fixed common item parameters (FCIP2; Kim, 2006) method. The FCIP2 method is based on the IRT principle of item parameter invariance. According to this principle, the equating items for both tests should have the same item parameters. After the item parameters for the non-essay items were put on the reference scale (the first step), the FCIP2 method was employed to place the essay items onto the operational scale (the second step). This method is performed by fixing the parameters of the "equating" items (in this case, all non-essay items) to their previously obtained on-scale values and then calibrating using flexMIRT to place the remaining items (in this case, the essay items) on scale.

Prior to implementing the SL method, two 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]"[4]) 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.

During the implementation of the SL method, a third evaluation of the equating items was conducted to check for parameter drift, as follows.

- 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 $\beta$, 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 3-36. Detailed results from each drift analysis, along with Delta and *b*-plots are presented in Appendix L.

[4] *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*

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 construct-irrelevant factors follows:

- 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 was removed from the equating set if a construct-irrelevant reason was identified in the content review. If a construct-relevant reason was identified or the content review does not find any reason, an item was kept as an equating item.

**Table 3-36. Year-to-year Equating Items Watch List**

| Content Area | Grade | Item ID | Statistical Reason | Content Reason | Action |
|---|---|---|---|---|---|
| ELA | 3 | IA00286 | beta | None identified | Retained |
| | 4 | IA00289 | beta | None identified | Retained |
| | 8 | IA00063 | beta | None identified | Retained |
| Mathematics | 4 | IA00961 | beta | None identified | Retained |
| | | IA01093 | beta | None identified | Retained |
| | 5 | IA00936 | beta | None identified | Retained |
| | | IA00865 | beta | None identified | Retained |
| | 8 | IA02495 | beta | None identified | Retained |
| | | IA05070 | beta | None identified | Retained |
| STE | 5 | IA05657 | beta | None identified | Retained |
| | | IA05702 | beta | None identified | Retained |
| | 8 | IA05245 | beta | Item used a term from the old standards | Removed from equating |
| | | IA05690 | beta | None identified | Retained |
| Biology | HS | IA10684 | beta | None identified | Retained |
| | | IA10989 | beta | None identified | Retained |
| | | IA11033 | beta | None identified | Retained |
| | | IA11054 | beta | None identified | Retained |
| Introductory Physics | HS | IA10704 | beta | None identified | Retained |
| | | IA10936 | beta | None identified | Retained |

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 2023 next-generation computer-based administration onto the target scale. The transformed item parameters were then used to build the raw score to scaled score look-up tables for the 2023 tests. The SL constants are presented in Table 3-37.

**Table 3-37. Stocking and Lord Constants**

| Content Area | Grade | Slope | Intercept |
|---|---|---|---|
| ELA | 3 | 1.14 | -0.21 |
| | 4 | 1.07 | -0.26 |
| | 5 | 1.14 | -0.21 |
| | 6 | 1.44 | -0.36 |
| | 7 | 1.22 | -0.30 |
| | 8 | 1.41 | -0.21 |
| | 10 | 1.16 | -0.17 |
| Mathematics | 3 | 1.08 | -0.03 |
| | 4 | 1.03 | 0.10 |
| | 5 | 1.01 | -0.02 |
| | 6 | 1.03 | -0.11 |
| | 7 | 1.11 | -0.14 |
| | 8 | 1.10 | -0.19 |
| | 10 | 0.97 | -0.18 |
| STE | 5 | 1.11 | -0.17 |
| | 8 | 1.06 | -0.17 |
| Biology | 10 | 0.84 | 0.26 |
| Introductory Physics | 10 | 0.94 | 0.29 |

## 3.6.4 Achievement Standards

Cutpoints for the next-generation MCAS tests were set via standard setting in 2017 for grades 3–8 ELA and mathematics tests, in 2019 for grade 10 ELA and mathematics tests and grades 5 and 8 STE tests, and in 2022 for biology and introductory physics (see the *2022 Next-Generation MCAS and MCAS-Alt Technical Report* for the 2022 standard-setting report, the *2019 Next-Generation MCAS and MCAS-Alt Technical Report* 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 3-38. Also shown in Table 3-38 are the cut scores on the reporting score scale. The operational $\theta$-metric and reporting score scale cut scores will remain fixed throughout the assessment program unless standards are reset.

**Table 3-38. 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 |
| 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 |
| Biology | HS | -0.850 | 0.210 | 1.300 | 440 | 470 | 500 | 530 | 560 |
| Introductory Physics | HS | -1.010 | 0.120 | 1.260 | 440 | 470 | 500 | 530 | 560 |

## 3.6.5 Reported Scale Scores

Because the $\theta$ scale used in IRT calibrations is not understood by most stakeholders, reporting scales were developed for the MCAS tests. The reporting scales are linear transformations of the underlying $\theta$ scale. As the three $\theta$ cutpoints from the standard setting have equal intervals, one single linear transformation was sufficient to transform the $\theta$ 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 obtained) on the 2023 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 2023 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, equating typically results in different raw cut scores across different administrations. The raw cut score between *Partially Meeting Expectations* and *Meeting Expectations* could be, for example, 35 in grade 3 mathematics in 2022 but 34 in 2023, 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 $\theta$. Since the $\theta$ 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 $\theta$ metric and their equivalent values on the scaled score metric. Students' ability estimates are obtained by mapping their raw scores through the TCC. Scaled scores are calculated using the following linear equation, representing the standard deviation of scaled scores on the first administration of the test:

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

where
*m* is the slope and
*b* is the intercept.

A separate linear transformation is used for each grade and content area combination. Table 3-39 shows the slope and intercept terms used to calculate the scaled scores for each grade and content area. Note that the values in Table 3-39 will not change unless the standards are reset.

Appendix L contains the raw-score-to-scaled-score look-up table for each test. The tables show the scaled score equivalent of each raw score for the 2023 next-generation MCAS tests. Additionally, Appendix L contains scaled score distribution graphs for each grade and content area for each testing form.

**Table 3-39. 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 |
| Biology | HS | 27.907 | 493.721 |
| Introductory Physics | HS | 26.432 | 496.696 |

# 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 that are 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 or height, 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 their true ability on a test.

There are several ways to estimate an assessment's reliability. The approach that was implemented to assess the reliability of the 2023 next-generation MCAS tests is the $\alpha$ coefficient of Cronbach (1951). This approach is most easily understood as an extension of a related procedure, split-half reliability. In the split-half approach, a test form 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 $\alpha$ 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 $\alpha$ is referred to as a coefficient of internal consistency. The term "internal" indicates that the index is measured internally to each test of interest, using data that come only from the test itself (Anastasi & Urbina, 1997). The formula for Cronbach's $\alpha$ is given 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 total number of items,
$\sigma^2_{(Y_i)}$ represents individual item variance, and
$\sigma^2_x$ represents the total test variance.

Note that the information presented in this section and associated appendices are based only on first-time test takers who are not first-year EL students and who took non-accommodated online forms.

## 3.7.1 Reliability and Standard Errors of Measurement

Table 3-40 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 0.88 to 0.94, which are generally in acceptable ranges.

**Table 3-40. 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 | 60,542 | 44 | 25.01 | 9.51 | 0.91 | 2.81 |
| | 4 | 61,836 | 44 | 25.59 | 8.43 | 0.88 | 2.86 |
| | 5 | 62,316 | 48 | 28.76 | 9.43 | 0.90 | 2.96 |
| | 6 | 63,574 | 50 | 27.63 | 9.90 | 0.90 | 3.17 |
| | 7 | 63,711 | 50 | 27.50 | 10.71 | 0.90 | 3.44 |
| | 8 | 65,553 | 50 | 30.67 | 9.96 | 0.90 | 3.22 |
| | 10 | 68,104 | 51 | 34.77 | 10.17 | 0.91 | 3.13 |
| Mathematics | 3 | 51,707 | 48 | 29.38 | 11.52 | 0.93 | 2.99 |
| | 4 | 52,554 | 54 | 33.18 | 11.79 | 0.93 | 3.22 |
| | 5 | 54,159 | 54 | 30.52 | 11.90 | 0.92 | 3.36 |
| | 6 | 56,389 | 54 | 28.84 | 13.12 | 0.93 | 3.56 |
| | 7 | 57,234 | 54 | 25.42 | 13.58 | 0.93 | 3.49 |
| | 8 | 59,572 | 54 | 28.94 | 13.10 | 0.93 | 3.47 |
| | 10 | 63,574 | 60 | 31.75 | 14.78 | 0.94 | 3.61 |
| STE | 5 | 49,237 | 54 | 31.91 | 10.22 | 0.90 | 3.18 |
| | 8 | 54,215 | 54 | 29.09 | 10.92 | 0.90 | 3.42 |
| Biology | HS | 49,403 | 60 | 33.94 | 13.08 | 0.92 | 3.61 |
| Introductory Physics | HS | 12,822 | 60 | 33.89 | 12.48 | 0.92 | 3.49 |

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 2023 next-generation MCAS tests. Appendix M presents reliability coefficients for various subgroups of interest. Cronbach's $\alpha$ coefficients were calculated using the formula defined above based only on the members of the subgroup in question in the computations; values are calculated only for subgroups with 10 or more students. The reliability coefficients for subgroups range from 0.78 to 0.95 across the tests, with a median of 0.91 and a standard deviation of 0.02, indicating that reliabilities are generally within a reasonable range.

For several reasons, the subgroup reliability results should be interpreted with caution. Reliability is dependent not only on the measurement properties of a test but also on the statistical distribution of the studied subgroup. For example, Appendix M shows that subgroup sizes may vary considerably, which results in natural variation in reliability coefficients. Alternatively, $\alpha$, 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 2023 next-generation MCAS content areas, which are described in section 3.2. Cronbach's $\alpha$ coefficients for subcategories were calculated via the same formula defined previously using only the items of a given subcategory in the computations. Results are presented in Appendix M. Lower reliabilities on subcategory scores are associated with lower numbers of items. For example, the grade 3 reporting category Geometry has only 4 items, resulting in a predictably very low reliability statistic of 0.49, the reliability coefficients for the reporting subcategories range from 0.49 to 0.88, with a median of 0.72 and a standard deviation of 0.09. Because they are based on a subset of items rather than the full test, subcategory reliabilities were typically lower than were overall test score reliabilities, approximately to the degree expected based on classical test theory (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 subcategories.

## 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 2023 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 N shows achievement-level distributions by content area and grade for the 2023 next-generation MCAS tests. Note that the information presented in Appendix N is based on all test takers reported with an achievement level.

