# Rhode Island and Vermont Multi-State Science Assessment

# 2018–2019

# Volume 1: Annual Technical Report

RIDE Rhode Island Department of Education

VERMONT AGENCY OF EDUCATION

# TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES



## LIST OF APPENDICES

# 1. INTRODUCTION

The Rhode Island and Vermont Multi-State Science Assessment (MSSA) is an assessment for grades 5, 8, and 11. The *2018–2019 MSSA Technical Report* is provided to document and make transparent all methods used in item development, test construction, psychometrics, standard setting, test administration, and score reporting, including summaries of student results, and evidence and support for intended uses and interpretations of the test scores. The technical report is reported as six separate, self-contained volumes:

1) **Annual Technical Report.** This volume is updated each year and provides a global overview of the tests administered to students each year.
2) **Test Development.** This volume summarizes the procedures used to construct test forms and provides summaries of the item bank and development process.
3) **Standard Setting.** This volume documents the methods and results of the MSSA standard-setting process.
4) **Evidence of Reliability and Validity.** This volume provides technical summaries of the test quality and special studies to support the intended uses and interpretations of the test scores.
5) **Test Administration.** This volume describes the methods used to administer all tests, security protocols, and modifications or accommodations available.
6) **Score Interpretation Guide.** This volume describes the score types reported and details the appropriate inferences that can be drawn from each score reported.

The Rhode Island Department of Education (RIDE) and the Vermont Agency of Education (VT AOE) communicates the quality of the MSSA by making these technical reports accessible to the public on the state's website.

## 1.1 BACKGROUND AND HISTORICAL CONTEXT OF TESTS

Rhode Island and Vermont adopted the Next Generation Science Standards (NGSS) in 2013. The RIDE, the VT AOE, and their assessment vendor, the American Institutes for Research (AIR), developed and administered new online assessments to measure the new standards. These new assessments, the Rhode Island Next Generation Science Assessment (RI NGSA) and the Vermont Science Assessment (VTSA), were developed jointly by both states and measure the science knowledge and skills of Rhode Island and Vermont students in grades 5, 8, and 11. In 2017–2018, the assessment was administered as an independent field test in Rhode Island and as an operational field test in Vermont. The MSSA was administered operationally for the first time in both states in 2018–2019.

The RIDE provides an overview of the RI NGSA at https://www.ride.ri.gov/InstructionAssessment/Assessment/NGSAAssessment.aspx and at https://ri.portal.airast.org/get-started/test-administration-guidance.stml.

The VT AOE provides an overview of the VTSA at https://education.vermont.gov/student-learning/assessments/state-and-local-assessments/science and also at https://vt.portal.airast.org/resources/vermont-science-assessment/.

Information about the NGSS is available at: www.nextgenscience.org.

In the remainder of this volume, the term Multi-State Science Assessment (MSSA) will refer to the Rhode Island Next Generation Science Assessment (RI NGSA) and the Vermont Science Assessment (VTSA) combined, unless stated explicitly otherwise.

## 1.2 Purpose and Intended Uses of the Multi-State Science Assessment

The MSSA is a criterion-referenced test established using principles of evidence-centered design to yield overall and discipline-level test scores at the student level and other levels of aggregation that reflect student achievement of the MSSA. The Next Generation Science Standards (NGSS) establish a set of knowledge and skills that all students need to have to be prepared for a wide range of high-quality post-secondary opportunities, including higher education and the workplace.

The NGSS reflect the latest research and advances in modern science and differ from previous science standards in multiple ways. First, rather than describe general knowledge and skills that students should know and be able to do, they describe specific performances that demonstrate what students know and can do. The NGSS refers to such performed knowledge and skills as *performance expectations*. Second, while unidimensionality is a typical goal of standards (and the assessments that measure them), the NGSS are intentionally multi-dimensional. Each performance expectation incorporates all three dimensions from the NGSS Framework—a science or engineering practice, a disciplinary core idea, and a crosscutting concept. Third, while traditional standards do not consider other subject areas, the NGSS connects to other subjects like the Common Core mathematics and English language arts (ELA) standards. Another unique feature of the NGSS is the assumption that students should learn all science disciplines rather than a select few, as is traditionally the expectation in many high schools, where students may elect, for example, to take biology and chemistry but not physics or astronomy.

