2021 Legacy MCAS Technical Report

9/13/22

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

cognia™

CELEBRATING 125 1895-2020 EXCELLENCE

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

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

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

www.doe.mass.edu

Table of Contents

List of Tables

CHAPTER 1. OVERVIEW

1.1 Purposes of the MCAS

This is the 2021 **Legacy** MCAS Technical Report, which is a companion to the 2021 **Next Generation** MCAS Technical report available at www.doe.mass.edu/mcas/tech/?section=techreports. While several hundred legacy tests are expected to be administered in 2022 and 2023, this will be the final edition of the annual Legacy technical report that describes the technical specifications for the first version of MCAS. While many psychometric qualities will continue, the paper-and-pencil-based version of the state tests first administered in 1998 has been replaced by online tests reported with new standards for grade-level expectations and scaled scores from 440 to 560.

The Massachusetts Comprehensive Assessment System (MCAS) was developed in response to provisions in the Massachusetts Education Reform Act of 1993, which established greater and more equitable funding to schools, accountability for student learning, and statewide standards and assessments for students, educators, schools, and districts. The Act specifies that the testing program must

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

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

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

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

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

1.2 Purpose of This Report

The purpose of this 2021 Legacy MCAS Technical Report is to document the technical quality and characteristics of the legacy MCAS operational tests that were administered in 2021: the STE tests in high school. The report presents evidence of the validity and reliability of test score interpretations, and

describes modifications made to the MCAS program in 2021. A companion document, the 2021 Next-Generation MCAS and MCAS-Alt Technical Report, provides information regarding the next-generation MCAS tests administered in 2021 in grades 3–8 and 10 ELA and mathematics and grades 5 and 8 STE.

Technical reports for previous testing years are made available by the DESE at www.doe.mass.edu/mcas/tech/?section=techreports. The previous technical reports, as well as other documents referenced in this report, provide additional background information about the MCAS program, its development, and administration.

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

1.3 Organization of This Report

This report provides detailed information regarding test design and development, scoring, and analysis and reporting of 2021 legacy MCAS results at the student, school, district, and state levels. This detailed information includes, but is not limited to, the following:

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

In addition, the appendices contain detailed item-level and summary statistics related to each 2021 legacy MCAS test and its results.

Chapter 1 of this report provides a brief overview of what is documented within the report, including updates made to the MCAS program during 2021. Chapter 2 explains the guiding philosophy, purposes, uses, components, and validity of the MCAS. Chapter 3 covers test design and development, test administration, scoring, and analysis and reporting of results for the tests, including information about characteristics of test items, how scores were calculated, the reliability of the scores, how scores were reported, and the validity of results. Numerous appendices, which appear after Chapter 3, are referenced throughout the report.

1.4 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. The 2021 MCAS administration marked a continuation of that transition.

Table 1-1 shows which MCAS tests were administered at each grade level in spring 2021 and whether the tests were next-generation (NG) or legacy (L) assessments. As the table shows, legacy MCAS tests continued to be used in 2021 for all high school STE assessments.

Table 1-1. Spring 2021 MCAS Tests Administered, by Grade Level

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

**Students may take one of four high school STE tests offered in biology, chemistry, introductory physics, and technology/engineering in grade 9 or grade 10. Results of the grades 9 and 10 tests are summarized and reported in grade 10.*

Because of the continuing transition, DESE has again, in 2021, published two separate technical reports for the MCAS assessments. This document focuses on the legacy MCAS assessments administered in high school STE and the retests in grade 10 ELA and mathematics.

The 2021 Legacy MCAS Technical Report will be the final document focused on the legacy MCAS. Legacy tests are expected to be offered for high school biology in February 2022 and for high school chemistry and technology/engineering in June 2022 and 2023, but a separate Legacy MCAS Technical Report will not be published for 2022.

Background information and technical information about the next-generation MCAS assessments is documented in the 2021 Next-Generation MCAS and MCAS-Alt Technical Report.

Chapter 2. The State Assessment System: Legacy

2.1 Guiding Philosophy

The MCAS program plays a central role in helping all stakeholders in the Commonwealth's education system—students, parents, teachers, administrators, policy leaders, and the public—understand the successes and challenges in preparing students for higher education, work, and engaged citizenship.

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

The MCAS testing program has been an important component of education reform in Massachusetts for over 15 years. The program continues to evolve with the introduction of next-generation tests.

2.2 Alignment to the Massachusetts Curriculum Frameworks

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

2.3 Uses of MCAS Results

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

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

2.4 Validity of MCAS

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

- test design and development;
- administration;
- scoring;
- technical evidence of test quality (classical item statistics, differential item functioning, item response theory statistics, reliability, dimensionality, decision accuracy and consistency); and
- reporting.

Validity information is described in detail in section 3.8 of this report.

Chapter 3. MCAS

3.1 Overview

MCAS tests have been administered to students in Massachusetts since 1998. In 1998, English language arts (ELA), mathematics, and science and technology/engineering (STE) were assessed at grades 4, 8, and 10. In subsequent years, additional grades and content areas were added to the testing program. Following the initial administration of each new test, performance standards were set.

Public school students in the graduating class of 2003 were the first students required to earn a Competency Determination (CD) in ELA and mathematics as a condition for receiving a high school diploma. Students in the class of 2010 and beyond are required to earn a CD in ELA, mathematics, and STE.

The MCAS program is managed by DESE staff with assistance and support from the assessment contractor, Cognia. Massachusetts educators play a key role in the MCAS through service on a variety of committees related to the development of MCAS test items, the development of MCAS performance level descriptors, and the setting of performance standards. The program is supported by a six-member national Technical Advisory Committee (TAC) as well as measurement specialists from the Center for Assessment and Boston College.

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

3.2 Legacy Test Design and Development

The June 2021 legacy MCAS test administration comprised high school STE end-of-course tests in biology, chemistry, introductory physics, and technology/engineering. A legacy February 2021 biology test was also administered. This test could be taken as a retest or as a first experience of MCAS STE for transfer students or students in block-scheduled science classes who completed their biology class in January.

The grade 10 ELA and mathematics retests offered to students in the classes of 2022 and prior in spring and November 2021 were also legacy tests (next-generation tests and retests were offered to the class of 2023 at the same times). These retests are typically given to students who have not yet met the CD requirements for high school graduation. However, students could take the retests for scholarship purposes for this retest administration only. Further, because students in the classes of 2020 to 2022 were granted modified CD requirements by the Board of Elementary and Secondary Education, students in these classes may not have had to pass an MCAS ELA or mathematics test or retest. More information can be found at www.doe.mass.edu/mcas/graduation.html and in Appendix A.

All legacy tests were paper-based tests; no computer-equivalent tests were available.

3.2.1 Test Specifications

3.2.1.1 Criterion-Referenced Test

Items used on the MCAS are developed specifically for Massachusetts and are aligned to Massachusetts content standards. These content standards are the basis for the reporting categories developed for each content area and are used to help guide the development of test items. The MCAS assesses only the content and skills described in the Massachusetts curriculum frameworks. All items on the STE legacy high school tests were coded to standards in the *2006 Massachusetts Science and*

Technology/Engineering Curriculum Framework. In June 2021, the biology and introductory physics tests were also coded to the standards in the *2016 Massachusetts Science and Technology/Engineering Curriculum Framework*.

The grade 10 ELA and mathematics retests were coded to standards in the 2001/2004 and 2011 ELA curriculum frameworks and the 2000/2004 and 2011 mathematics curriculum frameworks.

3.2.1.2 Legacy Item Types

The types of items used on the legacy MCAS tests, and their functions are described below.

- Multiple-choice items are used to provide breadth of coverage within a content area. Multiple-choice items make efficient use of limited testing time and allow for coverage of a wide range of knowledge and skills. Each multiple-choice item requires that students select the single best answer from four response options, and each item is aligned to one content standard. The items are machine-scored; correct responses are worth one score point, and incorrect and blank responses are assigned zero score points. Blank responses are coded to be discernable from incorrectly marked responses.

- Four-point open-response items require students to use higher-order thinking—including skills such as evaluation, analysis, and summarization—to construct satisfactory responses. Open-response items are distributed across the reporting categories. Open-response items are hand-scored by scorers trained in the specific requirements of each question. Students may receive up to four points per open-response item. Totally incorrect and blank responses receive a score of zero. Blank responses are coded to be discernable from incorrectly marked responses.

- One-point short-answer items are used as part of the mathematics retest to assess students' skills and abilities to work with brief, well-structured problems that have one or a very limited number of solutions (e.g., mathematical computations). The advantage of this type of item is that it requires students to demonstrate knowledge and skills by generating, rather than selecting, an answer. One-point short-answer items are hand-scored and assigned one point (correct) or zero points (blank or incorrect). The blanks are coded to be discernable from the incorrect responses.

- Writing prompts are administered to all students in grade 10 as part of the ELA retest. The writing assessment consists of two sessions. During the first session, students write a draft composition. In the second session, students write a final composition based on that draft.

3.2.1.3 Descriptions of Test Designs

The MCAS assessments are structured using both common and matrix items.

Common Items

Identical common items are administered to all students. Student scores are based on their performance on the common items only.

Matrix Items

The matrix portions of the STE tests are typically composed of field-test items that do not count toward student scores. In 2021, there were no field-test items, but each test form had matrix items to ensure consistency of the testing experience across all the high school STE tests.

The ELA and mathematics retests included no matrix items.

3.2.1.4 Test Design and Blueprints

High School STE

Each of the four high school STE tests focuses on one subject (biology, chemistry, introductory physics, or technology/engineering). Students in grade 9 who are enrolled in one of these subjects are eligible but not required to take the subject test in that subject. Typically, all students are required to take one of the four subject tests by the time they complete grade 10. However, the Board of Elementary and Secondary Education approved modified Competency Determination (CD) requirements due to the cancellation of testing opportunities during the COVID-19 pandemic. The modified CD for STE is in place for the classes of 2020–2023. More information about the modified CD requirements can be found here: www.doe.mass.edu/mcas/graduation.html and in Appendix A. Students were eligible to take the June and February 2021 tests for scholarship purposes.