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 examined 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 2023 next-generation MCAS tests because it is easily adaptable to all types of testing formats, including mixed formats.

The DAC estimates reported in Tables 3-41 and 3-42 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 2023 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 the same classification) signified overall consistency.

Cognia also measured consistency on the 2023 next-generation MCAS tests using Cohen's (1960) coefficient $\kappa$ (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 $\kappa$ 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 Tables 3-41 and 3-42. 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 (0.78–0.86), consistency (0.69–0.80), and kappa (0.56–0.69) indicate that most 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 3-41, the conditional accuracy value is 0.84 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, 84% would be expected to be in this classification when categorized according to their observed scores. Similarly, a consistency value of 0.73 indicates that 73% 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 3-42 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 3-42 range from 0.90–0.93 and 0.86–0.90, respectively. The false positive and false negative decision rates at the *Partially Meeting Expectations/Meeting Expectations* threshold range from 4%–5% and 3%–5%, respectively. 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 3-41. 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 | 0.82 (0.74) | 0.61 | 0.84 (0.73) | 0.81 (0.75) | 0.83 (0.77) | 0.77 (0.61) |
| | 4 | 0.81 (0.73) | 0.59 | 0.83 (0.72) | 0.82 (0.76) | 0.80 (0.73) | 0.74 (0.54) |
| | 5 | 0.82 (0.75) | 0.61 | 0.83 (0.74) | 0.81 (0.75) | 0.84 (0.78) | 0.72 (0.52) |
| | 6 | 0.78 (0.69) | 0.56 | 0.85 (0.77) | 0.77 (0.69) | 0.78 (0.70) | 0.64 (0.48) |
| | 7 | 0.81 (0.73) | 0.60 | 0.83 (0.73) | 0.81 (0.75) | 0.79 (0.72) | 0.78 (0.62) |
| | 8 | 0.78 (0.69) | 0.57 | 0.85 (0.76) | 0.78 (0.69) | 0.77 (0.69) | 0.70 (0.55) |
| | 10 | 0.80 (0.71) | 0.58 | 0.83 (0.71) | 0.77 (0.69) | 0.82 (0.75) | 0.78 (0.67) |
| Mathematics | 3 | 0.83 (0.76) | 0.65 | 0.84 (0.74) | 0.84 (0.78) | 0.83 (0.78) | 0.80 (0.67) |
| | 4 | 0.84 (0.78) | 0.66 | 0.85 (0.74) | 0.83 (0.77) | 0.85 (0.80) | 0.82 (0.70) |
| | 5 | 0.86 (0.80) | 0.67 | 0.76 (0.66) | 0.86 (0.82) | 0.87 (0.82) | 0.82 (0.68) |
| | 6 | 0.86 (0.80) | 0.69 | 0.85 (0.75) | 0.86 (0.81) | 0.86 (0.81) | 0.83 (0.72) |
| | 7 | 0.85 (0.78) | 0.68 | 0.86 (0.78) | 0.85 (0.79) | 0.84 (0.78) | 0.84 (0.73) |
| | 8 | 0.84 (0.78) | 0.67 | 0.80 (0.72) | 0.85 (0.79) | 0.85 (0.79) | 0.85 (0.73) |
| | 10 | 0.86 (0.80) | 0.69 | 0.73 (0.58) | 0.86 (0.82) | 0.88 (0.84) | 0.85 (0.76) |
| STE | 5 | 0.81 (0.73) | 0.59 | 0.81 (0.68) | 0.82 (0.76) | 0.80 (0.73) | 0.79 (0.65) |
| | 8 | 0.82 (0.75) | 0.62 | 0.84 (0.73) | 0.82 (0.76) | 0.83 (0.77) | 0.76 (0.59) |
| Biology | HS | 0.83 (0.76) | 0.65 | 0.82 (0.74) | 0.81 (0.75) | 0.85 (0.79) | 0.83 (0.72) |
| Introductory Physics | HS | 0.82 (0.75) | 0.64 | 0.68 (0.58) | 0.81 (0.75) | 0.85 (0.79) | 0.88 (0.79) |

**Table 3-42. 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 | 0.95 (0.93) | 0.02 | 0.03 | 0.91 (0.87) | 0.05 | 0.04 | 0.96 (0.94) | 0.02 | 0.01 |
| | 4 | 0.95 (0.92) | 0.02 | 0.03 | 0.90 (0.86) | 0.05 | 0.05 | 0.96 (0.95) | 0.03 | 0.01 |
| | 5 | 0.95 (0.93) | 0.02 | 0.03 | 0.91 (0.87) | 0.05 | 0.04 | 0.96 (0.95) | 0.03 | 0.01 |
| | 6 | 0.94 (0.91) | 0.03 | 0.03 | 0.91 (0.87) | 0.05 | 0.05 | 0.94 (0.91) | 0.03 | 0.03 |
| | 7 | 0.94 (0.92) | 0.03 | 0.03 | 0.91 (0.87) | 0.05 | 0.05 | 0.96 (0.94) | 0.03 | 0.02 |
| | 8 | 0.94 (0.91) | 0.03 | 0.03 | 0.91 (0.87) | 0.04 | 0.05 | 0.94 (0.91) | 0.03 | 0.03 |
| | 10 | 0.96 (0.94) | 0.02 | 0.02 | 0.91 (0.87) | 0.05 | 0.04 | 0.93 (0.90) | 0.04 | 0.03 |
| Mathematics | 3 | 0.96 (0.94) | 0.02 | 0.02 | 0.92 (0.88) | 0.04 | 0.04 | 0.96 (0.94) | 0.02 | 0.02 |
| | 4 | 0.96 (0.95) | 0.01 | 0.02 | 0.92 (0.89) | 0.04 | 0.04 | 0.96 (0.94) | 0.02 | 0.02 |
| | 5 | 0.97 (0.95) | 0.02 | 0.02 | 0.92 (0.88) | 0.05 | 0.04 | 0.97 (0.96) | 0.02 | 0.01 |
| | 6 | 0.96 (0.95) | 0.01 | 0.02 | 0.92 (0.89) | 0.04 | 0.04 | 0.97 (0.96) | 0.02 | 0.01 |
| | 7 | 0.95 (0.93) | 0.02 | 0.03 | 0.93 (0.90) | 0.04 | 0.04 | 0.97 (0.95) | 0.02 | 0.01 |
| | 8 | 0.95 (0.92) | 0.03 | 0.03 | 0.93 (0.90) | 0.04 | 0.03 | 0.97 (0.96) | 0.02 | 0.01 |
| | 10 | 0.97 (0.95) | 0.01 | 0.02 | 0.93 (0.90) | 0.04 | 0.03 | 0.97 (0.95) | 0.02 | 0.02 |
| STE | 5 | 0.95 (0.93) | 0.02 | 0.03 | 0.90 (0.87) | 0.05 | 0.05 | 0.95 (0.93) | 0.03 | 0.02 |
| | 8 | 0.95 (0.93) | 0.02 | 0.03 | 0.91 (0.87) | 0.05 | 0.04 | 0.96 (0.95) | 0.02 | 0.01 |
| Biology | HS | 0.96 (0.94) | 0.02 | 0.02 | 0.92 (0.89) | 0.05 | 0.03 | 0.96 (0.94) | 0.02 | 0.02 |
| Introductory Physics | HS | 0.95 (0.93) | 0.03 | 0.02 | 0.92 (0.89) | 0.05 | 0.03 | 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 3-41 and 3-42 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 in grades 3–8 and 10 and grades 5, 8 science and technology/engineering (STE) and high school biology and introductory physics: *Not Meeting*

*Expectations, Partially Meeting Expectations, Meeting Expectations,* and *Exceeding Expectations.* (This language is different than that used for the high school chemistry and technology/engineering tests.)

Parent/Guardian Reports and student results labels are the only printed reports; one copy of each was mailed to districts for distribution to schools. The schools disseminate the reports to parents/guardians. Parent/Guardian Reports were also made available to schools and districts online in PearsonAccess Next (PAN). See section 3.8.1 for additional details of the Parent/Guardian Report.

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 Parent/Guardian Report is generated for each student eligible to take the MCAS tests based on Reporting Business Requirements. It is a standalone 4-page (11” x 17” sheet of paper) color report that is folded in half. A sample report is provided in Appendix O.

The report is designed to present parents/guardians with a detailed summary of their student’s MCAS performance and to enable comparisons with other students at the school, district, and state levels. 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 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. Local Student ID was added to the report at all grades. When available, the student’s graduating class is printed on the report for high school students. The cover also presents general information about the test, and website information for parent/guardian resources. The front page also contains text from the Family Guide pertaining to the student’s grade in fall of 2023 for all subjects.

Each content area page of the report contains the achievement level, scaled score, and standard error of the scaled score for the content area. If the student does not receive a scaled score, the reason is displayed where the achievement level would be displayed. 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. All achievement level descriptors are presented as part of the scale score graphical display for each content area. A horizontal gray bar was used to represent the standard error for next-generation content areas. A vertical black bar was used to represent the standard error for legacy content areas.

For next-generation tests, the student’s scaled score is compared to the average scaled score at the school, district, and state levels, based on business requirements that document student inclusion rules for aggregations. These scaled score values are color-coded based on the corresponding achievement levels. The mode of testing—paper, or computer—for the subject is indicated on each content area page. Up to 3 years of scores, including the current year, are reported where available for ELA and mathematics. Growth percentiles are reported for ELA and mathematics in all grades except in grade 3.

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.
- At the bottom of each subject page, grade-specific resources are provided to help parents with the next steps.

Reporting Categories for each content area is reported in a table presenting the points earned by the student for the reporting category, the total points possible for the reporting category, average points earned in the school, district, and state and the average points earned by students at or near the Meeting Expectations cut score. Science practices are also summarized in the table.

The Science practices reported in the item grid illustrate the assignment of specific practices to items associated with that practice (practices are not assigned to all items). A '/' was used to indicate when an item does not have a practice assigned.

For students in grade 10 or higher, a template was created for students who previously passed or previously failed high school STE and did not take the spring 2023 STE tests. The science page for these students was replaced by the back page image that is provided by DESE. Students in grade 10 or higher taking ELA and mathematics and chemistry or technology/engineering were reported on the previously existing template for legacy sciences.

The fourth or back page of the report shows the results for science for students in grades 5, 8, or high school, for students in grades where science is not assessed (3, 4, 6, and 7), or high school students who did not participate in the science assessment, the back page shows the image that was provided by DESE. Report templates are used based on reporting rules provided by DESE.