The MSSA supports instruction and student learning by providing valuable feedback to educators and parents, which can be used to form instructional strategies to remediate or enrich instruction. An array of reporting metrics is provided so that achievement can be evaluated at the student level and at aggregate levels and to monitor improvement at the student and group levels over time.

The MSSA draws items from an item bank that consists of AIRCore items, and items owned by several other states that share a memorandum of understanding (MOU) to share content, leadership, and new ideas and methods. Full members of the MOU in 2019 were Connecticut, Hawaii, Idaho, Oregon, Rhode Island, Utah, Vermont, West Virginia, and Wyoming. AIR has a supporting and coordinating role. New Hampshire observes and participates in some activities. AIR, the RIDE, and the VT AOE worked together to ensure that the items in the test forms constructed for all grades within the states uniquely measure NGSS.

Table 1 outlines the required uses and citations for the MSSA based on the §18-2E-5-(d)(3) and the federal Every Student Succeeds Act (ESSA) plan. The MSSA fulfills all the requirements described in Table 1.

*Table 1. Required Uses and Citations for the MSSA*

| Required Use | Required Use Citation |
|---|---|
| Indicator of academic achievement and progress | ESSA Plan Section 1 A. i; ESSA Plan Section 4 4.1 A |
| Test administration frequency and grade levels | 15.1-21-08.1 |
| Compilation of test scores | 15.1-21-09 |
| Publication of test scores | 15.1-21-10 |
| Requirement for alignment of test to academic content standards | 15.1-21-11 |

## 1.3 Participants in the Development and Analysis of the Multi-State Science Assessment

The Rhode Island Department of Education (RIDE) and the Vermont Agency of Education (VT AOE) manage the Rhode Island and Vermont state assessment programs with the assistance of several participants, including Rhode Island and Vermont educators, a Technical Advisory Committee (TAC), and vendors. The RIDE and the VT AOE fulfill the diverse requirements of implementing Rhode Island's and Vermont's statewide assessments while meeting or exceeding the guidelines established in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014).

### 1.3.1 Rhode Island Department of Education and Vermont Agency of Education

The Office of Instruction, Assessment & Curriculum in the RIDE and Office of Assessment in the VT AOE manage test development, administration, scoring, and reporting of results for the statewide comprehensive assessment programs, including coordinating with other RIDE and VT AOE offices, Rhode Island and Vermont public schools, and vendors.

### 1.3.2 Rhode Island and Vermont Educators

Rhode Island and Vermont educators participate in most aspects of the conceptualization and development of the MSSA. Educators participate in the development of the academic standards, the clarification of how these standards are assessed, the test design, and the review of test questions and passages.

### 1.3.3 Technical Advisory Committee

The RIDE and the VT AOE convene an advisory committee panel several times a year to discuss psychometric, test development, administrative, and policy issues of relevance to current and future Rhode Island and Vermont assessments. This committee is composed of several nationally recognized assessment experts and highly experienced practitioners from several school districts.

### 1.3.4 American Institutes for Research

The American Institutes for Research (AIR) is the vendor that was selected through the state-mandated competitive procurement process. AIR is responsible for developing test content, building test forms, conducting psychometric analyses, administering and scoring test forms, and reporting test results for the MSSA described in this report. Additionally, AIR is responsible for developing and maintaining the AIRCore item bank.

### 1.3.5 Caveon Test Security

Caveon Test Security monitored web pages and social media during the spring 2019 test administration to ensure that any secure testing materials such as items and prompts were not leaked.

## 1.4 AVAILABLE TEST FORMATS AND SPECIAL VERSIONS

The MSSA is administered online using a linear-on-the-fly (LOFT) test design. Science items are centered on a scientific phenomenon. They can consist of shorter (stand-alone) items or items with several parts (item clusters) requiring the student to interact with the item in various ways. In Rhode Island, the assessment was administered as an independent field test in spring 2018 and as an operational test in spring 2019. In Vermont, the assessment was administered as an operational field test in spring 2018, and as an operational test in spring 2019. In 2020 and onwards, additional items will be field tested to build upon the item bank.