Reporting categories for each test are listed in Table 3-1.

Table 3-1. High School STE Reporting Categories by Content Area

Reporting Categories	Approximate % of Points (+/- 5%)
Biology	
Biochemistry & Cell Biology	25
Genetics	20
Anatomy & Physiology	15
Ecology	20
Evolution & Biodiversity	20
Total	**100**
Chemistry	
Properties of Matter & Thermochemistry	25
Atomic Structure & Periodicity	25
Bonding & Reactions	30
Solutions, Equilibrium, & Acid-Base Theory	20
Total	**100**
Introductory Physics	
Motion & Forces	40
Heat & Heat Transfer	15
Waves & Radiation	25
Electromagnetism	20
Total	**100**
Technology/Engineering	
Engineering Design	20
Construction & Manufacturing	20
Fluid & Thermal Systems	30
Electrical & Communications Systems	30
Total	**100**

Tables 3-2 and 3-3 list the distribution of common and matrix items in each test.

Table 3-2. Distribution of STE Common and Matrix Items by Grade and Item Type for High School

Grade	Test	# of Forms	Positions per Form: Common MC	Common OR	Matrix MC	Matrix OR
HS	Biology	1	40	5	12	2
	Chemistry	1	40	5	20	2
	Introductory Physics	1	40	5	12	2
	Technology/Engineering	1	40	5	20	2

ELA Retest

The grade 10 ELA retest is made up of a reading comprehension portion (three sessions, each approximately 45 minutes in length) and a composition portion. There are three long passages and three short passages with a total of 52 common points. Each long passage item set includes eight multiple-choice items and one 4-point open-response item. The three short passages include a combined total of twelve multiple-choice items and one 4-point open-response item. The composition portion of the ELA retest consists of one writing prompt with a total value of 20 points (12 points for topic development and 8 points for standard English conventions). The composition score accounts for 28% of a student's total raw score for ELA.

Mathematics Retest

The grade 10 mathematics retest includes multiple-choice, short-answer, and open-response items. Short-answer items require students to perform a computation or solve a simple problem. Open-response items are more complex. Multiple-choice and short-answer items are each worth 1 point; open-response items are worth 4 points.

Table 3-3. Distribution of Retest Common and Matrix Items by Grade and Item Type for High School

Grade	Test	# of Forms	Positions per Form: Common MC	Common OR	Common SA	Common WP	Matrix MC	Matrix OR
HS	ELA	1	36	4	N/A	1	N/A	N/A
	Mathematics	1	32	6	4	N/A	N/A	N/A

3.2.1.5 Cognitive Skills for STE Tests

The high school STE test items are coded using revised Bloom's cognitive descriptions. A list of the cognitive skills can be found in Appendix B. Each item on a STE test is assigned a cognitive skill according to the cognitive demand of the item. Cognitive skills are not synonymous with difficulty. The cognitive skill describes each item based on the complexity of the mental processing a student must use to answer the item correctly. Only one cognitive skill is designated for each common item.

3.2.1.6 Use of Calculators, Formula Sheets, and Rulers

STE Tests

Formula sheets are provided to students taking the high school chemistry, introductory physics, and technology/engineering tests. These sheets contain reference information that students may need to answer certain test items. Students taking the chemistry test also receive a copy of the Periodic Table of the Elements to use during the test.

Students taking the technology/engineering test receive an MCAS ruler. The use of calculators is allowed for all the STE tests, although the high school biology tests are designed to be taken without the aid of a calculator.

Mathematics Retest

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

Reference sheets are provided to students taking the grade 10 mathematics retest. These sheets contain information, such as formulas, that students may need to answer certain items. The reference sheets are published each year.

3.2.2 Item and Test Development Process

Table 3-4 provides a high-level view of the item and test development process in chronological order.

Table 3-4. Overview of Test Development Process

Development Step	Detail of the Process
Select reading passages (for ELA only)	Contractor's content specialists find potential passages and present them to DESE for initial approval; DESE-approved passages go to Assessment Development Committees (ADCs) composed of experienced educators, and then to a Bias and Sensitivity Review Committee (Bias) for review and recommendations. ELA items are not developed until the passages have been reviewed by an ADC and Bias. With the ADC and Bias recommendations, DESE makes the final determination as to which passages will be used. (See Appendix C for committee members).
Develop items	Contractor's content specialists and subcontractors develop draft items in ELA, mathematics, and STE aligned to specific Massachusetts standards.
DESE and educator review of items	DESE content specialists review and edit items prior to presenting the items to ADCs. ADCs review items and make recommendations. Bias and Sensitivity Committee reviews items and makes recommendations. DESE test developers make final decisions based on recommendations from ADCs and Bias.
Expert review of items	Experts from higher education and practitioners review all field-test items for content accuracy. Each item is reviewed by at least two independent expert reviewers.
Benchmark open-response items and compositions	DESE and contractor content specialists meet to determine appropriate benchmark papers for training of scorers of field-tested open-response items and compositions. Scoring rubrics and notes are reviewed and edited during benchmarking meetings based on a representative sample of student responses collected during field-testing. During the scoring process, the contractor contacts DESE content specialists with any unforeseen issues.
Item statistics meeting	ADCs review field-test statistics and recommend items for the common-eligible status, for re-field-testing (with edits), or for rejection. Bias also reviews items with elevated differential item functioning (DIF) statistics and recommends items to become common-eligible or to be rejected.
Test construction	Before test construction, DESE provides target performance-level cut scores to the contractor. The contractor proposes a set of common items (items that count toward student scores) and sets of matrix items. The common set of items is delivered to DESE content specialists with proposed cut scores, including Test Characteristic Curves (TCCs) and Test Information Functions (TIFs). DESE content specialists and editorial staff review and edit the proposed common items and sets of matrix items. Contractor and DESE content specialists and editorial staff meet to review edits and changes to tests. Psychometricians provide statistical information about the effect of any proposed 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	For February and Spring 2021, no items from the high school STE tests were released. In addition, the mathematics and ELA retests are not released.

3.2.2.1 Item Development

All items used on the MCAS tests are developed specifically for Massachusetts and are directly linked to the Massachusetts curriculum frameworks. The content standards contained within the frameworks are the basis for the reporting categories developed for each content area and are used to guide the development of assessment items. See section 2.2 for specific content standard alignment.

Item Development and Review

DESE ITEM REVIEW

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

- **Alignment**: Are the items aligned to the standards? Is there a better standard to which the item could be aligned?
- **Content**: Does the item show a depth of understanding of the subject?
- **Contexts**: Are contexts used when appropriate? Are they realistic?
- **Grade-level appropriateness**: Are the content, language, and contexts appropriate for the grade level?
- **Distractors**: Have the distractors for multiple-choice items been chosen based on common sources of error? Are they plausible?
- **Mechanics**: How well are the items written? Do they follow the conventions of item writing?

DESE content specialists, in consultation with Cognia test developers, then discuss and revise the proposed item sets in preparation for Assessment Development Committee (ADC) review.

ADC ITEM REVIEW

Once DESE has reviewed new items and scoring guides and requested changes have been made, the materials are submitted to content ADCs for further review. Committees review new items using the characteristics described above and provide insight into how standards are interpreted across the state.

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

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

All ADC committee recommendations remain with the item.

BIAS AND SENSITIVITY COMMITTEE ITEM REVIEW

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

- accept
- accept with edits (including the issues they have identified and their suggested edits), or
- reject (including their reasoning).

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

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

EXTERNAL CONTENT EXPERT ITEM REVIEW

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

3.2.2.2 Item Editing

DESE content specialists review the recommendations of the ADC and Bias committees and expert reviewers and determine whether to accept the suggested edits. The items are also reviewed and edited by DESE and Cognia editors to ensure adherence to style guidelines in *The Chicago Manual of Style*, to MCAS-specific style guidelines, and to sound testing principles. According to these principles, all items should:

- demonstrate correct grammar, punctuation, usage, and spelling;
- be written in a clear, concise style;
- contain unambiguous explanations that tell students what is required to attain a maximum score;
- be written at a reading level that allows students to demonstrate their knowledge of the subject matter being tested; and
- exhibit high technical quality regarding psychometric characteristics.

3.2.2.3 Field-Testing Items

Items that have made it through the reviews listed above are approved for field-testing. Field-test items appear in the matrix portion of the test. Each item is answered by a minimum of 1,500 students (except where noted), resulting in enough responses to yield reliable performance data.

3.2.2.4 Scoring of Field-Test Items

Each field-tested multiple-choice item is machine scored. Short-answer and open-response items are hand scored. To train scorers, DESE works closely with the scoring staff to refine the rubrics and scoring

notes and to select benchmark papers that exemplify different score points and the variations within each score point. See section 3.4 for additional information on scorers and scoring.

3.2.2.5 Data Review of Field-Test Items

Data Review by DESE

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

Data Review by ADCs

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

- item difficulty (or mean score for polytomous items)
- item discrimination
- Differential Item Functioning (DIF) (see sub-groups listed below)
- distribution of scores across answer options and score points
- distribution of answer options and score points across quartiles

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

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

If an item is significantly edited after it has been field-tested, the item cannot be used in the common portion of the test until it has been field-tested again. If the ADC recommends editing an item based on the item statistics, the newly edited item returns to the field-test-eligible pool to be field-tested again.

Data Review by the Bias and Sensitivity Review Committee

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

- female/male,
- black/white,
- Hispanic/white, and

- EL and former EL who have been transitioned out of EL for fewer than two years/native English speakers and former EL who have been transitioned from EL for two or more years.