## 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
- Local Student ID
- student's scaled score and achievement level for each content area (or the reason the student did not receive a score)
- Additionally, for high school, the student's graduating class and the CD status for each content area are printed.

## 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, 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 2023 next-generation MCAS administration is included in Appendix P.

### 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 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 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 the following criteria:

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

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

The quality-assurance group uses a checklist to implement its procedures. Once the checklist is completed, sample reports are circulated for review by psychometric and program management staff. The appropriate sample reports are then sent to 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 regarding 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, 3.3, and 3.4.

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 the *2019 Next-Generation MCAS and MCAS-Alt Technical Report*, Appendix S for study details), 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 most 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 most 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. 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 were used to filter out the results of students who did not spend sufficient time on their answers. The data from the remaining motivated students were 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 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, an 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 2018–19 administration. 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 grades 3–8 in 2017–18. 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. They first compared student results on the Next Generation MCAS tests to course grades and course-taking in middle school 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 were significantly associated with taking advanced mathematics courses. Convergent validity evidence was also reported between MCAS test portions and subjects.

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

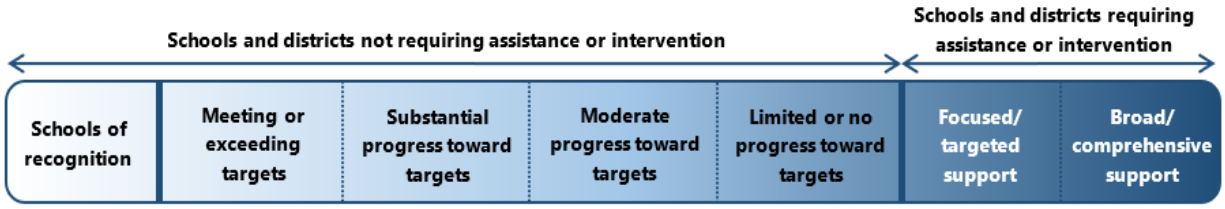
DESE takes many steps to support the intended uses of MCAS data. (The intended uses are listed in section 2.3 of this report.) This section will examine some of the reporting systems and policies designed to address each use.

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, which requires that students submit an MCAS-Alt Skills Survey as well as a collection of work samples and other documentation that demonstrates their proficiency on the state standards. Technical information on the MCAS-Alt is presented in Chapter 4 of this report.

2. Providing information to support program evaluation at the school and district levels.
3. Providing transparency into student performance through comprehensive reporting on the results of individual students, schools, districts, and the state.

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 O) 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.

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. These reports can help educators review classroom and school patterns, reflect on practice from last year, and plan for incoming classes based on previous performance. 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, low-income 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 *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**

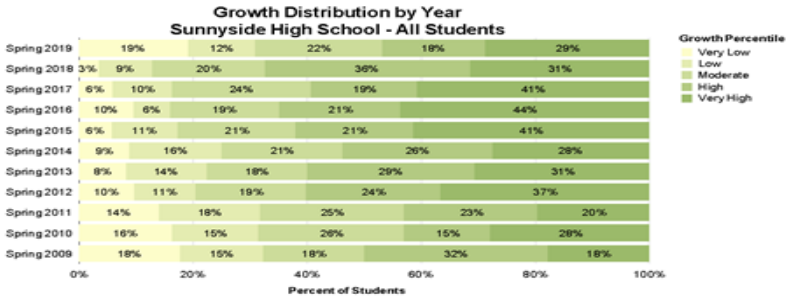
**All Students Students** (161)

**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 less than 10 and school results only include students enrolled in the school since October 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.

Aggregated 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 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 DESE's efforts to support the use of MCAS results by parents, educators, and leaders in addition to the use of MCAS results as a 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 conduct, 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 the most 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 to graduate from high school. Students with the most 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 MCAS-Alt is intended to document the student's achievement and progress in learning the skills, knowledge, and concepts outlined in the state's curriculum frameworks, and to

- provide a basis for including difficult-to-assess students in statewide assessment and accountability systems;
- determine whether students with the most 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; and
- assist teachers in providing challenging academic instruction.

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

### 4.1.3 Format

The MCAS-Alt consists of a structured set of "evidence" collected during instructional activities in each subject to be assessed during the school year, plus a standardized MCAS-Alt Skills Survey that measures the degree to which students have already learned the range of skills covered by a particular strand or domain of the frameworks. Teachers are required to use the results of the skills survey to identify particular standards and levels of complexity at which to begin assessing the student. The MCAS-Alt also includes the student's demographic information and weekly schedule, parent/guardian verification and signoff, and a school calendar, all of which are submitted to the state each spring. Preliminary 2023 results were reported to schools in June, with final results provided in September.

The Department's *Resource Guide to the Massachusetts Curriculum Frameworks for Students with Disabilities* (the *Resource Guide*) describes the content to be assessed by the 2023 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* provides strategies for adapting and using the state's learning standards to instruct and assess students taking the MCAS-Alt. The fall 2022 *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 are listed in Table 4-1.

**Table 4-1. MCAS-Alt Requirements in Each Category**

| Grade | ELA Strands Required | Mathematics Domains Required | STE Strands Required |
|---|---|---|---|
| 3 | ▪ Language<br>▪ Reading<br>▪ Writing | ▪ Operations and Algebraic Thinking<br>▪ Measurement and Data | |
| 4 | ▪ Language<br>▪ Reading<br>▪ Writing | ▪ Operations and Algebraic Thinking<br>▪ Numbers and Operations – Fractions | |
| 5 | ▪ Language<br>▪ Reading<br>▪ Writing | ▪ Number and Operations in Base Ten<br>▪ Number and Operations – Fractions | For any three of the four STE disciplines,* select one core idea in each discipline and assess six entry points within each core idea. |
| 6 | ▪ Language<br>▪ Reading<br>▪ Writing | ▪ Statistics and Probability<br>▪ The Number System | |
| 7 | ▪ Language<br>▪ Reading<br>▪ Writing | ▪ Ratios and Proportional Relationships<br>▪ Geometry | |
| 8 | ▪ Language<br>▪ Reading<br>▪ Writing | ▪ Expressions and Equations<br>▪ Geometry | For any three of the four STE disciplines,* select one core idea in each discipline and assess six entry points within each core idea. |
| 10 | ▪ Language<br>▪ Reading<br>▪ Writing | Any three of the five mathematics conceptual categories:<br>▪ Functions<br>▪ Geometry<br>▪ Statistics and Probability<br>▪ Number and Quantity<br>▪ Algebra | Select three core ideas in <u>one</u> of the following disciplines:<br>▪ Biology<br>▪ Chemistry<br>▪ Introductory Physics<br>or<br>▪ Technology/Engineering |

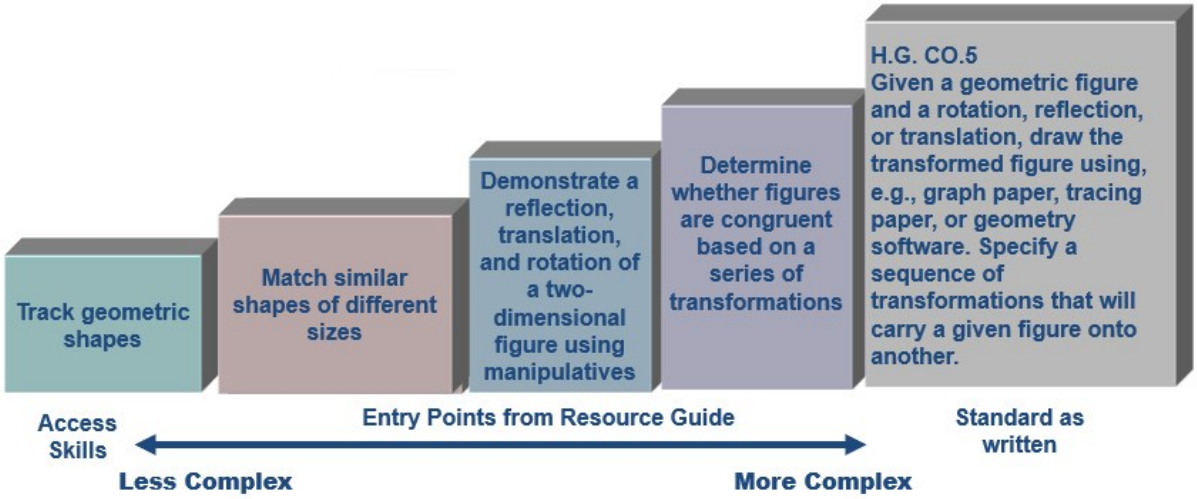
** Earth and Space Science, Life Science, Physical Sciences, Technology/Engineering*

### 4.2.1.1 Access to the Grade-Level Curriculum

Students with disabilities are expected to achieve the same standards as their peers who do not have disabilities. However, they may need extensive support to learn the necessary knowledge and skills and are likely to require instruction in smaller segments and at a slower pace. The *Resource Guides to the Massachusetts Curriculum Frameworks for Students with Disabilities* identify student-centered academic outcomes, called entry points, based on each grade-level content standard. The *Resource Guides* are intended to assist educators in teaching and assessing appropriately challenging, standards-based academic skills and content aligned with grade-level standards, as required by law. Entry points consist of academic outcomes based on the "essence" of the grade-level content but presented at modified levels of complexity and difficulty. Entry points provide a roadmap for students to make steady progress toward eventually meeting standards at grade-level complexity.

In a small number of cases where students with the most significant cognitive abilities cannot yet address entry points even at the lowest levels of complexity and even with the use of instructional accommodations, those students are instructed and assessed on the acquisition of access skills, which describe the communication and motor skills practiced during age-appropriate activities based on the standards. Entry points and access skills are listed in the *Resource Guides* in ELA, mathematics, and STE for every curriculum framework standard, available online at www.doe.mass.edu/mcas/alt/resources.html.

**Figure 4-1. Model of a Method to Access the Grade-Level Curriculum Using Entry Points That Address the Essence of the Standard for Students Who Take the MCAS-Alt (Mathematics Example)**




## How Resource Guides Were Developed

After each curriculum framework was developed or subsequently revised, DESE convened panels of experts in each of three content areas (ELA, mathematics, and STE) to adapt the general education curriculum standards for students with the most significant cognitive disabilities. Panelists included content specialists, assessment experts, special educators familiar with students with the most significant cognitive disabilities, higher education faculty, parents and advocates, and members of the state's contractor team. Panelists are listed for each content area on the acknowledgements page of each on the Resource Guides here: www.doe.mass.edu/mcas/alt/resources.html.