Students unable to participate in the online administration have the option to use print-on-demand—a feature that provides the same items administered to students online in a paper format. Spanish versions of the MSSA (developed to meet the same content standards as the English versions) are available for all tested grades. Students participating in the computer based MSSA can use standard online testing features in the test delivery system (TDS), which include a selection of font color and size and the ability to zoom in and zoom out or highlight text. In addition to the resources available to all students, options are available to accommodate students with an Individualized Education Program (IEP) or Section 504 Plan. These include braille, American Sign Language (ASL), closed captioning, and large print. Students with disabilities have the option to take the MSSA with or without accommodations or to take an alternate assessment.

## 1.5 STUDENT PARTICIPATION

All students in Rhode Island and Vermont public schools are required to participate in the statewide assessments. The MSSA is administered in the spring.

Table 2 shows the number of students who were tested (number tested) and the number of students whose scores were included for the analyses in this technical report (number reported), while Table 3 and Table 4 show the number of students who were tested in Rhode Island and Vermont, respectively. Table 5 shows the demographic characteristics of the student population, in counts and in percentages, in the spring administration of the 2018–2019 assessments. Table 6 shows the demographic characteristics for Rhode Island students, and Table 7 shows the demographic characteristics for Vermont students. The subgroups reported here are gender, ethnicity, students with limited English proficiency (LEP), students with economic disadvantages, and students qualified for special education.

*Table 2. Total Number of Students Participating in the MSSA, Spring 2019*

| Grade | Number Tested | Number Reported |
|---|---|---|
| 5 | 16,904 | 16,867 |
| 8 | 16,514 | 16,482 |
| 11 | 15,186 | 15,156 |

*Table 3. Number of Rhode Island Students Participating in the MSSA, Spring 2019*

| Grade | Number Tested | Number Reported |
|---|---|---|
| 5 | 10,816 | 10,798 |
| 8 | 10,553 | 10,546 |
| 11 | 9,748 | 9,733 |

*Table 4. Number of Vermont Students Participating in the MSSA, Spring 2019*

| Grade | Number Tested | Number Reported |
|---|---|---|
| 5 | 6,088 | 6,069 |
| 8 | 5,961 | 5,936 |
| 11 | 5,438 | 5,423 |

*Table 5. Combined Distribution of Demographic Characteristics of Student Population*

| Group | Grade 5 | | Grade 8 | | Grade 11 | |
|---|---|---|---|---|---|---|
| | *N* | % | *N* | % | *N* | % |
| All Students | 16,867 | 100 | 16,482 | 100 | 15,156 | 100 |
| Female | 8,317 | 49.31 | 8,061 | 48.91 | 7,400 | 48.83 |
| Male | 8,547 | 50.67 | 8,420 | 51.09 | 7,753 | 51.15 |
| African American | 1,102 | 6.53 | 1,074 | 6.52 | 960 | 6.33 |
| American Indian/Native Alaskan | 90 | 0.53 | 100 | 0.61 | 95.00 | 0.63 |
| Asian | 496 | 2.94 | 425 | 2.58 | 464 | 3.06 |

| Group | Grade 5 | | Grade 8 | | Grade 11 | |
|---|---|---|---|---|---|---|
| | *N* | % | *N* | % | *N* | % |
| Hispanic | 3,128 | 18.55 | 2,842 | 17.24 | 2,458 | 16.22 |
| Multi-Racial | 663 | 3.93 | 593 | 3.60 | 402 | 2.65 |
| Pacific Islander | 23 | 0.14 | 25 | 0.15 | 21 | 0.14 |
| White | 11,365 | 67.38 | 11,423 | 69.31 | 10,756 | 70.97 |
| Limited English Proficiency | 1,384 | 8.21 | 898 | 5.45 | 868 | 5.73 |
| Special Education | 2,654 | 15.73 | 2,549 | 15.47 | 1,791 | 11.82 |
| Economically Disadvantaged | 2,460 | 14.58 | 2,135 | 12.95 | 1,618 | 10.68 |