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

3.2.2.6 Item Selection and Operational Test Assembly

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

- **Content coverage/match to test design and blueprints.** The test designs and blueprints stipulate a specific number of items per item type for each content area. Item selection for the embedded field test is based on the depth of items in the existing pool of items that are eligible for the common portion of the test. Should a certain standard have few items aligned to it, then more items aligned to that standard will be field-tested to ensure a range of items aligned to that standard are available for use.

- **Item difficulty and complexity.** Item statistics drawn from the data analysis of previously field-tested items are used to ensure similar levels of difficulty and complexity from year to year as well as high-quality psychometric characteristics. Since 2011, items can be reused if they have not been released. When an item is reused in the common portion of the test, the latest usage statistics accompany that item.

- **"Clueing" items**. Items are reviewed for any information that might "clue" or help the students answer another item within the same testing session.

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

- **Key patterns.** The sequence of keys (correct answers) is reviewed to ensure that the key order appears random.

- **Option balance.** Items are balanced across forms so that each form contains a roughly equivalent number of key options (As, Bs, Cs, and Ds).

- **Page fit**. Item placement is modified to ensure the best fit and arrangement of items on any given page.

- **Facing-page issues.** For multiple-choice items associated with a stimulus (reading passages and high school biology modules) and for multiple-choice items with large graphics, consideration is given to whether those items need to begin on a left- or right-hand page and to the nature and amount of material that needs to be placed on facing pages. These considerations serve to minimize the amount of page flipping required of students.

- **Relationships among forms**. Although field-test items differ from form to form, these items must take up the same number of pages in all forms so that sessions begin on the same page in every

form. Therefore, the number of pages needed for the longest form often determines the layout of all other forms.

- **Visual appeal.** The visual accessibility of each page is always taken into consideration, including such aspects as the amount of "white space," the density of the test, and the number of graphics.

3.2.2.7 Operational Test Draft Review

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

3.2.2.8 Special Edition Test Forms

Students with Disabilities

MCAS is accessible to students with disabilities through the provision of special edition test forms and the availability of a range of accommodations for students taking the standard tests. To be eligible to receive a special edition test form, a student must have a disability that is documented either in an individualized education program (IEP) or in a 504 plan. All 2021 MCAS legacy operational tests and retests were available in the following special editions for students with disabilities:

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

- **Braille—**This form included only the common items found in the operational test. If an item indicates bias toward students with visual disabilities (e.g., if it includes a complex graphic that a student taking the Braille test could not reasonably be expected to comprehend as rendered), then simplification of the graphic is considered, with appropriate rewording of the item text, as necessary. If a graphic such as a photograph cannot be rendered in Braille, or if the graphic is not needed for the student to respond to the item, the graphic is replaced with descriptive text or a caption or eliminated altogether. Three-dimensional shapes that are rendered in two dimensions in print are rendered on the Braille test as "front view," "top view," and/or "side view," and are accompanied where necessary by a three-dimensional wooden or plastic manipulative wrapped in a Braille-labeled plastic bag.

  Modifications to original test items for the Braille version of the test are made only when necessary, as determined by the Braille test subcontractor, blind consumers, 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.

- **Electronic text reader CD**—Test versions were offered on a CD for students with disabilities who require a read-aloud function, using locally installed Kurzweil-3000 software. This edition contained only the common items found in the operational test. The items were not modified and were read aloud to the student as they appear in the standard test booklet. For items or passages that included graphics, the captions and words in the graphics were read aloud verbatim to the student. Students typically use headphones with this format but may also be tested individually in a separate setting to minimize distractions to other students (from hearing what is being read aloud).

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

Appendix D details student accommodations that did not require a special test form. After testing was completed, DESE received a list with the number of students who participated in 2021 legacy MCAS with each accommodation. No identifying information was provided (in keeping with confidentiality practices).

Spanish-Speaking Students

Spanish/English editions of the March and November mathematics retests in grade 10 were available for Spanish-speaking EL students who had been enrolled in school in the continental United States for fewer than three years and could read and write in Spanish at or near grade level. The Spanish/English editions of the mathematics retests were not made available in any other special format.

3.3 Test Administration

3.3.1 Test Administration Schedule

The legacy MCAS tests for high school STE were administered during June in spring 2021. In addition, a biology test was administered in February 2021. Because of a modification to the state's competency determination (CD) requirements for graduation, only grade 9 students and former students were eligible to participate in the February biology administration, as well as in one of the four tests administered in June (students in grades 10–12 were eligible for a modified CD, for which students completed relevant coursework in lieu of participating in MCAS testing).

The 2021 MCAS administration also included retest opportunities in ELA and mathematics for students in grades 11 and 12 and former students who exited high school and who did not previously pass one or both grade 10 tests.

Table 3-5 shows the complete 2020–2021 legacy MCAS test administration schedule. See Part III of the legacy *Principal's Administration Manual* for information about scheduling test administration, including make-up sessions for students who were absent on the day of testing.

Table 3-5. High School End-of-Course STE and Retest Test Administration Windows

Content Area	Sessions	Prescribed Test Administration Windows*	Deadline for Return of Materials to Contractor
Biology	Session 1 Session 2	February 8–26	March 1
STE (Biology, Chemistry, Introductory Physics, Technology/Engineering)	Session 1 Session 2	June 1–11	June 15
ELA Spring Retest	Composition Reading Sessions 1 & 2 Reading Session 3	May 3–June 4	June 17
Mathematics Spring Retest	Session 1 Session 2	May 3–June 4	June 17
ELA November Retest	Composition Reading Sessions 1 & 2 Reading Session 3	November 16 November 17 November 18	November 23
Mathematics November Retest	Session 1 Session 2	November 9 November 10	November 23

**In 2021, in recognition of challenges presented by Covid and health and safety protocols, testing windows were used for the high school tests for the first time.*

3.3.2 Security Requirements

Principals were responsible for ensuring that all test administrators complied with the requirements and instructions contained in the administration-specific legacy *Test Administrator's Manuals*. In addition, other administrators, educators, and staff within the school were responsible for complying with the same requirements. Schools and school staff who violated the test security requirements were subject to numerous possible sanctions and penalties, including delays in reporting of test results, the invalidation of test results, the removal of school personnel from future MCAS administrations, and possible licensure consequences for licensed educators.

Full security requirements, including details about responsibilities of principals and test administrators, examples of testing irregularities, guidance for establishing and following a document tracking system, and lists of approved and unapproved resource materials, can be found in the spring 2021 *Principal's Administration Manual,* the January[1] 2021 *Principal's Administration Manual*, the winter/spring 2021 *Principal's Administration Manual*, and the 2021 *Test Administrator's Manuals*.

3.3.3 Participation Requirements

In spring 2021, students educated with Massachusetts public funds were required by state and federal laws to participate in MCAS testing. The 1993 Massachusetts Education Reform Act mandates that **all** students in the tested grades who are educated with Massachusetts public funds participate in the MCAS, including the following groups of students:

[1] *The Department had planned a special high school administration in January 2021, which was subsequently canceled. The January manuals were then repurposed to serve as the manuals for the legacy ELA and Mathematics tests in the spring.*

- students enrolled in public schools
- students enrolled in charter schools
- students enrolled in innovation schools
- students enrolled in a Commonwealth of Massachusetts Virtual School
- students enrolled in educational collaboratives
- students enrolled in private schools receiving special education that is publicly funded by the Commonwealth, including approved and unapproved private special education schools within and outside Massachusetts
- students enrolled in institutional settings receiving educational services
- students in military families
- students in the custody of either the Department of Children and Families (DCF) or the Department of Youth Services (DYS)
- students with disabilities
- 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)

Students were eligible to participate in the 2021 high school STE tests according to the following criteria, which were posted online:

- Students in grade 9 (class of 2024) were required to take a high school STE test by the end of grade 10 and earn a passing score on one of the four STE tests to meet their CD requirement to earn a high school diploma in addition to meeting all local graduation requirements. Schools were required to make this testing opportunity available to students.
- Grade 9 students were encouraged to participate in an STE test if they were enrolled in a course that aligns to one of the four tests. Parents/guardians, in consultation with principals and other school personnel, had the option to request that their child participate in testing at grade 10 instead.
- Students in the classes of 2020–2023 were eligible for the modified CD in STE.

To certify that all eligible students were offered the opportunity to participate in testing and to certify that test security protocols were met, principals were required to complete the online Principal's Certification of Proper Test Administration (PCPA) following each test administration. For a summary of participation rates, see the 2021 MCAS Participation Report on DESE's School and District Profiles website, at profiles.doe.mass.edu/statereport/participation.aspx.

3.3.4 Administration Procedures

It was the principal's responsibility to coordinate the school's 2021 MCAS test administration. This 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 eligible high school students are given the opportunity to participate in testing
- coordinating the school's test administration schedule and ensuring that legacy tests in 2020–2021 were administered during prescribed windows
- ensuring that accommodations are properly provided and that transcriptions, if required for any accommodation, are done appropriately (Accommodation frequencies during 2021 testing can be found in Appendix E. For a list of test accommodations, see Appendix D.)
- 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 *s*pring 2021 *Principal's Administration Manual*, the winter/spring 2021 *Principal's Administration Manual*, and the January 2021 *Principal's Administration Manual*.

The MCAS program is supported by the MCAS Service Center, which includes a toll-free telephone line and email answered by staff members who provide support to schools and districts. The MCAS Service Center operates weekdays from 7:00 a.m. to 5:00 p.m. (Eastern Time), Monday through Friday.

3.4 Scoring

For paper-based tests (including all legacy tests), Cognia scanned each MCAS student answer booklet into an electronic imaging system called iScore—a secure server-to-server interface designed by Cognia. For computer-based tests (next-generation tests only), images of the student answers were transferred to iScore from the test administration platform and sorted at the item level.

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

3.4.1 Machine-Scored Items

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

3.4.2 Hand-Scored Items

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

3.4.2.1 Scoring Operations and Staff

The iScore database, its operation, and its administrative controls were all based in Dover, New Hampshire, and all MCAS item responses were scored applying a virtual scoring model maintaining the same stringent quality control measures that were applied in a center-based scoring environment.