Each panel reviewed the standards in their respective content area and identified the big ideas, key skills, and content knowledge—the so-called "essence"—contained in each standard. Once panelists agreed upon the essence, they determined "entry points," standards-based outcomes at successively lower levels of complexity than are typically expected of students who are achieving the grade-level standards as originally written. First, the panels determined entry points at the lowest level of complexity at which a student could address the standard without losing its essence. Then, they determined additional entry points at successively higher levels of complexity so teachers could identify and select the entry point at a challenging and attainable level of complexity appropriate for each student. This "continuum of complexity" allows teachers to progress to higher levels of complexity once lower complexity entry points are mastered by the student.

The process of developing the essence and entry points was repeated in each of the three content areas and was replicated each time revisions were made to the curriculum frameworks (1999; 2001; 2006; 2011; 2016; 2017). Subsequently, special educators familiar with students with the most significant cognitive disabilities developed access skills appropriate for students who are unable to address the content and skills at even the lowest level of complexity. Access skills include only motor and communication skills addressed during a standards-based activity in the required strand/domain and are intended for a very small number of students with the most unique, complex, and significant cognitive disabilities. Each Resource Guide lists the standards as written for students in each grade together with entry points and access skills intended for students with the most cognitive disabilities who are designated to participate in the MCAS-Alt.

### 4.2.1.2 Assessment Design

The MCAS-Alt assessments for ELA–language, ELA–reading, mathematics, and legacy high school STE (chemistry and technology/engineering only) consist of a completed MCAS-Alt Skills Survey, a collection of primary evidence, supporting documentation, and other required information.

## MCAS-Alt Skills Survey

The MCAS-Alt Skills Survey (see Appendix Q) is a standardized component of the MCAS-Alt that must be administered by the teacher to each student before selecting an entry point or access skill in the subject required for assessment. The survey determines a student's current level of academic knowledge, skills, and abilities across a broad range of standards. The results of the skills survey are intended to be used as the basis for selecting an entry point or access skill listed in the *Resource Guide* in each subject scheduled for assessment. The survey is also intended to familiarize teachers with the range of entry points in a strand/domain that are available for the assessment.

The survey lists the important skills in each strand/domain/conceptual category/discipline from least to most complex. The skills represented on each survey were identified in collaboration with content experts in order to assess students with the most significant cognitive disabilities on skills that represent the "knowledge of most worth" within each strand ranging from low to high complexity.

To complete the skills survey, teachers may use the sample tasks provided on the survey, design their own simple tasks, use classroom observations, class assignments, progress reports, or locally administered assessments to determine the degree to which the student can perform each skill listed in the survey. A sample strand from the survey is shown in Figure 4-2.

A follow-up skills survey, though not required, is recommended *after* the selected skill has been taught to note the student's progress, especially if the student will attend a different classroom the following year.

**Figure 4-2. MCAS-Alt Skills Survey–Reading Sample Strand**

**Reading (Informational or Literary Text)**

| | **Based on a literary or informational text** read by or to the student, student can: | A 0% (unable) | B Up to 25% (rarely) | C Up to 50% (occasionally) | D Up to 75% (more often than not) | E Up to 100% (almost always) |
|---|---|---|---|---|---|---|
| 1. | Identify the main character(s) in the text. | | | | | |
| 2. | Identify the setting of the text. | | | | | |
| 3. | State key details from the text. | | | | | |
| 4. | Identify events (or ideas) presented in the text. | | | | | |
| 5. | Identify the central (main) idea of the text. | | | | | |
| 6. | Explain *why* or *how* something occurred in the text. | | | | | |
| 7. | Identify and define unknown words in the text; or match words or phrases from the text to their meaning. | | | | | |
| 8. | Differentiate between a fact and the author's opinion. | | | | | |
| 9. | Describe the author's point of view. | | | | | |

**Instructions for Completing the Skills Survey**

Teachers are instructed to conduct the MCAS-Alt Skills Survey for the most significant entry points listed in the Resource Guide in the required strand/domain for a student in that grade. Next, they check one box (A–E) for each skill in the required strand/domain(s).

Teachers may use any combination of the following methods to conduct a brief assessment of each skill:

a) observations, informal assessments, progress reports, or classroom work

b) 2–4 tasks, based on the examples provided in the survey form or designed by the teacher that are accommodated for each student's instructional level and needs

If using specific tasks or activities to assess the student, the following protocol should be used:

1) Present the first task to the student.

2) If the student does not respond on the first attempt, repeat the task with a verbal reminder or other prompt (if needed), but do not give the answer. (Note: If a prompt is given, the response may be accurate, but is not independent.)

3) If the student responds to the first task, give a second, more complex task. Repeat with a prompt if needed. Make notes on the survey form to remind you of the student's performance of each task.

4) If the student does not respond to the second task, even with a prompt, do not introduce a third task. Simply mark an "X" in the column (A, B, C, D, or E) that most closely describes their performance of the skill.

5) Introduce the next task in the survey. Repeat steps 2 through 4 until all skills in the required strand/domain are assessed.

Once the survey has been completed for each required strand/domain, review the results, and proceed as follows:

- Select a related or higher-level-of-complexity entry point from the Resource Guide based on any skill that has been checked in columns A, B, or C.
- Do not select an entry point for any skills checked in columns D or E.
- If column A ("unable to perform the skill") is checked for all skills in the strand/domain, consider assessing an access skill (i.e., a motor or communication skill).
- If columns D and/or E are checked for most of the skills in the strand/domain, then the IEP team should consider whether the standard MCAS test (paper or online) would be more appropriate for the student in that subject.

Submit a completed MCAS-Alt Skills Survey for each assessed strand just after the Strand Cover Sheet in each student's MCAS-Alt. A strand without a completed Skills Survey will receive a score of *Incomplete*.

**MCAS-Alt Skills Survey Pilot**

In fall 2018, 55 MCAS-Alt training specialists (i.e., special educators selected to be peer trainers) were asked to conduct a pilot study of the MCAS-Alt Skills Survey with one or more students in at least one of three content areas and provide responses to the following questions.

- How difficult was it to administer the skills survey?
- How much time did it take to administer each strand of the survey?
- Did conducting the skills survey help you gain a better understanding of your students' abilities?

- Was the skills survey helpful in guiding you to select appropriate entry points to assess?
- Was the skills survey rating system useful in determining a student's performance?
- Do you have suggestions for how the DESE should communicate this new requirement to teachers for the 2019–2020 school year? (Note: The survey was introduced in 2019–2020, but the state's academic assessments were cancelled due to the impact of the pandemic in spring 2020. The survey was first implemented and scored in the 2020–2021 school year.)

DESE received 48 written responses to the questions listed above. Most respondents said the skills survey was easy to administer, though the duration of administration varied widely (from 5–30 minutes per strand, depending on the student's abilities—surveying lower functioning students was completed more quickly while higher functioning students took longer).

Several said it seemed redundant of other broad-based skills assessments they routinely conduct at the start of each school year, though many said the MCAS-Alt Skills Survey was more formal, sequential, systematic, and standards-based. Respondents were about equally divided on the question of its effectiveness in helping gain a better understanding of their student(s), though many said it helped them identify the standards on which to focus for instruction and assessment. A few said their students surprised them with new skills they hadn't been aware they had mastered, and many said it was most helpful in cases when surveying students with whom they were less familiar. Many felt the survey helped them expand their understanding of possible entry points to select for assessment and the range of skills they were willing to teach and assess.

While most respondents acknowledged that the survey would require additional time to conduct, a large proportion said it was not overly time-consuming to administer. A few said it had saved them time, since it revealed the areas that needed the greatest instructional focus and gave them ideas for areas to assess. Some suggested the survey would be a good informal pre- and post-assessment conducted at different points throughout the school year, which could assist with progress monitoring and passing along orientation information to a new teacher the following year. Most felt the skills survey process will make sense to teachers when it is introduced, though they might be unhappy about the additional work requirement and suggested it be made optional.

As a result of feedback from the pilot study, the following adjustments were made to the operational MCAS-Alt Skills Survey:

- The skills survey was incorporated into the online MCAS-Alt forms and graphs application so it could be completed online.
- Multiple skills that had been combined were separated into separate skills.
- A training unit was developed to prepare teachers for implementation.
- The designations used in headers for columns A through E to rate each skill were edited to include both percentages of independence AND descriptors of the students' achievement of the skill (see Figure 4-3 below).
- Additional consultation occurred with content specialists to develop examples of assessment activities, ensure fidelity to the standards, and provide coverage of the most significant entry points across all ability levels.
- Instructional examples were added to the listed skills in Science and Technology/Engineering.

**Figure 4-3. Descriptors for Each Column Used on the Skills Survey**

| A | B | C | D | E |
|---|---|---|---|---|
| Student is **unable** to perform this skill.<br>---------OR--------<br>Teacher is unable to assess student on this skill. | Student is just starting to learn this skill and demonstrates the skill only **rarely** without support.<br><br>Student performs this skill accurately with **0–25% independence**.<br>---------OR--------<br>Student performs this skill independently with **0–25% accuracy**. | Student demonstrates this skill **intermittently** and only **occasionally** without support.<br><br>Student performs this skill accurately with **26–50% independence**.<br>---------OR--------<br>Student performs this skill independently with **26–50% accuracy**. | Student demonstrates this skill **more often than not** without support.<br><br>Student performs this skill accurately with **51–75% independence**.<br>---------OR--------<br>Student performs this skill independently with **51–75% accuracy**. | Student demonstrates this skill **almost all the time** without support.<br><br>Student performs this skill accurately with **76–100% independence**.<br>--------OR---------<br>Student performs this skill independently with **76–100% accuracy**. |

## Primary Evidence

For the evidence collection portion of the MCAS-Alt, the ELA, mathematics, and STE assessments require the inclusion of an instructional data chart and two or more pieces of primary evidence in each assessed strand, plus other supporting documentation that shows or describes the student's performance of the targeted skill.

The ELA–language, ELA–reading, and all required mathematics strands must include a data chart (e.g., field data chart, line graph, or bar graph) that indicates the following:

- the student's performance of 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
- progress over time, including an indication that the student has attempted a new skill

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

- work samples (created by the student or dictated to a scribe using the student's primary mode of communication)
- photographs of one or more classroom activities
- audio or video clips of the student performing the targeted activity

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 what the student was asked to do and how the task or activity was conducted,
- the percentage of accuracy for the task or activity, and

- the percentage of independence during the task or activity (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 comprise the "core set of evidence" required in each strand, with the exception (noted below) of the ELA–Writing strand and next-generation STE strands.