*Table 6. Distribution of Demographic Characteristics of Rhode Island Student Population*

| Group | Grade 5 | | Grade 8 | | Grade 11 | |
|---|---|---|---|---|---|---|
| | *N* | % | *N* | % | *N* | % |
| All Students | 10,798 | 100 | 10,546 | 100 | 9,733 | 100 |
| Female | 5,349 | 49.54 | 5,215 | 49.45 | 4,748 | 48.78 |
| Male | 5,449 | 50.46 | 5,331 | 50.55 | 4,985 | 51.22 |
| African American | 938 | 8.69 | 920 | 8.72 | 832 | 8.55 |
| American Indian/Native Alaskan | 78 | 0.72 | 78 | 0.74 | 79 | 0.81 |
| Asian | 397 | 3.68 | 291 | 2.76 | 317 | 3.26 |
| Hispanic | 2,986 | 27.65 | 2,721 | 25.80 | 2,328 | 23.92 |
| Multi-Racial | 494 | 4.57 | 422 | 4.00 | 310 | 3.19 |
| Pacific Islander | 8 | 0.07 | 14 | 0.13 | 15 | 0.15 |
| White | 5,897 | 54.61 | 6,100 | 57.84 | 5,852 | 60.13 |
| Limited English Proficiency | 1,239 | 11.47 | 826 | 7.83 | 792 | 8.14 |
| Special Education | 1,457 | 13.49 | 1,491 | 14.14 | 1,071 | 11.00 |
| Economically Disadvantaged | – | – | – | – | – | – |

*Table 7. Distribution of Demographic Characteristics of Vermont Student Population*

| Group | Grade 5 | | Grade 8 | | Grade 11 | |
|---|---|---|---|---|---|---|
| | *N* | % | *N* | % | *N* | % |
| All Students | 6,069 | 100 | 5,936 | 100 | 5,423 | 100 |
| Female | 2,968 | 48.90 | 2,846 | 47.94 | 2,652 | 48.90 |

| Group | Grade 5 | | Grade 8 | | Grade 11 | |
|---|---|---|---|---|---|---|
| | *N* | *%* | *N* | *%* | *N* | *%* |
| Male | 3,098 | 51.05 | 3,089 | 52.04 | 2,768 | 51.04 |
| African American | 164 | 2.70 | 154 | 2.59 | 128 | 2.36 |
| American Indian/Native Alaskan | 12 | 0.20 | 22 | 0.37 | 16 | 0.30 |
| Asian | 99 | 1.63 | 134 | 2.26 | 147 | 2.71 |
| Hispanic | 142 | 2.34 | 121 | 2.04 | 130 | 2.40 |
| Multi-Racial | 169 | 2.78 | 171 | 2.88 | 92 | 1.70 |
| Pacific Islander | 15 | 0.25 | 11 | 0.19 | 6 | 0.11 |
| White | 5,468 | 90.10 | 5,323 | 89.67 | 4,904 | 90.43 |
| Limited English Proficiency | 145 | 2.39 | 72 | 1.21 | 76 | 1.40 |
| Special Education | 1,197 | 19.72 | 1,058 | 17.82 | 720 | 13.28 |
| Economically Disadvantaged | 2,460 | 40.53 | 2,135 | 35.97 | 1,618 | 29.84 |

# 2. Summary of Operational Procedures

## 2.1 Test Administration

Table 8 shows the testing window for the 2018–2019 Multi-State Science Assessment (MSSA) in Rhode Island and Vermont.

*Table 8. MSSA Testing Windows by State*

| State | Grades | Testing Window |
|---|---|---|
| Rhode Island | 5, 8, 11 | April 22, 2019–May 24, 2019 |
| Vermont | 5, 8, 11 | April 15, 2019–May 31, 2019 |

The key personnel involved with the Rhode Island and Vermont test administration included the District Test Coordinators (DCs), School Test Coordinators (SCs), and test administrators (TAs) who proctored the test. Test Administration Manuals were provided so that personnel involved with the statewide assessment administrations could maintain both standardized administration conditions and test security.