The following staff members were involved with scoring the 2021 MCAS responses:

- The **Scoring Project Manager** oversaw communication and coordination of MCAS scoring across all scoring sites, scheduling of activities, and oversight of contractual work.
- The **iScore Operations Manager** coordinated all technical aspects associated with scoring.
- **Scoring Operations Managers** provided logistical coordination.
- **Scoring Content Specialists** ensured consistency of content area benchmarking and scoring across all grade levels and monitored the quality and accuracy of scoring.
- **Several Scoring Supervisors**, selected from a pool of experienced **Scoring Team Leaders (STLs)**, participated in training, scoring, and cleanup activities for specified content areas and grade levels. Scoring Supervisors monitored and performed read-behinds on STLs.
- **STLs**, selected from a pool of skilled and experienced scorers, monitored and performed read-behinds on scorers of their respective teams. STLs generally monitored between 5 and 11 scorers.

3.4.2.2 Benchmarking Meetings

No Legacy items underwent the benchmarking process or were discussed in benchmarking meetings. All Legacy scoring materials had been approved by DESE during prior years.

3.4.2.3 Scorer Recruitment and Qualifications

MCAS scorers, a diverse group of individuals with a wide range of backgrounds, ages, and experiences, were recruited by Cognia's HR department with support from a temporary employment agency, Kelly Services. All MCAS scorers were required to hold a four-year baccalaureate. Additionally, scorers had to have either a degree related to the content area being scored, or two classes related to the content area being scored with demonstrated experience in scoring the content area. Additional stipulations are that teachers, tutors, and administrators (e.g., principals, guidance counselors) who were under contract or

employed by or in Massachusetts schools, and people under 18 years of age, were not eligible to score MCAS responses.

Cognia verified that all leadership staff and scorers working on high school MCAS Legacy items were returning associates with demonstrated prior scoring experience on MCAS high school items in the respective content area.

3.4.2.4 Methodology for Scoring Hand-Scored Polytomous Items

The legacy MCAS tests included polytomous items requiring students to generate written responses. Polytomous items included open-response items requiring a longer or more complex response, with assigned scores of 0–4.

Scorers could assign a score-point value to a response or, if not, designate the response as one of the following:

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

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

Scorers could also flag a response as a "Crisis" response, which would be sent to scoring leadership for immediate attention. A response could be flagged as a "Crisis" response if it indicated:

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

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

Student responses were double-blind scored (each response was independently read and scored by two different scorers) for all high school operational items.

Double-Blind Scoring

In double-blind scoring, neither scorer knew whether the response had been scored before, and if it had been scored, what score it had been given. A double-blind response with discrepant scores between the two scorers (i.e., a difference greater than one point if there are three or more score points) was sent to

the arbitration queue and read by an STL or a Scoring Supervisor. For a double-blind response with adjacent scores within one point of each other, the higher score was used.

Read-Behind Scoring

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

Table 3-6 illustrates how the rules were applied for instances when the two read-behind or two double-blind scores were not an exact match.

Table 3-6. Read-Behind and Double-Blind Resolution Charts

Double-Blind Scoring* of 4-Point Item			
Scorer #1	Scorer #2	Scoring Leadership Resolution	Final
4	4	--	4
4	1	2	2
0	1	--	1
2	4	3	3
1	2	--	2
2	0	2	2

** If double-blind scores are adjacent (only 1 point different), the higher score is used as the final score. If two scores are neither exact nor adjacent, the resolution score is used as the final score.*

3.4.2.6 Scorer Training

Scoring content specialists had overall responsibility for ensuring that scorers scored responses consistently, fairly, and according to the approved scoring guidelines. Scoring materials were carefully compiled and checked for consistency and accuracy. The timing, order, and manner in which the materials were presented to scorers were planned and carefully standardized to ensure that all scorers had the same training environment and scoring experience, regardless of scoring location, content, grade level, or item scored.

Cognia used a range of methods to train scorers to score MCAS hand-scored items. The training methods were:

- live group training via Zoom;
- recording of live group training;
- pre-recorded interactive modules.

Scorers started the training process by receiving an overview of the MCAS; this general orientation included the purpose and goal of the testing program and any unique features of the test and the testing population. Scorer training for a specific item to be scored always started with a thorough review and discussion of the scoring guide, which consisted of the task, the scoring rubric, and any specific scoring notes for that task. All scoring guides were previously approved by the 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 actual student responses, many of which had been used to train scorers when the item was a field-test item:

- **Anchor sets** are DESE-approved sets consisting of three sample responses at each score point. Each response is a typical response, rather than an unusual or uncommon one; is solid, rather than controversial; and has a true score, meaning that this response has a precise score that will not be changed. Anchor sets are used to exemplify each score point.

- **Practice sets** include unusual, discussion-provoking responses, illustrating the range of responses encountered in operational scoring (e.g., exceptionally creative approaches; extremely short or disorganized responses; responses that demonstrate attributes of both higher-score anchor papers and lower-score anchor papers; and responses that show traits of multiple score points). Practice sets are used to refine the scorers' understanding of how to apply the scoring rules across a wide range of responses.

- **Qualifying sets** consist of 10 responses that are clear, typical examples of each of the score points. Qualifying sets are used to determine if scorers are able to score consistently according to the DESE-approved scoring rubric.

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

3.4.2.7 Leadership Training

Scoring content specialists also had overall responsibility for ensuring that scoring leadership (scoring supervisors and STLs) continued their history of scoring consistently, fairly, and only according to the approved scoring guidelines. Once they had completed their item-specific leadership training, scoring leadership was required to meet or surpass a qualification standard of at least 80% exact and 90% exact-plus-adjacent.

3.4.2.8 Monitoring of Scoring Quality Control

Once MCAS scorers met or exceeded the minimum standard on a qualifying set and were allowed to begin scoring, they were constantly monitored throughout the entire scoring window to ensure they scored student responses as accurately and consistently as possible. If a scorer fell below the minimum standard on any of the quality-control tools, the scorer was counseled or retrained. Scorers were required to meet or exceed the minimum standard of 70% exact and 90% exact-plus-adjacent agreement on the following:

- recalibration sets (Recals);

- embedded responses;

- read-behind scoring (RBs); and

- compilation reports, an end-of-shift report with recalibration sets and RBs.

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

begin scoring. Scorers who had zero or one out of the five exact were typically reassigned to another item or sent home for the day.

Embedded responses were approved by the scoring content specialist and loaded into iScore for blind distribution to scorers at random points during the scoring of their first 100 operational responses. While the number of embedded Committee Review Responses (CRRs) ranged from 5 to 30, depending on the item, for most items MCAS scorers received 10 of these previously scored responses during the first day of scoring an item. Scorers who fell below the 70% exact and 90% exact-plus-adjacent accuracy standard were counseled and, if approved by the STL, given extra monitoring assignments (such as additional RBs), and allowed to resume scoring.

RBs involved responses that were first read and scored by a scorer, then read and scored by an STL. STLs would, at various points during the scoring shift, command iScore to forward the next one, two, or three responses to be scored by a particular scorer. After the scorer scored each response, and without knowing the score given by the scorer, the STL would submit his or her own score to the response and then compare his or her score to the scorer's score. RBs 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 match standard were counseled, given extra monitoring assignments such as additional RBs, and allowed to resume scoring if they demonstrated the ability to meet the scoring standards after the intervention.

Double-blind scoring involved responses scored independently by two different scorers. While scorers were aware that some of the responses they scored were going to be scored by others, they had no way of knowing what responses would be scored by another scorer, or if they were the first, second, or only scorer. Double-blind scoring served as an indicator for agreement of scoring between two scorers.

Compilation reports displayed all the statistics for each scorer, including the percentage of exact, adjacent, and discrepant scores on the Recals, as well as that scorer's percentage of exact, adjacent, and discrepant scores on the RBs. As the STL conducted RBs, the scorers' overall percentages on the compilation reports were automatically calculated and updated. If the compilation report at the end of the scoring shift listed 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.

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

3.4.2.9 Interrater Consistency for Operational Items

As described above, double-blind scoring was one of the processes implemented to ensure valid and reliable hand-scoring of items and, as such, provide evidence of scoring stability. All the open-response and composition items were double-scored on the high school test.

A summary of the interrater consistency statistics for operational items is presented in Table 3-7. 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 percent exact agreement, percent adjacent agreement, correlation between the first two sets of scores, and the percent of responses that required a third score. This same information is provided at the item level in Appendix F.

Table 3-7. Summary of Interrater Consistency Statistics for Operational Items, Organized across Items by Content Area and Grade

Content Area	Grade	Number of		Percentage*		Correlation	% Third Scores
		Score Categories	Included Scores	Exact	Adjacent		
Biology	HS	5	176,194	69.45	27.37	0.87	3.17
Chemistry	HS	5	79	87.34	11.39	0.94	1.27
Introductory Physics	HS	5	56,067	65.98	30.55	0.86	3.47
Technology/Engineering	HS	5	1,752	72.15	25.17	0.83	2.68

**Values may not total 100% due to rounding.*

3.5 Classical Item Analysis

As noted in Brown (1983), "A test is only as good as the items it contains." A complete evaluation of a test's quality must include an evaluation of each item. Both *Standards for Educational and Psychological Testing* (AERA et al., 2014) and the *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 2004) include standards for identifying quality items. Items should assess only knowledge or skills that are identified as part of the domain being tested and should avoid assessing irrelevant factors. Items should also be unambiguous and free of grammatical errors, potentially insensitive content or language, and other confounding characteristics. In addition, items must not unfairly disadvantage students—in particular, racial, ethnic, or gender groups.