The MCAS-Alt for ELA–Writing consists of a skills survey, 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. Final writing samples are included in the final score.

The MCAS-Alt assessments for STE in grades 5 and 8, and high school biology and introductory physics consist of primary evidence in three STE disciplines. Each discipline includes evidence of three entry points within the same core idea. STE evidence consists of the MCAS-Alt Skills Survey plus three work samples that integrate the STE content with three of the eight science practices described in the 2016 Massachusetts Curriculum Framework for STE. The STE assessment for high school consists of a skills survey and three different core ideas in one discipline (either biology or introductory physics). Each core idea consists of three work samples documenting three different science practices, one for each summary sheet.

A detailed description of the instructions given to educators who are conducting the MCAS-Alt is provided in section 4.3, MCAS-Alt Test Administration.

### 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 context of the instructional 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 self-awareness, perceptions, choices, decision-making, and self-assessment of the work he or she created and/or the learning that occurred as a result. For example, a student may respond to questions such as these:
    - 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 2023 MCAS-Alt* (see Appendix R), using the MCAS-Alt Rubric for Scoring Each Strand, to produce a subscore for the strand based on the following:

- **completeness** of assessment 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** of 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 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 the Analysis and Reporting Business Requirements (Appendix P) 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.*

## 4.2.2 Test Development

### 4.2.2.1 Rationale

AA-AAAS 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 statewide MCAS so their performance of skills and knowledge of content described in the state's curriculum frameworks can be assessed and so that they are 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 the most significant cognitive disabilities. This requirement ensures that students with the most 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 use of curriculum resources provided by DESE, 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 informally reported unanticipated gains in student achievement.

### 4.2.2.2 Test Specifications

#### MCAS-Alt Skills Survey

Each strand must include a completed MCAS-Alt Skills Survey indicating the results of the student's performance in a broad range of skills. The information compiled in the skills survey must be used by the educator to select a targeted skill from the *Resource Guide* in the content area and strand(s) required for assessment. Only those skills (i.e., entry points and access skills) that the student was unable to perform

accurately and independently at least 50 percent of the time on the MCAS-Alt Skills Survey may be selected by the student's teacher for the MCAS-Alt.

## Evidence for English Language Arts (Language and Reading only), Mathematics, and Legacy STE (Chemistry and Technology/Engineering) Strands

Each portfolio strand must include a data chart documenting the student's performance of the targeted skill being assessed in the required content area (i.e., the percentage of accuracy and independence of each performance). Data are collected on at least eight different dates to determine the degree to which the skill has been mastered. On each date, the data must indicate the percentage of correct versus inaccurate responses given by the student, and whether the student required cues, prompts, or other assistance to respond (i.e., the overall percentage of independent responses by the student). Each data chart must include a brief description of activities conducted on each date and must describe how the activity addressed the measurable outcome being assessed. Data are 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 from either a single activity or task; or from a series of responses to specific tasks summarized for each date.

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

The following information must be provided either on a Work Description or on the evidence itself:

- student's name
- date
- content area, strand/domain, and learning standard being assessed
- entry point being assessed during the activity
- a summary of the percent of student's accuracy and independence during the activity
- description of the activity

## Evidence for ELA–Writing

The ELA–Writing strand requires a completed MCAS-Alt Skills Survey and at least three writing samples that demonstrate the student's expressive communication skills, based on any combination of the following text types:

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

In addition to three writing samples, one *baseline* sample must be submitted, which may include either an outline, completed graphic organizer, or draft of a writing assignment. The baseline sample should provide information to guide additional instruction in writing in that text type. Teachers are also required to pre-score the student's three final writing samples using a rubric provided by DESE for that purpose. See Appendix S for the Scoring Rubric for ELA–Writing.

## Evidence for Next-Generation Science and Technology/Engineering (STE) Strands (Grades 5 and 8)

The format described below is intended to encourage the teaching of units of science based on a core idea, rather than assessing isolated skills. Teachers are directed to complete these steps:

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

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

**Step 2: Conduct the STE Skills Survey** available to determine the optimal grade-span at which to select entry points for the student. The STE Skills Survey must be conducted once for the entire STE content area, not for each discipline, and must include all eight science practices.

**Step 3: Select a core idea** within the chosen discipline that is relevant and that engages and challenges the student.

**Step 4: Select three (3) entry points or access skills.** Three (3) different science practices must be addressed within the selected entry points or access skills. This step encourages teachers to design inter-related activities that address a theme or unit of study.

**Step 5: List the following information on each STE Summary Sheet:**

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

**Step 6: Attach three pieces of primary evidence (i.e., work samples) to its corresponding completed STE Summary Sheet.** Photographs and/or videos may be submitted as primary evidence if they are labeled and clearly show the final product of instruction.

## Evidence for High School STE Strands

Assessment formats differ depending on the educator's selection of either the next-generation or legacy disciplines described below.

**Step 1: Choose one (1)** of the following next-generation STE disciplines:

- Biology **OR** Introductory Physics

**Step 2: Conduct the MCAS-Alt STE Skills Survey** to determine the grade-span at which to select entry points in each science practice for the student. Only one skills survey is required for high school Biology and Introductory Physics.

**Step 3: Select three (3) *core ideas* within the chosen discipline from the Next-Generation STE Resource Guide** that engage and challenge the student.

**For *each* core idea:**

**Step 4: Select three (3) entry points or access skills. Three (3) different science practices must be addressed** within the selected entry points or access skills. If entry points seem too complex at the grade level of the student, select entry points from earlier grade-level clusters in the same core idea. Use the information in the STE skills survey to assist with selection.

**Follow Steps 5 and 6 above for each of the three core ideas.**

# 4.3 MCAS-Alt Test Administration

## 4.3.1 Preparing the MCAS-Alt for Submission

The student's MCAS-Alt must include all elements listed below. Required forms can either be photocopied from those found in the *2023 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 binder
- **MCAS-Alt cover sheet** containing important information about the student
- **Student's introduction** to his/her MCAS-Alt produced as independently as possible by the student using his/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"
- **Verification form** signed by a parent, guardian, or primary care provider signifying that he or she has reviewed the student's completed MCAS-Alt materials 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 student's completed MCAS-Alt materials.)
- **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; the calendar is used to verify the dates specified on the data chart and in other evidence.
- **MCAS-Alt Skills Survey** completed for each strand/domain/discipline required for assessment
- **Strand cover sheet** describing the accompanying set of evidence for a particular strand
- **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.)
- **Writing scoring rubric** for ELA–Writing only completed by the teacher for each of three final writing samples
- **STE Summary Sheet** completed by the teacher (as detailed in section 4.2.2.2)

The contents listed above, plus all primary evidence and supporting documentation, constitute the student's MCAS-Alt.

## 4.3.2 Participation Requirements

### 4.3.2.1 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 statewide assessments that correspond with the grades in which they are reported in DESE's Student Information Management System (SIMS). Students with the most 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 their 504 plan.

### 4.3.2.2 Participation Guidelines

A student's IEP team (or 504 plan coordinator, in consultation with other staff) determines how the student will participate in MCAS and other state- and district-wide assessments 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 their cognitive disabilities. The participation guidelines and the characteristics to consider for students taking the MCAS-Alt are described below and in the participation section of the *Educator's Manual for MCAS-Alt* (available at www.doe.mass.edu/mcas/alt/resources.html). Information on how a student with a disability will participate in state- and district-wide testing must be documented in the student's IEP or 504 plan and 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.

A decision-making flow chart, entitled the MCAS Decision-Making Tool for MCAS Participation (see Appendix T), was developed in 2003 and updated in 2020 and is intended for use by IEP teams to make annual decisions regarding appropriate student participation in MCAS in each content area. Recent revisions to the tool included the addition of specific criteria determining which students may be considered for accommodations when taking the standard MCAS and which are eligible to participate in the MCAS-Alt. The criteria are located in Appendix U. IEP teams are strongly encouraged to use the tool to guide the team's discussion and decision-making regarding statewide assessments.

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

- Can the student demonstrate knowledge and skills, either fully or partially, on the **standard MCAS test under routine conditions**?
- Can the student demonstrate knowledge and skills, either fully or partially, on the **standard MCAS test with accommodations**? If so, which accommodations are necessary for the student to participate?
- If no to the above questions and the student has a significant cognitive disability, see the options below to determine whether the student qualifies to take the **alternate assessment** (MCAS-Alt). (**Note:** Alternate assessments are intended only for students with the most significant cognitive disabilities who are unable to take standard MCAS tests, even with accommodations. Students should not be identified for alternate assessments based solely on a particular disability, a placement in a specific classroom or program, previous low achievement on the tests, or EL status.)

The student's team must review the options provided in Figure 4-4. 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 4-4. Participation Guidelines**

OPTION 1

| Characteristics of Student's Instructional Program and Local Assessment | Recommended Participation in MCAS |
|---|---|
| *If the student is*<br>a) generally able to demonstrate knowledge and skills on a computer- or paper-based test, either with or without test accommodations, ***and*** is<br>b) working on learning standards **at, near, or somewhat below grade-level expectations,** | *Then*<br>the student should take the computer- or paper-based **MCAS test**, either with or without accommodations. |

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*<br>a) **generally unable** to demonstrate knowledge and skills on a paper-and-pencil test, even with accommodations; ***and*** is<br>b) working on learning standards that have been **substantially modified** due to the nature and severity of their disability; ***or*** is<br>c) receiving **intensive, individualized instruction** in order to acquire, generalize, and demonstrate knowledge and skills, | *Then*<br>the student should take the **MCAS Alternate Assessment** (MCAS-Alt) in this subject. |

### 4.3.2.3 2023 MCAS-Alt Participation Rates

In ELA, 5,822 students took the MCAS-Alt (1.2 percent); in mathematics, 5,889 students took the MCAS-Alt (1.2 percent); and in STE, 2,382 students took the MCAS-Alt (1.1 percent).

Additional information about MCAS-Alt participation rates is provided in the 2023 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.