The AIR Secure Browser developed by AIR was required to access the online Rhode Island and Vermont tests. The online browser provided a secure environment for student testing by disabling the hot keys, copy, and screen capture capabilities and preventing access to the desktop (Internet, email, and other files or programs installed on school machines). During the online assessment, students could pause a test, review previously answered questions, and modify their response if the test had not been paused for more than 20 minutes. Students do not have a required time limit for each test session, but schools are given approximate time estimates for how long each test may take for most students for test administration planning purpose.

## 2.2 Simulations

Before the operational testing window opens, AIR employs a simulation approach. Simulations are performed for all MSSA tests. AIR delivers the MSSA under a linear-on-the-fly (LOFT) test design. The test is delivered using the same item selection algorithm that AIR uses to deliver adaptive tests, except that only the blueprint of a test is considered during the item-selection process. Simulations were carried out to configure the algorithm settings and to evaluate whether individual tests adhered to the test blueprint and monitor item exposure rates. The simulation approaches and results are discussed in Volume 2 of this technical report.

## 2.3 Designated Supports and Accommodations

The accessibility supports discussed in this document include embedded (digitally provided) and non-embedded (non-digitally or locally provided) universal features that are available to all students as they access instructional or assessment content; designated features that are available to those students for whom the need has been identified by an informed educator or team of educators; and accommodations that are generally available for students for whom there is documentation on an Individualized Education Program (IEP) or Section 504 Plan. For English learners (ELs), Spanish language versions of the MSSA are available.

Scores achieved by students using designated supports are included for federal accountability purposes. All educators making these decisions were trained on the process and understand the range of designated supports available.

Accommodations are changes in procedures or materials that ensure equitable access to instructional and assessment content and generate valid assessment results for students who need them. Embedded accommodations (e.g., text-to-speech) are provided digitally through instructional or assessment technology, and non-embedded designated features (e.g., scribe) are non-digital. State-approved accommodations do not compromise the learning expectations, constructs, or grade-level standards. Such accommodations help students with a documented need generate valid assessment outcomes so that they can fully demonstrate what they know and are able to do. From the psychometric point of view, the purpose of providing accommodations is to "increase the validity of inferences about students with disabilities by offsetting specific disability-related, construct-irrelevant impediments to performance" (Koretz & Hamilton, 2006, p. 562).

The TAs and SCs in Rhode Island and Vermont are responsible for ensuring that arrangements for accommodations are made before the test administration dates. The available accommodation options for eligible students are listed on the follow page. Descriptions for each of these accommodations can be found in Volume 5 of this technical report.

Table 9 through Table 12 list the number of testing sessions in which a student was provided with each designated support or accommodation during the spring 2019 test administration sessions in Rhode Island and Vermont, respectively.

*Table 9. Number of Testing Sessions with Accessibility Features, Rhode Island*

| Accessibility Features | Grade | | |
|---|---|---|---|
| | 5 | 8 | 11 |
| **Embedded** | | | |
| Color Choices | – | 4 | – |
| Masking | 365 | 123 | 14 |
| Mouse Pointer | 3 | – | 1 |
| Print Size | 8 | 3 | 1 |
| Text-to-Speech: Stimuli and Items | 1,544 | 569 | 56 |
| **Non-Embedded** | | | |
| Magnification | – | 1 | 1 |

*Table 10. Number of Testing Sessions with Allowed Accommodations, Rhode Island*

| Accommodations | Grade | | |
|---|---|---|---|
| | 5 | 8 | 11 |
| **Non-Embedded** | | | |
| AT/ACC Devices (Requires Permissive Mode) | – | 1 | – |
| Speech-to-Text (Requires Permissive Mode) | 18 | 25 | – |

*Table 11. Number of Testing Sessions with Allowed Designated Supports, Vermont*

| Designated Supports | Grade | | |
|---|---|---|---|
| | 5 | 8 | 11 |
| **Embedded** | | | |
| Color Choices | 52 | 3 | 15 |
| Masking | 4,993 | 4,963 | 5,145 |
| Mouse Pointer | 10 | – | 1 |
| Print Size | 14 | 8 | 6 |
| Text-to-Speech: Stimuli and Items | 1,161 | 836 | 336 |
| **Non-Embedded** | | | |
| Amplification | 4 | 1 | 1 |
| Bilingual Dictionary | 2 | 1 | |
| Color Contrast | 5 | 7 | 6 |
| Color Overlay | 7 | 1 | 2 |