Both qualitative and quantitative analyses have been conducted to ensure that MCAS items meet these standards. Qualitative analyses are described in earlier sections of this chapter; this section focuses on quantitative evaluations. Statistical evaluations are presented in four parts: (1) difficulty indices, (2) item-test correlations, (3) DIF statistics, and (4) dimensionality analyses. The item analyses presented here are based on the statewide administration of the legacy MCAS in spring 2021. Note that the information presented in this section is based on the items common to all forms, since those are the items on which student scores are calculated. (Item analyses, 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 multiple-choice and open-response items are evaluated in terms of item difficulty according to standard classical test theory practices. Difficulty is defined as the average proportion of points achieved on an item and is measured by obtaining the average score on an item and dividing it by the maximum possible score for the item. Multiple-choice items are scored dichotomously (correct vs. incorrect), so, for these items, the difficulty index is simply the proportion of students who correctly answered the item. Open-response items are scored polytomously, meaning that a student can achieve scores other than just 0 or 1 (e.g., 0, 1, 2, 3, or 4 for a 4-point open-response item). By computing the difficulty index as the average proportion of points achieved, the indices for the different item types are placed on a similar scale, ranging from 0.0 to 1.0 regardless of the item type. Although this index is traditionally described as a measure of difficulty, it is properly interpreted as an easiness index, because larger values indicate easier items. An index of 0.0 indicates that all students received no credit for the item, and an index of 1.0 indicates that all students received full credit for the item (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 (0.25 for four-option multiple-choice items or essentially zero for open-response items) to 0.90, with the majority of items generally falling between 0.40 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.

A desirable characteristic of an item is for higher-ability students to perform better on the item than lower-ability students. The correlation between student performance on a single item and total test score is a commonly used measure of this 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 2021 legacy MCAS open-response items, the item discrimination index used was the Pearson product-moment correlation; for multiple-choice items, the corresponding statistic is commonly referred to as a point-biserial correlation. The theoretical range of these statistics is -1.00 to 1.00, with a typical observed range for multiple-choice items from 0.20 to 0.60.

Discrimination indices can be thought of as measures of how closely an item assesses the same knowledge and skills assessed by the other items contributing to the criterion total score on the assessment. When an item has a high discrimination index, it means that students selecting the correct response are students with higher total scores, and students selecting incorrect responses are associated with lower total scores. Given this definition, an item can discriminate between low-performing examinees and high-performing examinees. Very low or negative point-biserial coefficients computed after field-testing new items can help identify items that are flawed and should not be considered for the operational tests.

A summary of the item difficulty and item discrimination statistics for each grade and content area combination is presented in Table 3-8. Note that the statistics are presented for all items as well as by item type, multiple-choice (MC) and open-response (OR). The mean difficulty (*p*-value) and discrimination values shown in the table are within generally acceptable and expected ranges and are consistent with results obtained in previous administrations.

Table 3-8. 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
Biology	HS	ALL	45	0.68	0.13	0.42	0.13
		MC	40	0.70	0.11	0.39	0.09
		OR	5	0.48	0.09	0.69	0.05
Chemistry*	HS	ALL	45	--	--	--	--
		MC	40	--	--	--	--
		OR	5	--	--	--	--
Introductory Physics	HS	ALL	45	0.64	0.15	0.44	0.12
		MC	40	0.66	0.15	0.41	0.08
		OR	5	0.51	0.08	0.69	0.04
Technology/ Engineering	HS	ALL	45	0.55	0.17	0.34	0.13
		MC	40	0.58	0.16	0.31	0.11
		OR	5	0.35	0.07	0.53	0.09

**Item difficulty and discrimination statistics cannot be reported for Chemistry because the sample size of students was too small in 2021.*

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

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

In addition to the item difficulty and discrimination summaries presented above, these same statistics were also calculated at the item level along with item-level score point distributions. These classical statistics, item difficulty and discrimination, are provided in Appendix G for each item. On these legacy MCAS items, the item difficulty and discrimination indices are within generally acceptable and expected ranges. Very few items were answered correctly at near-chance or near-perfect rates. Similarly, the positive discrimination indices indicate that students who performed well on individual items tended to perform well overall. There are a small number of items with discrimination indices below 0.20, but none were negative. While it is acceptable to include items with low discrimination values or with very high or very low item difficulty values when their content is needed to ensure that the content specifications are appropriately covered, there were very few such cases on the 2021 legacy MCAS. Item-level score point distributions are provided for open-response items in Appendix H; for each item, the percentage of students who received each score point is presented.

3.5.2 DIF

The *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 2004) explicitly states that subgroup differences in performance should be examined when sample sizes permit and that actions should be taken to ensure that differences in performance are attributable to construct-relevant, rather than irrelevant, factors. *Standards for Educational and Psychological Testing* (AERA et al., 2014) includes similar guidelines. As part of the effort to identify such problems, psychometricians evaluated the 2021 legacy MCAS items in terms of DIF statistics.

For the 2021 legacy MCAS, the standardization DIF procedure (Dorans & Kulick, 1986) was employed to evaluate subgroup differences. (Subgroup differences denote significant group-level differences in performance for examinees with equivalent achievement levels on the test.) The standardization DIF procedure is designed to identify items for which subgroups of interest perform differently, beyond the impact of differences in overall achievement. The DIF procedure calculates the difference in item performance for two groups of students (at a time) matched for achievement on the total test. Specifically, average item performance is calculated for students at every total score. Then an overall average is calculated, weighting the total score distribution so that it is the same for the two groups.

For all content areas in high school STE, DIF statistics were calculated for all subgroups that include at least 50 students. To enable calculation of DIF statistics for the limited English proficient/formerly limited English proficient (LEP/FLEP) comparison, the minimum was set at 50 for all grade levels.

When differential performance between two groups occurs on an item (i.e., a DIF index in the “low” or “high” categories explained below), it may or may not be indicative of item bias. Course-taking patterns or differences in school curricula can lead to low or high DIF, but for construct-relevant reasons. However, if

subgroup differences in performance can be traced to differential experience (such as geographical living conditions or access to technology), the inclusion of such items is reconsidered during the item review process.

Computed DIF indices have a theoretical range from -1.0 to 1.0 for multiple-choice items, and the index is adjusted to the same scale for open-response items. Dorans and Holland (1993) suggested that index values between -0.05 and 0.05 denote negligible DIF. The majority of 2021 legacy MCAS items fell within this range. Dorans and Holland further stated that items with values between -0.10 and -0.05 and between 0.05 and 0.10 (i.e., "low" DIF) should be inspected to ensure that no possible effect is overlooked, and that items with values outside the -0.10 to 0.10 range (i.e., "high" DIF) are more unusual and should be examined very carefully before being used again operationally.[2]

For the 2021 legacy 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,
- white compared with African American/black,
- white compared with Hispanic or Latino,
- not economically disadvantaged compared with economically disadvantaged,
- not LEP-FLEP compared with LEP-FLEP[3], and
- without disabilities compared to with disabilities.

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

3.5.3 Dimensionality Analysis

Because tests are constructed with multiple content area subcategories and their associated knowledge and skills, the potential exists for a large number of dimensions being invoked beyond the common primary dimension. Generally, the subcategories are highly correlated with each other; therefore, the primary dimension they share typically explains an overwhelming majority of variance in test scores. In fact, the presence of just such a dominant primary dimension is the psychometric assumption that provides the foundation for the unidimensional item response theory (IRT) models that are used for calibrating, linking, scaling, and equating the 2021 legacy MCAS test forms.

The purpose of dimensionality analysis is to investigate whether violation of the assumption of test unidimensionality is statistically detectable and, if so, (a) the degree to which unidimensionality is violated and (b) the nature of the multidimensionality. Dimensionality analyses were performed on common items

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

[3] LEP = limited English proficient/English learners, FLEP = formerly limited English proficient/English learners who have been transitioned from EL for two or more years.

for the legacy MCAS high school Biology, Chemistry, Introductory Physics, and Technology/Engineering tests administered during spring 2021. The results for these analyses are reported below, including a comparison with the results from 2018.

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

DIMTEST is a hypothesis-testing procedure for detecting violations of local independence. The data are first randomly divided into a training sample and a cross-validation sample. Then an exploratory analysis of the conditional covariances is conducted on the training sample data to find the cluster of items that displays the greatest evidence of local dependence. The cross-validation sample is then used to test whether the conditional covariances of the selected cluster of items display local dependence, conditioning on total score on the 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 common items of the four legacy MCAS tests administered during spring 2021. The data for each grade were split into a training sample and a cross-validation sample. For high school science tests, there were over 55,000 students for biology, over 15,000 for introductory physics, over 1,800 for technology/engineering, and over 400 for chemistry. Because DIMTEST had an upper limit of 24,000 students, the training and cross-validation samples for the tests that had over 24,000 students were limited to 12,000 each, randomly sampled from the total sample. DETECT, on the other hand, had an upper limit of 500,000 students, so every training sample and cross-validation sample used all the available data. After randomly splitting the data into training and cross-validation samples, DIMTEST was applied to each dataset to see if the null hypothesis of unidimensionality would be rejected. DETECT was then applied to each dataset for which the DIMTEST null hypothesis was rejected in order to estimate the effect size of the multidimensionality.

3.5.3.1 DIMTEST Analyses

The results of the DIMTEST analyses indicated that the null hypothesis was rejected at a significance level of 0.05 for every dataset except for high school chemistry. The nonrejection for chemistry was likely due to the combined effects of the presence of weak multidimensionality (as evidenced in analyses from years prior to spring 2013) and small sample size (the sample size dropped from about 2,300 in spring 2008 to about 800 in spring 2016). Because strict unidimensionality is an idealization that almost never holds exactly for a given dataset, the statistical rejections in the DIMTEST results were not surprising. Indeed, because of the very large sample sizes (over 14,000) involved in six of the datasets, DIMTEST would be expected to be sensitive to even quite small violations of unidimensionality.

3.5.3.2 DETECT Analyses

Next, DETECT was used to estimate the effect size for the violations of local independence for all the tests. Table 3-9 below displays the multidimensionality effect-size estimates from DETECT.