## 4.3.3 Educator Training

During October 2022, a total of 1,308 educators and administrators received training on conducting the 2023 MCAS-Alt. Attendees had the option to participate in one of three sessions: an introduction to MCAS-Alt for educators new to the alternate 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
- alternate assessment 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 previous year
- updates and reminders for 2023
- completing quality alternate assessments
- ensuring appropriate designation of students for MCAS-Alt

Topics for the administrators' overview session included the following:

- MCAS-Alt overview
- MCAS-Alt statewide results from previous year
- who should take the MCAS-Alt
- supporting teachers who conduct the MCAS-Alt, principal's role in MCAS-Alt
- the federally mandated cap on the percentage of students who may be assessed through an alternate assessment based on alternate academic achievement standards
- grade-level and Competency Portfolios

In January–March 2023, educators attended virtual training and in-person review sessions during which they were able to discuss their students' alternate assessments that were under development and have their questions answered by MCAS-Alt training specialists (i.e., expert teachers).

### 4.3.4 Support for Educators

A total of 60 MCAS-Alt training specialists were trained by DESE in the 2022–2023 school year to assist and support teachers conducting the MCAS-Alt in their districts, as well as to assist DESE at Department-sponsored assessment training and review sessions in January–March 2023. In addition, DESE staff provided ongoing technical assistance throughout the year via email and telephone to educators with specific questions about their students' alternate assessments.

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

The MCAS-Alt reflects the degree to which a student has learned and applied the knowledge and skills outlined in the Massachusetts curriculum frameworks. The MCAS-Alt 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 *Guidelines for Scoring 2023 MCAS-Alt,* available in Appendix R. The *MCAS-Alt Rubric for Scoring Each Strand* has been used as the basis for scoring the MCAS-Alt 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 Portsmouth, New Hampshire, from April 9 through May 10. DESE and Cognia trained and closely monitored scorers to ensure that scores were accurate.

Each student's MCAS-Alt was 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 assessments to DESE and Cognia staff, and by locking assessments 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 assigned scorers. Communication and coordination among scorers were maintained through daily meetings between FMs, TLs, and scoring leadership to ensure that critical information and uniform scoring rules were implemented across all grade clusters.

## 4.4.2 Recruitment, Training, and Qualification of Scoring Personnel

### 4.4.2.1 Scorer Training Materials

The MCAS-Alt Project Leadership Team (PLT), including DESE and Cognia staff plus four contracted teacher consultants, met daily over the course of scoring in 2023 and periodically throughout the 2022–2023 school year to accomplish the following:

- nominate prospective MCAS-Alt training specialists to serve as scoring specialists for the 2023 scoring institute
- select sample strands to use to train, calibrate, and qualify scorers in 2023
- discuss which recurring issues and concerns to address during the following fall educator training sessions

All sample strands were scored using the *2023 Guidelines for Scoring MCAS-Alt*, noting any scoring concerns or discrepancies that arose during the review. Concerns were resolved by referring to guidelines and requirements in the *2023 Educator's Manual for MCAS-Alt* and by following additional scoring rules agreed upon by the PLT.

Of the alternate assessments 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 scored by all four MCAS-Alt Teacher Consultants. Of the 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 multiple hiring 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 scoring specialists, to assist DESE and Cognia. Eight MCAS-Alt training specialists were selected to participate in scoring and were designated as scoring specialists to assist 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 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 the AltScore computer program. 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, as well as maintaining the accuracy, reliability, and consistency of scorers at tables under their supervision.

#### 4.4.2.4 Qualification of Scorers

Before scoring actual student assessments, each potential scorer was required to take a qualifying assessment consisting of eight sample strands that contained a total of 178 score points. The threshold percentage for qualification on the 178 available score points was 85%.

### Scorers

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 assessments. 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

Originally, a statewide task force comprised of DESE staff (from Special Education and Student Assessment offices), members of the contractor team (then Measured Progress and the University of Kentucky), and the Massachusetts Alternate Assessment Statewide Advisory Committee (a diverse stakeholder group) provided recommendations to DESE on how alternate assessments should be scored, including the criteria on which to base the scores. Some advised DESE to develop scoring criteria based only on student performance, since that is what the standard MCAS assessments measured, rather than assessing how well the student's program provided opportunities to learn and demonstrate knowledge and skills. Others felt that student achievement could not be separated from program effectiveness. In the end, a scoring rubric was developed in which three of five categories are based on student performance; two reflect the effectiveness of the student's program; and one on whether the evidence submitted was sufficient in scope and quantity to allow a score to be determined.

- Completeness: whether the submitted evidence was sufficient to allow a score to be determined
- Level of Complexity: the relative difficulty of academic tasks and knowledge attempted by the student (counts toward the final overall score)
- Demonstration of Skills and Concepts: the accuracy of the student's performance (counts toward the final overall score)
- Independence: cues, prompts, and other assistance provided to the student during tasks and activities being assessed (counts toward the final overall score)
- Self-Evaluation: the extent to which opportunities were provided for the student to evaluate, reflect upon, self-correct, set goals, and select examples of the student's own performance (context of the instruction; not counted toward the final overall score)
- Generalized Performance: the number of contexts and instructional approaches provided to and used by the student to perform tasks and demonstrate knowledge and skills (program quality; not counted toward the final overall score)

#### 4.4.3.1 Scoring English Language Arts (except ELA–Writing), Mathematics, and Legacy Science and Technology/Engineering

Guided by a TL, scorers at each table reviewed and scored assessments from the same grade. Scorers were permitted to ask TLs questions as they reviewed assessments. 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 an assessment to score 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 assessment was verified, including the submission of a completed MCAS-Alt Skills Survey, each strand was scored in the following dimensions:

A. Level of Complexity (LOC)
B. Completeness
C. Demonstration of Skills and Concepts (DSC)
D. Independence (Ind)
E. Self-Evaluation (S-E)
F. Generalized Performance (GP)

The 2023 MCAS-Alt score distributions for all scoring dimensions are provided in Appendix J.

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 and provide scorer comments to the school submitting the assessment. Use of the computer-based scoring application allowed scorers to 1) focus exclusively and sequentially on each assessment product and record the necessary information, rather than keeping track of products they had previously reviewed, and 2) automatically calculate the scores.

### 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 with assessment requirements and with descriptions of the activities documented in the assessment 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 incorporated the criteria listed above.

**Table 4-2. Scoring Rubric for Level of Complexity**

| Score Point | | | | |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |
| The 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 evidence was correctly labeled with the following information:

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

If evidence was not labeled correctly, or if pieces of evidence did not address the measurable outcome stated on the Strand Cover Sheet or work description, that evidence 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 *2023 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 not likely to be considered sufficient to document the relationship between the activity and the measurable outcome and therefore would result in the exclusion of those data points from being scored.

A score of M (i.e., evidence was missing or was insufficient to determine a score) was given in both Demonstration of Skills and Concepts and Independence if

- a completed data chart documenting the student's performance of the same skill on at least eight dates was not submitted; and/or
- at least two pieces of scorable primary evidence were not submitted.

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

- A completed MCAS-Alt Skills Survey was not submitted for the strand.
- The data chart listed the percentages of *both* accuracy and independence at or above 80 percent at the beginning of the data collection period, indicating that the student was not learning 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 eight different 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 and counted 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 strand contains insufficient information to determine a score. | Student's performance is primarily inaccurate and demonstrates minimal understanding in this strand.<br>(0%–25% accurate) | Student's performance is limited and inconsistent with regard to accuracy and demonstrates limited understanding in this strand.<br>(26%–50% accurate) | Student's performance is mostly accurate and demonstrates some understanding in this strand.<br>(51%–75% accurate) | Student's performance is accurate and is of consistently high quality in this strand.<br>(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 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 activity.

- 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 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 strand that involve self-correction, task-monitoring, goal-setting, reflection, and overall awareness by the student of their 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 this strand. | 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 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 Writing Scoring Rubric in Appendix S, according to the appropriate text type:

- Opinions/Arguments

- Informative/Explanatory texts
- Narrative (including Poetry)

MCAS-Alt scorers verified the completion of the MCAS-Alt Skill Survey for the strand and that the scores submitted by the teacher were based on the writing sample generated by the student, and not based on any text generated by the teacher. The rubric scores were lowered by scorers in cases where writing rubric scores did not accurately reflect the student's own 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, that must have been reflected in the rubric scores, particularly in the 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 required to be 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 strand.

#### 4.4.3.3 Next-Generation Science and Technology/Engineering

The requirements for STE in grades 5 and 8 included teachers selecting any three (3) of the following STE disciplines:

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

Teachers were required to create one STE strand within each of the three selected disciplines, each based on a different learning standard and core idea.

High school next-generation STE included a selection of either biology or introductory physics. Teachers were required to create three strands within the one selected discipline, each based on a different learning standard and core idea.

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

1. One MCAS-Alt next-generation STE Skills Survey was submitted for the entire content area.
2. The student's name, valid date, % of accuracy, and % independence were listed on at least three STE Summary Sheets.
3. The activities on the three STE Summary Sheets reflected the same core idea.
4. Three different science practices were represented from the selected entry points or access skills.
5. At least three STE Summary Sheets had primary evidence attached.

After verifying the above, the scorer used the *AltScore* program to rate complexity, accuracy, independence, and self-evaluation for the three STE Summary Sheets.

#### 4.4.3.4 Monitoring Scoring Quality

The FM oversees the general workflow in the scoring room and monitors overall scoring consistency and accuracy, particularly among TLs. The TLs ensure that scorers at their table are consistent and accurate in their scoring. Scoring consistency and accuracy are 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 how responses will be (or have already been) scored by another randomly selected scorer. For each scored assessment, at least one was double-scored for each scorer each morning and afternoon or, at minimum, every fifth assessment each day (i.e., 20% of the total scored by a 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 or if they might resume scoring.

Table 4-10 in section 4.7.4 shows the percentages of interrater agreement for the 2023 MCAS-Alt.

#### 4.4.3.6 Resolution Scoring

*Resolution scoring* refers to the rescoring of an assessment 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 assessments 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 assessment 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 assessment 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 strands and complete assessments scored, plus their 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.

## 4.5 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 2023 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 Appendix I. Item-level score distributions for each item (i.e., the percentage of students who received each score point) are provided in Appendix J. Note that the Self-Evaluation and Generalized Performance dimension scores are also included in Appendix J.

### 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 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 most items. Because the nature of alternate assessments is different from that of general assessments, and because 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.