| Designated Supports | Grade | | |
|---|---|---|---|
| | **5** | **8** | **11** |
| Magnification | 12 | 2 | 4 |
| Noise Buffer | 43 | 13 | 8 |
| Read Aloud Items | 287 | 120 | 81 |
| Read-Aloud Items - Spanish | 2 | 5 | 1 |
| Read-Aloud Stimuli | 249 | 106 | 79 |
| Read-Aloud Stimuli - Spanish | 2 | – | – |
| Scribe Items (Non-Writing) | 175 | 120 | 10 |
| Separate Setting | 689 | 597 | 410 |
| Simplified Test Directions | 135 | 78 | 42 |

*Table 12. Number of Testing Sessions with Allowed Accommodations, Vermont*

| Accommodations | Grade | | |
|---|---|---|---|
| | **5** | **8** | **11** |
| **Embedded** | | | |
| Permissive Mode | 107 | 87 | 41 |
| Streamlined Mode | 128 | 92 | 64 |
| **Non-Embedded** | | | |
| Alternate Response Options (Requires Permissive Mode) | 2 | 2 | 9 |
| Large Print | – | 2 | 1 |
| Paper Test Booklet | 2 | 2 | 3 |
| Read-Aloud Passages | 74 | 19 | 36 |
| Scribe Items (Writing) | 53 | 15 | 21 |
| Speech-to-Text (Requires Permissive Mode) | 89 | 70 | 29 |
| Word Prediction | 34 | 10 | 2 |

# 3. ITEM BANK AND TEST DESIGN

## 3.1 ITEM BANK

The American Institutes for Research (AIR) works with a group of states to develop science assessments to assess the Next Generation Science Standards (NGSS) and other standards influenced by the same science framework. Many of these states have signed a memorandum of understanding (MOU) to share item specifications and items. AIR has coordinated this group of states and holds contracts to develop and deliver the items for most of them.

AIR also built the AIRCore science item bank in partnership with these states. These AIR-owned items make up a substantial part of the item bank and are shared with partner states. Rhode Island and Vermont signed the MOU, and therefore, the item pool available for Rhode Island and Vermont includes items from three sources:

- Items owned by Rhode Island and Vermont (referred throughout as MSSA items)
- Items shared by other MOU states
- Items shared from the AIRCore item bank

All these items follow the same specifications, test development processes, and review processes. In 2018, AIR field tested more than 540 item clusters and stand-alone items, of which 451 (including items from all sources) were accepted and made available as operational items in 2019. In 2019, 347 item clusters and stand-alone items were field tested, of which 268 have passed rubric validation and item data review.

The science item bank is used for operational accountability tests in seven states (in 2019), including Rhode Island and Vermont. An additional three state tests will become operational in 2020, and four other states are scheduled to become operational in 2021.

AIR's process for developing and field testing science items is detailed in Volume 2 of this technical report. Here, note that best practices have been implemented at every turn:

- The goals, uses, and claims that the test would be designed to support were identified in a collaborative meeting over August 22 and 23, 2016, as an attempt to facilitate the transition from NGSS content standards to statewide summative assessments for science. AIR invited content and assessment leaders from 10 states (most of them participating in the MOU) as well as four nationally recognized experts who helped co-author the NGSS standards. Two nationally recognized psychometricians also participated.
- AIR staff and participating states collaborated to develop items and test specifications. The item specifications generally were accompanied by sample items meeting those specifications. All specifications and sample items were reviewed by state content experts and committees of educators in at least one state.
- Items were reviewed by science experts in at least one state.
- Every item has been reviewed by a content advisory committee (composed of state educators) in at least one state, or in a cross-state educator review process.
- Every item has been reviewed by a committee of educators charged with evaluating language accessibility, bias, and sensitivity in at least one state or a cross-state educator review.
- Every item was field tested, and items with questionable data were re-reviewed by committees of educators.

## 3.2 Field Testing

All items that were part of the 2019 operational pool were field tested in 2018, as described in Section 3.2.1. Additional items were field tested in 2019, described in Section 3.2.2.