Table 3-9. Multidimensionality Effect Sizes by Grade and Content Area

Content Area	Grade	Multidimensionality Effect Size		
		2018	2019	2021*
Biology	HS	0.08	0.08	0.10
Chemistry	HS	0.07	0.08	0.18
Introductory Physics	HS	0.08	0.05	0.10
Technology/Engineering	HS	0.10	0.11	0.12
Average		**0.0825**	**0.08**	**0.12**

**Testing not conducted in 2020 due to COVID-19.*

The DETECT values indicate weak multidimensionality for all the tests for the 2021 legacy MCAS forms. Also shown in Table 3-9 are the values reported in last year's dimensionality analyses. 2019's results are similar to those from this year. Although the DETECT values are slightly greater than in previous administrations, the interpretation is like previous years in that multidimensionality is generally weak.

In summary, for the 2021 dimensionality analyses, the violations of local independence, as evidenced by the DETECT effect sizes, were weak in all cases. Thus, these effects do not seem to warrant any changes in test design or scoring. In addition, the magnitude of the violations of local independence have been consistently low over the years and continue to be low despite the small increases in effect size for 2021.

3.6 MCAS IRT Scaling and Equating

The purpose of equating is to ensure that scores obtained from different forms of a test are equivalent to each other. Equating may be used if multiple test forms are administered in the same year, as well as to equate one year's forms to those given in the previous year. Equating ensures that students are not advantaged or disadvantaged because the test form they took is easier or harder than those taken by other students.

All MCAS 2021 high school STE tests used item pre-equating methodology[4] as described in Kolen and Brennan (2014). Item pre-equating allows the raw to scaled score conversion to be produced before the form is administered, which in turn allows for faster reporting and turnaround times. In item pre-equating, new forms are built from a pool of pre-existing IRT-calibrated items. Those items were calibrated in previous field-test administrations, where the field-test items were included on the same form as the operational items. The operational items were used as a set of common items for transforming the item parameters of the field-test items so that they would be on the same theta scale as the IRT-calibrated item pool. This allows for the item pool to be expanded continually.

However, with pre-equating, a number of cautions need to be taken into consideration. Kolen and Brennan (2014) state that to ensure that items behave the same on each administration the items should appear in the same contexts and positions operationally as they did non-operationally. Thus, care was taken to avoid significant shifts in position and context during the construction of the test forms.

Item parameters for the 2021 operational administration were calibrated after the 2018 MCAS operational administration. As such, no new calibrations were run for the 2021 operational items on these pre-equated tests prior to the reporting of scores. Raw score to scaled score lookups are displayed in Appendix J.

Typically, post-equating procedures were implemented after the operational administration to check the drift of the pre-equated item parameters and update them when needed. However, given that the 2021 administration is the last year of the legacy MCAS program for the four high school science tests, post-equating was not conducted.

3.6.1 IRT

All MCAS items were calibrated using IRT. IRT uses mathematical models to define a relationship between an unobserved measure of student performance, usually referred to as theta (θ), and the probability ($P(\theta)$) of getting a dichotomous item correct or of getting a particular score on a polytomous item (Hambleton, Swaminathan, & Rogers, 1991; Hambleton & Swaminathan, 1985). In IRT, it is assumed that all items are independent measures of the same construct (i.e., of the same θ). Another way to think of θ is as a mathematical representation of the latent trait of interest. Several common IRT models are used to specify the relationship between θ and $P(\theta)$ (Hambleton & van der Linden, 1997; Hambleton & Swaminathan, 1985). The process of determining the mathematical relationship between θ and $P(\theta)$ is called *item calibration*. After items are calibrated, they are defined by a set of parameters that specify a nonlinear, monotonically increasing relationship between θ and $P(\theta)$. Once the item parameters are known, an estimate of θ for each student can be calculated. This estimate, $\hat{\theta}$, is considered to be an estimate of the student's true score or a general representation of student performance. IRT has characteristics that may be preferable to those of raw scores for equating purposes because it specifically models examinee responses at the item level and facilitates equating to an IRT-based item pool (Kolen & Brennan, 2014).

For the 2021 legacy MCAS, the graded-response model (GRM) was used for polytomous items (Nering & Ostini, 2010) for all grade and content area combinations. The three-parameter logistic (3PL) model was used for dichotomous items for all grade and content area combinations except high school technology/engineering, which used the one-parameter logistic (1PL) model (Hambleton & van der

[4] Only one item in biology was post-equated because of an update in the scoring rubric after field-test administration. Post-equating was conducted by fixing the parameters of the remaining items and freely estimating the parameter for the one item.

Linden, 1997; Hambleton, Swaminathan, & Rogers, 1991). The 1PL model was chosen for high school technology/engineering because there was concern that the tests might have too few examinees to support the 3PL model in future administrations.

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

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

where
U indexes the scored response on an item,
i indexes the items,
j indexes students,
α represents item discrimination,
b represents item difficulty,
c is the pseudo guessing parameter,
θ is the student proficiency, and
D is a normalizing constant equal to 1.701.
For high school technology/engineering, this reduces to the following:

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

In the GRM for polytomous items, an item is scored in k + 1 graded categories that can be viewed as a set of k dichotomies. At each point of dichotomization (i.e., at each threshold), a two-parameter model can be used to model the probability that a student's response falls at or above a particular ordered category, given θ. This implies that a polytomous item with k + 1 categories can be characterized by k item category threshold curves (ICTCs) of the two-parameter logistic form:

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

where
U indexes the scored response on an item,
i indexes the items,
j indexes students,
k indexes threshold,
θ is the student ability,
α represents item discrimination,
b represents item difficulty,
d represents threshold, and
D is a normalizing constant equal to 1.701.

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

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

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

The GRM is also commonly expressed as:

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

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

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

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

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

3.6.2 IRT Results

The tables in Appendix K give the IRT item parameters and standard errors of all operational scoring items on the 2021 MCAS tests by grade and content area. Note that the standard errors for the parameters are equal to zero because the parameter's value was fixed in the pre-equating described above. In addition, Appendix L contains graphs of the TCCs and TIFs, which are defined below.

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

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

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

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

The TIF displays the amount of statistical information that the test provides at each value of θ_j. Information functions depict test precision across the entire latent trait continuum. There is an inverse relationship between the information of a test and its standard error of measurement (SEM). For long tests, the SEM at a given θ_j is approximately equal to the inverse of the square root of the statistical information at θ_j (Hambleton, Swaminathan, & Rogers, 1991), as follows:

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

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

3.6.3 Achievement Standards

Cutpoints for all MCAS tests were set via standard setting in 2007, establishing the theta cuts used for reporting each year. These theta cuts are presented in Table 3-10. The operational θ-metric cut scores will remain fixed throughout the assessment program unless standards are reset. Also shown in the table are the cutpoints on the reporting score scale (*2007 Standard Setting Report*).

Table 3-10. Cut Scores on the Theta Metric and Reporting Scale by Content Area and Grade

Content Area	Grade	Theta			Scaled Score				
		Cut 1	Cut 2	Cut 3	Min	Cut 1	Cut 2	Cut 3	Max
Biology	HS	-1.436	-0.554	0.686	200	220	240	260	280
Chemistry	HS	-0.134	0.425	1.150	200	220	240	260	280
Introductory Physics	HS	-0.714	0.108	1.133	200	220	240	260	280
Technology/ Engineering	HS	-0.366	0.201	1.300	200	220	240	260	280

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

3.6.4 Reported Scaled Scores

Because the θ scale used in IRT calibrations is not understood by most stakeholders, reporting scales were developed for the MCAS. The reporting scales are linear transformations of the underlying θ scale within each performance level. Student scores on the MCAS tests are reported in even-integer values from 200 to 280. Because there are four separate transformations (one for each achievement level), shown in Table 3-11, a 2-point difference between scaled scores in the *Failing* level does not mean the same thing as a 2-point difference in the *Needs Improvement* level. Because the scales differ across achievement levels, it is not appropriate to calculate means and standard deviations with scaled scores.

By providing information that is more specific about the position of a student's results, scaled scores supplement achievement level scores. Students' raw scores (i.e., total number of points) on the 2021 MCAS tests were translated to scaled scores using a data analysis process called scaling. Scaling simply converts from one scale to another. In the same way that a given temperature can be expressed on either

the Fahrenheit or Celsius scale, or the same distance can be expressed in either miles or kilometers, student scores on the 2021 MCAS tests can be expressed in raw or scaled scores.

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

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

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

where
m is the slope and
b is the intercept.

A separate linear transformation is used for each grade and content area combination and for each achievement level. Table 3-11 shows the slope and intercept terms used to calculate the scaled scores for each grade, content area, and achievement level. Note that the values in Table 3-11 will not change unless the standards are reset.