Considering 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 | 0.76 | 0.20 | 0.37 | 0.09 |
| | 4 | 9 | 0.77 | 0.20 | 0.41 | 0.08 |
| | 5 | 9 | 0.77 | 0.20 | 0.38 | 0.08 |
| | 6 | 9 | 0.77 | 0.20 | 0.41 | 0.07 |
| | 7 | 9 | 0.77 | 0.20 | 0.42 | 0.06 |
| | 8 | 9 | 0.77 | 0.19 | 0.43 | 0.10 |
| | 10 | 9 | 0.78 | 0.19 | 0.34 | 0.10 |
| Mathematics | 3 | 6 | 0.83 | 0.19 | 0.53 | 0.17 |
| | 4 | 6 | 0.83 | 0.18 | 0.57 | 0.14 |
| | 5 | 6 | 0.83 | 0.19 | 0.59 | 0.12 |
| | 6 | 6 | 0.82 | 0.19 | 0.59 | 0.11 |
| | 7 | 6 | 0.83 | 0.19 | 0.63 | 0.06 |
| | 8 | 6 | 0.83 | 0.19 | 0.57 | 0.13 |
| | 10 | 15 | 0.83 | 0.18 | 0.34 | 0.12 |
| STE | 5 | 12 | 0.81 | 0.17 | 0.44 | 0.18 |
| | 8 | 12 | 0.80 | 0.17 | 0.44 | 0.16 |
| Biology | HS | 9 | 0.81 | 0.17 | 0.43 | 0.20 |
| Chemistry | HS | 9 | 0.86 | 0.19 | 0.34 | 0.18 |
| Introductory Physics | HS | 9 | 0.80 | 0.17 | 0.41 | 0.20 |
| Technology/Engineering | HS | 9 | 0.83 | 0.18 | 0.54 | 0.17 |

## 4.5.3 Structural Relationships Among Dimensions

By design, the achievement-level classification of the MCAS-Alt is based on three of the five scoring dimensions (Level of Complexity, Demonstration of Skills and Concepts, and Independence). As with any assessment, it is important that these dimensions be carefully examined. This was achieved by exploring the relationships among student dimension scores with Pearson correlation coefficients. A very low correlation (near zero) would indicate that the dimensions are not related; a low negative correlation (approaching -1.00) indicates that they are inversely related (i.e., that a student with a high score on one dimension had a low score on the other); and a high positive correlation (approaching 1.00) indicates that the information provided by one dimension is similar to that provided by the other dimension. The average correlations among the three dimensions by content area and grade level are shown in Table 4-8.

**Table 4-8. 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 | 0.07 | 0.17 | 0.11 | 0.04 | 0.16 | 0.05 |
| | 4 | 3 | 0.17 | 0.20 | 0.14 | 0.05 | 0.09 | 0.04 |
| | 5 | 3 | 0.15 | 0.15 | 0.11 | 0.08 | 0.15 | 0.07 |
| | 6 | 3 | 0.21 | 0.22 | 0.15 | 0.03 | 0.14 | 0.07 |
| | 7 | 3 | 0.21 | 0.32 | 0.20 | 0.04 | 0.09 | 0.05 |
| | 8 | 3 | 0.15 | 0.25 | 0.15 | 0.01 | 0.15 | 0.02 |
| | 10 | 3 | 0.06 | 0.20 | 0.15 | 0.01 | 0.13 | 0.09 |
| Mathematics | 3 | 2 | 0.07 | 0.05 | 0.10 | 0.02 | 0.04 | 0.05 |
| | 4 | 2 | 0.15 | 0.12 | 0.07 | 0.03 | 0.01 | 0.00 |
| | 5 | 2 | 0.20 | 0.12 | 0.15 | 0.04 | 0.03 | 0.00 |
| | 6 | 2 | 0.21 | 0.15 | 0.09 | 0.02 | 0.04 | 0.08 |
| | 7 | 2 | 0.19 | 0.32 | 0.16 | 0.00 | 0.02 | 0.06 |
| | 8 | 2 | 0.19 | 0.15 | 0.09 | 0.04 | 0.00 | 0.05 |
| | 10 | 5 | 0.11 | 0.12 | 0.10 | 0.10 | 0.10 | 0.10 |
| STE | 5 | 4 | 0.07 | 0.09 | -0.06 | 0.10 | 0.03 | 0.02 |
| | 8 | 4 | 0.11 | 0.13 | -0.01 | 0.05 | 0.02 | 0.05 |
| Biology | HS | 3 | 0.14 | 0.06 | -0.06 | 0.08 | 0.11 | 0.04 |
| Chemistry | HS | 3 | -0.02 | 0.26 | 0.12 | 0.01 | 0.12 | 0.22 |
| Introductory Physics | HS | 3 | 0.01 | 0.13 | -0.03 | 0.05 | 0.09 | 0.07 |
| Technology/ Engineering | HS | 3 | -0.01 | 0.05 | 0.50 | 0.06 | 0.13 | 0.27 |

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

The average correlations between every two dimensions range from very weak (absolute values between 0.00 and 0.20) to weak (absolute values between 0.20 and 0.40) for most tests. 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.

### 4.5.5 Measuring Intended Cognitive Processes

Appendix V (Summary of Alt Score Frequencies) provides the frequency of scores in each strand's rubric area by grade and content area. Note that not all grades and content areas will use all strands and scores in the table. Where not applicable, the table cell is marked as blank. Although scores tend toward the center of the rubric, this is an expected outcome for the population taking the alternate assessment. There is still the expected frequency of scores at the highest or lowest ends of the rubric when a substantial population has taken the test, indicating that the tests elicit evidence across the full expected range of rubric areas and measure the full range of intended cognitive processes.

## 4.6 MCAS-Alt Bias/Fairness

Fairness is validated through the assessment development processes, and in the development of the standards themselves, which were thoroughly vetted for bias and sensitivity. The *Resource Guides to the Massachusetts Curriculum Frameworks for Students with Disabilities* provide instructional and assessment strategies for teaching students with disabilities the same learning standards (by grade level) as general education students. The *Resource Guides* are intended to promote access to the general curriculum, as required by law, and to assist educators in planning instruction and assessment for students with the most significant cognitive disabilities. *Resource Guides* were developed by diverse 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 conduct the MCAS-Alt 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 alternate assessments necessitate teaching challenging skills based on grade-level content standards to all students. Thus, all students taking the MCAS-Alt are taught academic skills based on the standards at an appropriate level of complexity.

Issues of fairness are also addressed in the 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 assessment 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 section 4.5, 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 section 4.5, individual item characteristics of the 2023 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 2023 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
$\sigma^2_x$ represents the total test variance.

Table 4-9 presents Cronbach's $\alpha$ 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 | 856 | 39 | 28.01 | 3.65 | 0.62 | 2.24 |
| | 4 | 917 | 39 | 28.24 | 3.83 | 0.67 | 2.21 |
| | 5 | 802 | 39 | 28.62 | 3.37 | 0.61 | 2.12 |
| | 6 | 757 | 39 | 28.33 | 3.65 | 0.68 | 2.07 |
| | 7 | 719 | 39 | 28.25 | 3.74 | 0.70 | 2.04 |
| | 8 | 722 | 39 | 28.66 | 3.48 | 0.67 | 2.00 |
| | 10 | 694 | 39 | 28.16 | 4.02 | 0.61 | 2.51 |
| Mathematics | 3 | 747 | 26 | 21.08 | 1.49 | 0.55 | 1.00 |
| | 4 | 809 | 26 | 21.02 | 1.56 | 0.58 | 1.01 |
| | 5 | 733 | 26 | 21.15 | 1.46 | 0.65 | 0.86 |
| | 6 | 702 | 26 | 20.97 | 1.61 | 0.63 | 0.99 |
| | 7 | 654 | 26 | 21.05 | 1.64 | 0.65 | 0.97 |
| | 8 | 636 | 26 | 21.03 | 1.58 | 0.59 | 1.01 |
| | 10 | 676 | 39 | 30.41 | 3.64 | 0.82 | 1.56 |
| STE | 5 | 758 | 39 | 30.31 | 3.17 | 0.72 | 1.68 |
| | 8 | 695 | 39 | 30.18 | 3.09 | 0.78 | 1.46 |
| Biology | HS | 417 | 39 | 30.61 | 3.22 | 0.67 | 1.84 |
| Chemistry | HS | 114 | 39 | 32.01 | 2.43 | 0.71 | 1.31 |
| Introductory Physics* | HS | 74 | 39 | 30.26 | 3.00 | 0.67 | 1.72 |
| Technology/ Engineering* | HS | 72 | 39 | 31.23 | 3.15 | 0.82 | 1.32 |

**Due to the small sample size of the tested population, the calculations do not produce meaningful values.*

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 consider the characteristics of the testing sample (such as the variability within the sample) and the test (such as the test length). For MCAS-Alt, considering the special population and the short test length, the range of the $\alpha$ coefficient in the 2023 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 2023 MCAS-Alt. Appendix M presents reliabilities for various subgroups of interest taking MCAS-Alt. Subgroup Cronbach's $\alpha$ 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, reliability coefficients are dependent not only on the measurement properties of a test but also on the statistical distribution of the studied subgroup. For example, it can be readily seen in Appendix M that subgroup sample sizes may vary considerably, which results in natural variation in reliability coefficients. Moreover $\alpha$, 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 Achievement-Level SEM

The SEM and reliability statistics discussed in section 4.7.1 were based on various groups of interest taking MCAS-Alt. Tables M-18 through M-24 in Appendix M present SEM for populations of students analyzed by achievement level. These results show a range of SEM from 0.48–4.83, which is reasonable and relatively stable over each grade and achievement-level category, demonstrating that the precision of the MCAS-Alt is consistent across the full achievement continuum.

As above, and for the same reasons, the results documented in this section should be interpreted with caution. Limiting the analyses to individual performance levels will reduce the variability for each subgroup when compared to the whole, which would likely indicate greater measurement error estimates in comparison to the true measurement error within the group, if it were known.

## 4.7.4 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 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 1 and scorer 2. A third score was also required as a confirmation score when either scorer 1 and/or scorer 2 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 score categories, 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 H-18 through H-21 of Appendix H.