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

Table 3-11. Scaled Score Slopes and Intercepts by Content Area and Grade

Content Area	Grade	Cut Score Index	Theta Cut	Scaled Score	Slope	Intercept
Biology	HS	1	-4.000	200	0.924	200.000
		2	-3.000	212	5.393	227.745
		3	-1.436	220	22.689	252.581
		4	-0.554	240	16.126	248.942
		5	0.686	260	8.642	254.074
Chemistry	HS	1	-4.000	200	0.774	200.000
		2	-3.000	207	4.532	220.607
		3	-0.134	220	35.778	224.794
		4	0.425	240	27.586	228.276
		5	1.150	260	10.811	247.568
Introductory Physics	HS	1	-4.000	200	0.909	200.000
		2	-3.000	210	6.015	228.373
		3	-1.392	220	24.516	254.128
		4	-0.576	240	19.660	251.330
		5	0.441	260	7.815	256.554
Technology/ Engineering	HS	1	-4.000	200	0.794	200.000
		2	-3.000	201	7.370	222.697
		3	-0.366	220	35.273	232.910
		4	0.201	240	18.198	236.342
		5	1.300	260	11.765	244.706

3.7 MCAS Reliability

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

There are a number of ways to estimate an assessment's reliability. The approach that was implemented to assess the reliability of the 2021 legacy MCAS tests is the α coefficient of Cronbach (1951). This approach is most easily understood as an extension of a related procedure, the split-half reliability. In the split-half approach, a test is split in half, and students' scores on the two half-tests are correlated. To estimate the correlation between two full-length tests, the Spearman-Brown correction (Spearman, 1910; Brown, 1910) is applied. If the correlation is high, this is evidence that the items complement one another and function well as a group, suggesting that measurement error is minimal. The split-half method requires psychometricians to select items that contribute to each half-test score. This decision may have an impact on the resulting correlation since each different possible split of the test into halves will result in a different correlation. Cronbach's α eliminates the item selection by comparing individual item variances to total test variance, and it has been shown to be the average of all possible split-half correlations. Along with the split-half reliability, Cronbach's α is referred to as a coefficient of internal consistency. The term "internal" indicates that the index is measured internal to each test of interest, using data that come only from the test itself (Anastasi & Urbina, 1997). The formula for Cronbach's α is given as follows:

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

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

3.7.1 Reliability and Standard Errors of Measurement

Table 3-12 presents descriptive statistics, Cronbach's α coefficient, and raw score SEMs for each content area and grade. (Statistics are based on common items only.) The raw score SEM is calculated by the definition of reliability:

$$SEM = \sqrt{\sigma^2_x(1-\alpha)}$$

Table 3-12 shows that the reliability estimates range from 0.88 to 0.92. These estimates are within acceptable ranges and are consistent with results obtained in previous administrations of the tests.

Table 3-12. Raw Score Descriptive Statistics, Cronbach's Alpha, and SEMs by Content Area and Grade

Content Area	Grade	Number of Students	Raw Score			Alpha	SEM
			Maximum	Mean	Standard Deviation		
Biology	HS	36,383	60	37.39	12.22	0.91	3.59
Chemistry	HS	16	60	29.69	10.76	0.89	3.53
Introductory Physics	HS	11,563	60	36.40	12.71	0.92	3.59
Technology/ Engineering	HS	372	60	30.13	10.10	0.88	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 sample characteristics such as the range of individual differences in the group (i.e., variability of the sample), average ability level of the sample that took the exams, test designs, test difficulty, test length, ceiling or floor effect, and influence of guessing. Hence, "the reported reliability coefficient is only applicable to samples similar to that on which it was computed" (Anastasi & Urbina, 1997, p. 107).

3.7.2 Subgroup Reliability

The reliability coefficients discussed in the previous section were based on the overall population of students who took the 2021 legacy MCAS tests. Appendix O presents reliabilities for various subgroups of interest. Cronbach's α coefficients were calculated using the formula defined above based only on the members of the subgroup in question in the computations; values are calculated only for subgroups with 10 or more students. The reliability coefficients for subgroups range from 0.83 to 0.93 across the tests,

with a median of 0.90 and a standard deviation of 0.02, indicating that reliabilities are generally within a reasonable range.

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

3.7.3 Reporting Subcategory Reliability

Reliabilities were calculated for the reporting subcategories within the 2021 legacy MCAS content areas, which are described in section 3.2. Cronbach's α coefficients for subcategories were calculated via the same formula defined previously using just the items of a given subcategory in the computations. Results are presented in Appendix O. The reliability coefficients for the reporting subcategories range from 0.42 to 0.83, with a median of 0.69 and a standard deviation of 0.10. Because they are based on a subset of items rather than the full test, subcategory reliabilities were typically lower than were overall test score reliabilities, approximately to the degree expected based on classical test theory (Haertel, 2006), and interpretations should take this into account. Qualitative differences among grades and content areas once again preclude valid inferences about the reliability of the full test score based on statistical comparisons among subtests.

3.7.4 Reliability of Achievement Level Categorization

The accuracy and consistency of classifying students into achievement levels are critical components of a standards-based reporting framework (Livingston & Lewis, 1995). For the 2021 legacy MCAS tests, students were classified into one of four achievement levels: *Failing*, *Needs Improvement*, *Proficient*, or *Advanced*.

Cognia conducted decision accuracy and consistency (DAC) analyses to determine the statistical accuracy and consistency of the classifications. This section explains the methodologies used to assess the reliability of classification decisions and gives the results of these analyses.

Accuracy refers to the extent to which achievement classifications based on test scores match the classifications that would have been assigned if the scores did not contain any measurement error. Accuracy must be estimated because errorless test scores do not exist. Consistency measures the extent to which classifications based on test scores match the classifications based on scores from a second, parallel form of the same test. Consistency can be evaluated directly from actual responses to test items if two complete and parallel forms of the test are administered to the same group of students. In operational testing programs, however, such a design is usually impractical. Instead, techniques have been developed to estimate both the accuracy and consistency of classifications based on a single administration of a test. The Livingston and Lewis (1995) technique was used for the 2021 legacy MCAS tests because it is easily adaptable to all types of testing formats, including mixed formats.

The DAC estimates reported in Tables 3-13 and 3-14 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 2021 legacy MCAS tests, after various technical adjustments (described in Livingston & Lewis, 1995), a four-by-four contingency table of accuracy was created for each content area and grade, where cell [*i,j*] represented the estimated proportion of students whose true score fell into classification i (where i = 1 to 4) and observed score fell into classification j (where j = 1 to 4). The sum of the diagonal entries (i.e., the proportion of students whose true and observed classifications matched) signified overall accuracy.

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

Cognia also measured consistency on the 2021 legacy MCAS tests using Cohen's (1960) coefficient κ (kappa), which assesses the proportion of consistent classifications after removing the proportion of consistent classifications that would be expected by chance. It is calculated using the following formula:

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

where

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

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

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

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

3.7.5 Decision Accuracy and Consistency Results

Results of the DAC analyses described above are provided in Table 3-13. The table includes overall accuracy indices with consistency indices displayed in parentheses next to the accuracy values, as well as overall kappa values. It is important to note that the DAC results for chemistry are based on the small sample of students (N=16) who took the test in this administration. DAC calculations are not likely appropriate for use with such a small sample size, which is the likely reason for the aberrative consistency and accuracy statistics in this subject. For subjects other than chemistry, overall ranges for accuracy (0.64–0.80), consistency (0.72–0.73), and kappa (0.60) indicate that the vast majority of students were classified accurately and consistently with respect to measurement error and chance. Accuracy and consistency values conditional on achievement level are also given. For these calculations, the denominator is the proportion of students associated with a given achievement level. For example, the conditional accuracy value is 0.72 for *Needs Improvement* for grade 10 Biology. This figure indicates that among the students whose true scores placed them in this classification, 72% would be expected to be in this classification when categorized according to their observed scores. Similarly, a consistency value of 0.59 indicates that 59% of students with observed scores in the *Needs Improvement* level would be expected to score in this classification again if a second, parallel test form were taken.

For some testing situations, the greatest concern may be decisions around achievement level thresholds. For example, for tests associated with 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 *Needs Improvement/Proficient* threshold is critically important, which summarizes the percentage of students who are correctly classified either above or below the particular cutpoint. Table 3-14 provides accuracy and consistency estimates for the 2021 legacy MCAS tests at each cutpoint, as well as false positive and false negative decision rates. (A false positive is the proportion of students whose observed scores were above the cut and whose true scores were below the cut. A false negative is the proportion of students whose observed scores were below the cut and whose true scores were above the cut.)

Once again disregarding the chemistry content area, the accuracy and consistency indices at the *Needs Improvement/Proficient* threshold range from 0.83–0.96 and 0.84–0.95. The false positive and false negative decision rates at the *Needs Improvement/Proficient* threshold both range from 1–14% across all tests. These results indicate that nearly all students were correctly classified with respect to being above or below the *Needs Improvement/Proficient* cutpoints.

Table 3-13. 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			
				Failing	*Needs Improvement*	*Proficient*	*Advanced*
Biology	HS	0.80 (0.72)	0.60	0.73 (0.56)	0.72 (0.59)	0.77 (0.68)	0.82 (0.76)
Chemistry	HS	0.40 (-.013)	0.10	0.50 (0.50)	0.36 (0.27)	0.33 (0.38)	0.53 (0.50)
Introductory Physics	HS	0.80 (0.72)	0.60	0.38 (0.18)	0.80 (0.72)	0.78 (0.70)	0.81 (0.73)
Technology/ Engineering	HS	0.64 (0.73)	0.60	1.00 (1.00)	0.77 (0.68)	0.68 (0.67)	0.70 (0.87)

Table 3-14. Summary of Decision Accuracy (and Consistency) Results by Content Area and Grade—Conditional on Cutpoint

Content Area	Grade	*Failing / Needs Improvement*			*Needs Improvement / Proficient*			*Proficient / Advanced*		
		Accuracy (consistency)	False		Accuracy (consistency)	False		Accuracy (consistency)	False	
			Pos	Neg		Pos	Neg		Pos	Neg
Biology	HS	0.99 (0.98)	0.01	0.01	0.96 (0.95)	0.01	0.02	0.94 (0.91)	0.03	0.03
Chemistry	HS	0.93 (0.93)	0.06	0.02	0.80 (0.74)	0.15	0.05	0.52 (0.51)	0.31	0.17
Introductory Physics	HS	0.99 (0.98)	0.00	0.01	0.96 (0.94)	0.02	0.03	0.94 (0.92)	0.03	0.03
Technology/ Engineering	HS	1.00 (1.00)	0.00	0.00	0.83 (0.84)	0.03	0.14	0.88 (0.90)	0.02	0.10

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 small groups can be expected to be lower than those calculated based on larger groups. For groups as small as this administration's chemistry test, the results may even be improbable or out of range. For this reason, the values presented in Tables 3-13 and 3-14 should be interpreted with caution. In addition, it is important to remember that it is inappropriate to compare DAC statistics across grades and content areas.

3.7.6 Reporting of Results

The 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: *Failing*, *Needs Improvement*, *Proficient*, and *Advanced*. Students receive a separate achievement level classification in each content area. In 2021, the only legacy tests administered were for high school STE: Introductory Physics, Biology, Chemistry, and Technology/Engineering. Students in grade 9 taking only a science test received only a legacy *Parent/Guardian Report*. In grades 10 and higher, students taking science tests received reports on the redesigned *Parent/Guardian Report* template. Reports are generated at the student level. *Parent/Guardian Reports* and student results labels are printed and mailed to districts for distribution to schools. The details of the reports are presented in the sections that follow. See Appendix P for a sample *Parent/Guardian Report*.

The DESE also provides numerous reports to districts, schools, and teachers through its Edwin Analytics reporting system. Section 3.8.5 provides more information about the Edwin Analytics system, along with examples of commonly used reports.

3.7.7 Parent/Guardian Report

For students in grade 9 taking only a high school STE test, the legacy *Parent/Guardian Report* is a standalone single page (11" x 17") report with a center fold. New in 2021, the grade 9 *Parent/Guardian Report* was available online in Pearson Access Next (PAN). One black-and-white copy of the report is printed. The report is designed to present parents/guardians with a detailed summary of their child's MCAS performance. In 2021, the achievement level summary was not reported on the *Parent/Guardian Report*. The high school February biology test results and the ELA and mathematics retest results are reported as legacy tests as in the past. The *Parent/Guardian Report* for the February biology is similar to the spring grade 9 *Parent/Guardian Report*. The spring grade 9 *Parent/Guardian Report* uses the same back-page image as the *Parent/Guardian Report* for next-generation spring non-STE grades. The ELA and mathematics results for all tested grades, except the retests, are reported according to the redesigned next-generation reports. The ELA and Mathematics Retest *Parent/Guardian Report* is a different design from the spring tests. It is a combined ELA and Mathematics report. The scale on the retest uses the legacy scale of 200–280. In 2021, a special administration of the ELA and Mathematics High School Legacy test was administered in the spring. The *Parent/Guardian Report* for this administration followed the design used for the November 2019 administration.

The front page of the *Parent/Guardian Report* for students with results for only a high school STE test provides student identification information, including student name, grade, birth date, Student ID (SASID), school name, and district name. The front page also presents the Commissioner's letter to parents/guardians, general information about the test, and website information for parent/guardian resources. The inside of the report contains the achievement level, scaled score, and standard error of the scaled score for the science test taken by the student. If the student does not receive a scaled score, the reason is displayed under the heading "Achievement Level." Information concerning the student's performance on individual test questions, a sub-content area summary for the content area, and a note

stating whether a student has met the graduation requirement for science also appear on the inside of the report.

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

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

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

3.7.8 Analysis and Reporting Business Requirements

To ensure that MCAS results are processed and reported accurately, a document defining analysis and reporting business requirements is prepared each year. The 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 students to be excluded from school-, district-, and state-level summary computations, if applicable. The *Analysis and Reporting Business Requirements* document is included in Appendix Q.

3.7.9 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 dataset, the first step is to verify the accuracy of the data. Once report designs have been approved by the DESE, reports are run using demonstration data to test the application of the analysis and reporting business requirements. These reports are then approved by the DESE.

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

The third aspect of quality control involves procedures to check the accuracy of reported data. Using a sample of schools and districts, the quality-assurance group verifies that the reported information is

correct. The selection of sample schools and districts for this purpose is very specific because it can affect the success of the quality-control efforts. There are two sets of samples selected that may not be mutually exclusive. The first set includes samples that satisfy all the following criteria:

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

The second set of samples includes districts or schools that have unique reporting situations that require the implementation of an analysis and reporting business requirement. This set is necessary to ensure that each requirement is applied correctly.

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

3.8 MCAS Validity

One purpose of this report is to describe the technical and reporting aspects of the MCAS program that support valid score interpretations. According to the *Standards for Educational and Psychological Testing* (AERA et al., 2014), considerations regarding establishing intended uses and interpretations of test results—and conforming to these uses—are of paramount importance 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, this technical report provides a comprehensive presentation of validity evidence associated with the MCAS program.

3.8.1 Test Content Validity Evidence

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

3.8.2 Response Process Validity Evidence

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

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

3.8.3 Internal Structure Validity Evidence

Evidence of test validity based on internal structure is presented in great detail in the discussions of item analyses, scaling, equating, and reliability in sections 3.5 through 3.7. Technical characteristics of the internal structure of the assessments are presented in terms of classical item statistics (item difficulty, item-test correlation), DIF analyses, dimensionality analyses, reliability, SEM, and IRT parameters and procedures. Each test is equated to the previous year's test in that grade and content area to preserve the meaning of scores over time. In general, item difficulty and discrimination indices were within acceptable and expected ranges. Very few items were answered correctly at near-chance or near-perfect rates. Similarly, the positive discrimination indices indicate that most items were assessing consistent constructs, and students who performed well on individual items tended to perform well overall. See the individual sections for more complete results of the different analyses.

3.8.4 Validity Evidence in Relationships to Other Variables

Massachusetts has accumulated a substantial amount of evidence of the criterion-related validity of the MCAS tests. This evidence shows that MCAS test results are correlated strongly with relevant measures of academic achievement.

3.8.5 Efforts to Support the Valid Use of MCAS Data

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

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

MCAS results and student growth percentiles are used as two categories of information in the DESE's accountability formulas for schools and districts.[5] The accountability formulas also consider the following variables when making accountability determinations for schools and districts: the rate of assessment participation, graduation rates (for high schools and districts), and student demographic group. Information on the state's accountability system is available on the DESE website at: www.doe.mass.edu/accountability/.

As documented on the accountability web page above, the DESE carefully weighs all available evidence prior to rendering accountability decisions for schools and districts. No school, for instance, is placed in

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

Level 4 or 5 without an agency-wide review of data, including (but not limited to) four years of assessment data. Assignment to a lower accountability level comes with increased involvement between the DESE and the local education agencies (LEAs). The different levels of engagement are explained in the State's System of Support, presented here: www.doe.mass.edu/accountability/. Among the supports, districts with schools in Level 3 get assistance with data analysis from one of the six regional District and School Assistance Centers (DSACs). The supports for LEAs in Levels 4 and 5 and documented outcomes associated with these supports are available here: www.doe.mass.edu/turnaround/howitworks/.

2. Providing information to support program evaluation at the school and district levels
3. Determining whether high school students have demonstrated the knowledge and skills required to earn a Competency Determination (CD)—one requirement for earning a high school diploma in Massachusetts

No student can be reported as a high school graduate in Massachusetts without first earning a CD. The typical path to earning a CD is to pass three MCAS high school exams—an ELA exam, a mathematics exam, and one of four STE exams. Most examinees in the state (around 90%, in a typical year) score *Needs Improvement* or higher on all three exams on their first try.[6] Examinees who have not earned a CD are given many opportunities to retake the exams during the retest and spring test administrations, with no limit to reexaminations. Examinees who are not awarded a CD may also appeal the decision. The DESE has instituted a rigorous appeals process that can afford some examinees the opportunity to demonstrate their competency on the state standards through the successful completion of high school course work. (Additional information on the appeals process can be found at www.doe.mass.edu/mcasappeals/.) Finally, students with significant disabilities who are unable to take the MCAS exams can participate in the MCAS-Alt program, which allows students to submit a portfolio of work that demonstrates their proficiency on the state standards.

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

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

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

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

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

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

An example of the *MCAS School Results by Standards Report* is shown in Figure 3-1. This report 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 shows the difference between the school and state columns, which allows educators to compare their school results to the state results. Filters provide educators with the capacity to compare student results across nine demographic categories, which include gender, race/ethnicity, economically disadvantaged status, and special education status.

Figure 3-1. Example of School Results by Standards Report—Mathematics, Grade 7

All Students **Students :** (161)

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

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

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

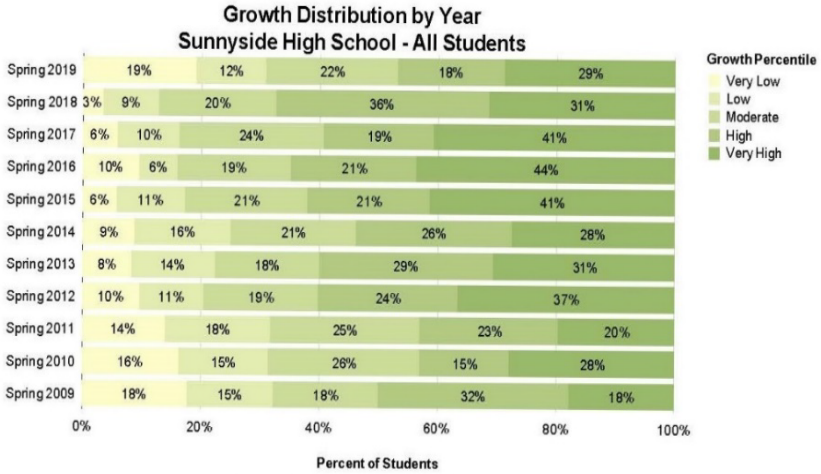
School results only include students enrolled in the school since Oct.1.

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

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

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

Figure 3-2. 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.

References

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

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

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

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

Brown, F. G. (1983). *Principles of educational and psychological testing* (3rd ed.). Fort Worth, TX: Holt, Rinehart and Winston.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Nering, M., & Ostini, R. (2010). *Handbook of polytomous item response theory models*. New York, NY: Routledge.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 221–262). New York, NY: Macmillan Publishing Company.

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

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

Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika 52,* 589–617.

Stout, W. F., Froelich, A. G., & Gao, F. (2001). Using resampling methods to produce an improved DIMTEST procedure. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on Item Response Theory* (pp. 357–375). New York, NY: Springer-Verlag.

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

Appendices