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

| Content Area | Grade | Items | Number of Score Categories | Included Scores | Percentage | | Correlation | % Third Scores |
|---|---|---|---|---|---|---|---|---|
| | | | | | Exact | Adjacent | | |
| ELA | 3 | 6 | 4 | 1,138 | 97.89 | 2.02 | 0.99 | 3.78 |
| | | 3 | 5 | 651 | 98.62 | 1.38 | 0.90 | 3.23 |
| | 4 | 6 | 4 | 990 | 97.58 | 2.22 | 0.98 | 4.04 |
| | | 3 | 5 | 567 | 99.12 | 0.88 | 0.91 | 2.29 |
| | 5 | 6 | 4 | 1,216 | 98.93 | 0.99 | 0.99 | 1.89 |
| | | 3 | 5 | 669 | 98.95 | 1.05 | 0.89 | 1.94 |
| | 6 | 6 | 4 | 1,056 | 98.86 | 0.95 | 0.99 | 1.99 |
| | | 3 | 5 | 591 | 98.82 | 1.18 | 0.92 | 2.20 |
| | 7 | 6 | 4 | 2,352 | 97.28 | 2.21 | 0.97 | 4.25 |
| | | 3 | 5 | 1,397 | 98.21 | 1.72 | 0.84 | 3.94 |
| | 8 | 6 | 4 | 998 | 98.30 | 1.60 | 0.99 | 2.71 |
| | | 3 | 5 | 546 | 98.72 | 1.28 | 0.91 | 2.01 |
| | 10 | 6 | 4 | 1,818 | 97.63 | 2.31 | 0.98 | 3.74 |
| | | 3 | 5 | 1,110 | 98.56 | 1.44 | 0.84 | 2.97 |

continued

| Content Area | Grade | Items | Number of Score Categories | Included Scores | Percentage | | Correlation | % Third Scores |
|---|---|---|---|---|---|---|---|---|
| | | | | | Exact | Adjacent | | |
| Mathematics | 3 | 4 | 4 | 722 | 99.58 | 0.42 | 0.99 | 0.55 |
| | | 2 | 5 | 440 | 98.86 | 1.14 | 0.90 | 1.14 |
| | 4 | 4 | 4 | 632 | 98.89 | 1.11 | 0.98 | 1.74 |
| | | 2 | 5 | 373 | 98.39 | 1.61 | 0.84 | 1.61 |
| | 5 | 4 | 4 | 772 | 98.58 | 1.17 | 0.96 | 2.46 |
| | | 2 | 5 | 439 | 99.32 | 0.68 | 0.93 | 0.68 |
| | 6 | 4 | 4 | 724 | 98.90 | 1.10 | 0.98 | 1.93 |
| | | 2 | 5 | 397 | 98.99 | 1.01 | 0.94 | 1.01 |
| | 7 | 4 | 4 | 1,602 | 97.63 | 1.94 | 0.92 | 3.12 |
| | | 2 | 5 | 949 | 98.84 | 1.16 | 0.89 | 1.37 |
| | 8 | 4 | 4 | 636 | 98.90 | 1.10 | 0.98 | 2.20 |
| | | 2 | 5 | 362 | 98.62 | 1.38 | 0.91 | 1.38 |
| | 10 | 10 | 4 | 1,788 | 98.38 | 1.57 | 0.97 | 2.29 |
| | | 5 | 5 | 1,103 | 99.27 | 0.73 | 0.92 | 0.82 |
| STE | 5 | 8 | 4 | 1,212 | 98.84 | 1.16 | 0.99 | 2.06 |
| | | 4 | 5 | 667 | 98.80 | 1.20 | 0.90 | 1.35 |
| | 8 | 8 | 4 | 928 | 99.25 | 0.75 | 0.99 | 1.29 |
| | | 4 | 5 | 533 | 98.87 | 1.13 | 0.93 | 1.31 |
| Biology | HS | 6 | 4 | 1,146 | 98.34 | 1.57 | 0.98 | 2.44 |
| | | 3 | 5 | 697 | 99.00 | 1.00 | 0.91 | 1.00 |
| Chemistry | HS | 6 | 4 | 266 | 99.62 | 0.38 | 0.98 | 0.38 |
| | | 3 | 5 | 157 | 100.00 | 0.00 | -- | 0.00 |
| Introductory Physics | HS | 6 | 4 | 168 | 98.81 | 1.19 | 0.98 | 2.38 |
| | | 3 | 5 | 87 | 100.00 | 0.00 | 1.00 | 0.00 |
| Technology/ Engineering | HS | 6 | 4 | 178 | 100.00 | 0.00 | 1.00 | 0.00 |
| | | 3 | 5 | 102 | 99.02 | 0.98 | 0.70 | 0.98 |

# 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 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 transitioned to next-generation achievement levels in grades 3–8 and 10, the description of each alternate academic achievement level (shown in Table 4-11) remains relatively consistent, because alternate academic achievement standards (i.e., levels) signify those students taking alternate assessments who perform well below the expectations of students taking the standard MCAS assessments. Therefore, this ensures that the meaning of students' alternate assessment scores is comparable from one year to the next. Names and descriptors for next-generation alternate and grade-level academic achievement standards are shown in Appendix W. Table 4-11 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 strand scores. In addition, achievement-level distributions for each of the last four years are provided in Appendix N.

**Table 4-11. MCAS-Alt Strand Achievement-Level Look-Up Table**

| 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 |
| 4 | 1 | 1 | 1 |

| Level of Complexity | Demonstration of Skills | Independence | Achievement Level |
|---|---|---|---|
| 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 *MCAS-Alt Feedback Form* and the *Parent/Guardian Report*.

## 4.9.2 Feedback Forms

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

## 4.9.3 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 the MCAS-Alt. 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 the report are provided: one for the parent/guardian and one to be kept in the student's school record. A sample report is 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 that 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.

### 4.9.4 Reporting Business Requirements

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

### 4.9.5 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 *2023 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

Test content validity is determined by identifying how well the assessment tasks represent the curriculum and standards for each content area and grade level. The primary evidence described in section 4.2.1 describes how the range and level of complexity of the standards being assessed have been modified to fit the needs of the MCAS-Alt testing population yet retain 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 are listed in Table 4-1, providing evidence the assessment is well-aligned to the same content standards applied to all Massachusetts students.

## 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 Validity Based on Cognitive Processes

Evidence based on cognitive processes is presented in section 4.5.5 and in Appendix V. An examination of score frequencies by content area by grade by subject shows that student scores are most common in the expected ranges for the population and that the tests measure the full range of intended cognitive processes.

## 4.10.4 Adequate Precision Across the Full Performance Continuum

Evidence indicating precision across the full performance continuum is presented in section 4.7.3 and in Appendix M. Standard errors of measurement calculated over students at each achievement level indicate that the tests provide an adequately precise estimate of student achievement across the full performance continuum.

## 4.10.5 Validity Based on Relations to Other Variables

The *Resource Guides to the Massachusetts Curriculum Framework for Students with Disabilities* (described in sections 4.1.3, 4.2.1.1, and 4.6) are used by Massachusetts educators to identify standards-based instructional goals for students. The guides also serve as the basis for the selection of the specific knowledge and skills on which the student will be assessed on the MCAS-Alt. These *Resource Guides* are developed through extensive collaboration with educators and experts. In essence, the *Resource Guides* capture the judgments of educators and experts about the curricular expectations and, as such, constitute a form of external criteria. By basing each student's assessment on the guides, the educator

implementing the MCAS-Alt brings their skills survey results and evidence collection into alignment with these judgments.

## 4.10.6 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 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.

## 4.10.7 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.

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 ensures that students with the most significant cognitive 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 important for all students in terms of post-secondary, career, and community success, 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.

Appendix X shows the report provided to parents and guardians for students assessed on the MCAS-Alt. The achievement-level descriptors provided on the first page of that report, as well as in Appendix W, describe the students' performance at each alternate academic achievement standard.

## 4.10.8 Summary

The *Standards for Educational and Psychological Testing* (2014) define validity as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (p. 11). Elaborating on that definition, the *Standards* assert that "it is the interpretations of test scores for proposed uses that are evaluated, not the test itself" (p. 11) and that "validation logically begins with an

explicit statement of the proposed interpretation of test scores, along with a rationale for the relevance of the interpretation to the proposed use" (p. 11). This definition applies specifically to intended interpretations and uses of test scores, rather than to the broader program of curriculum and instruction in which a testing program is embedded or to the surrounding education and school improvement policies and aspirations for student learning.

Further, the *Standards* state that "a sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretations of test scores for specific uses" (p. 21).

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.

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 also includes reliability and dimensionality characteristics and other evidence specific to the alternate assessment, as described in Table 4-12.

**Table 4-12. Summary of Validity Evidence for MCAS-Alt**

| Type of Validity Evidence | Section | Description of Information Provided |
|---|---|---|
| **Content-related validity evidence** | 4.2.1<br>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 |
| **Cognitive processes** | 4.5.5<br>Appendices V and W | Distributions of score frequencies indicate that the tests elicit the expected range of cognitive processes for this population |
| **Precision Over the Full Continuum** | 4.7.3<br>Appendix M | Measurement error calculated over respondent subgroups at each performance level indicate that the tests are sufficiently precise over the full performance continuum |
| **Validity Based on Other Variables** | 4.10.5, 4.1.3, 4.2.1.1, and 4.6 | Resource Guides capturing the judgments of educators and experts about the curricular expectations |
| **Reliability and classical item analyses; and subgroup statistics and scoring consistency** | 4.4, 4.7.4, and 4.8<br>Appendices H, N, R, and S | Procedures to ensure consistent scoring; interrater scoring statistics |
| | 4.5<br>Appendices I and J | Classical item statistics |
| | 4.7.1, 4.7.2, and 4.7.3<br>Appendix M | 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 |

# References


Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Wadsworth, Inc.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Prentice-Hall.

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). Marcel Dekker, Inc.

Brown, F. G. (1983). *Principles of educational and psychological testing* (3rd ed.). 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). University of Chicago Press.

Cai, L. (2012). flexMIRT: Flexible multilevel item factor analysis and test scoring [Computer software]. Vector Psychometric Group, LLC.

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

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*, 213–220.

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). 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.). John Wiley and Sons, Inc.

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.

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

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and Applications*. Kluwer Academic Publishers.

Hambleton, R. K., Swaminathan, H., & Rogers, J. H. (1991). *Fundamentals of item response theory*. Sage Publications, Inc.

Hambleton, R. K., & van der Linden, W. J. (1997). *Handbook of modern item response theory*. Springer-Verlag.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning.* Lawrence Erlbaum Associates.

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

Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education*. http://www.apa.org/science/programs/testing/fair-testing.pdf

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

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

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, *32*, 179–197.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Addison-Wesley.

Nering, M., & Ostini, R. (2010). *Handbook of polytomous item response theory models*. 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). Macmillan Publishing Company.

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

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

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). Springer-Verlag.

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

van der Linden, W. J. (2016). *Handbook of item response theory, volume one*. Chapman and Hall/CRC.

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

Wang, X., & Roussos, L. A. (2018, April). *A Simple Parametric Procedure for Detecting Drift in Anchor Items* [Conference presentation]. National Council on Measurement in Education Conference, New York, NY.

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