### 3.2.1 2018 Field Test

In 2018, a large pool of items was field tested in nine states. For three states (Hawaii, Oregon, and Wyoming), unscored field-test items were added as an additional segment to the operational (scored) legacy science test. Two other states (Connecticut and Rhode Island) conducted an independent field test in which all students participated and were administered a full set of items, but no scores were reported. In the remaining four states (New Hampshire, Utah, Vermont, and West Virginia), an operational field test was administered, meaning tests consisted of field-test items, but items became operational and were scored after the test administration if they were not rejected during rubric validation or item data review. In total, 340 item clusters and 205 stand-alone items were administered in the elementary, middle, and high school grade bands. Table 13 presents the number of item clusters and stand-alone items administered in each grade for each state.

*Table 13. Number of Item Clusters and Stand-Alone Items Administered in Spring 2018*

| Grade Band and Item Type | CT | HI | MSSA | NH | OR | UT | WV | WY | Entire Bank |
|---|---|---|---|---|---|---|---|---|---|
| **Elementary School** | **135** | **24** | **69** | **58** | **26** | – | **91** | **14** | **153** |
| Cluster | 78 | 13 | 40 | 34 | 20 | – | 56 | 6 | 86 |
| Stand-Alone | 57 | 11 | 29 | 24 | 6 | – | 35 | 8 | 67 |
| **Middle School** | **174** | **27** | **56** | **55** | **28** | **98** | **123** | **17** | **241** |
| Cluster | 115 | 13 | 26 | 30 | 22 | 98 | 90 | 5 | 171 |
| Stand-Alone | 59 | 14 | 30 | 25 | 6 | – | 33 | 12 | 70 |
| **High School** | **149** | **23** | **75** | **60** | **38** | **–** | **–** | **14** | **151** |
| Cluster | 81 | 14 | 34 | 33 | 30 | – | – | 6 | 83 |
| Stand-Alone | 68 | 9 | 41 | 27 | 8 | – | – | 8 | 68 |
| **Total** | **458** | **74** | **200** | **173** | **92** | **98** | **214** | **45** | **545** |

For the states with a separate field-test segment (states with a legacy science test) and one of the states with an operational field test (Utah), fixed field-test forms were constructed (using a balanced incomplete design except for Utah) and spiraled across students. For the independent and operational field tests (except for Utah), including Rhode Island and Vermont, items were administered using a linear-on-the-fly (LOFT) test design. The difference between the test design for the independent field tests and operational field tests depended on the test blueprint. For the independent field tests, the only blueprint constraint imposed was that students received four stand-alone items and two cluster items for each of the three science disciplines, whereas a full blueprint was implemented for the states with an operational field test. The blueprint for the MSSA is discussed in Section 3.3.

For any given state, a minimum sample size of 1,500 students per item was targeted. Most items were administered in two or more states so that the item pools for all individual states were linked through common items. Table 14 through Table 16 present the number of item clusters and stand-alone items that were in common between the item pools of any two states. The numbers below the diagonal represent the numbers for all the field-test items, and the numbers above the diagonal represent the number of common items at the time of the 2018 calibration. The shaded diagonal elements represent the number of items that were administered only in the given state (in parentheses, the number of unique items at the time of calibration). Table 14 presents the results for elementary school, Table 15 presents the results for middle school, and Table 16 presents the results for high school. The numbers at field testing are slightly different from the numbers at calibration for a variety of reasons, such as items being rejected during rubric validation, and versioning issues for some items in some states.

*Table 14. Number of Common Elementary School Items Administered and Calibrated in Spring 2018*

| | State | Connecticut | Hawaii | MSSA | New Hampshire | Oregon | Utah | West Virginia | Wyoming |
|---|---|---|---|---|---|---|---|---|---|
| **Cluster** | **CT** | **3 (3)** | 9 | 36 | 28 | 16 | 0 | 49 | 6 |
| | **HI** | 10 | **0 (0)** | 7 | 8 | 5 | 0 | 12 | 1 |
| | **MSSA** | 36 | 8 | **0 (2)** | 15 | 12 | 0 | 26 | 2 |
| | **NH** | 30 | 8 | 17 | **1 (3)** | 5 | 0 | 22 | 2 |
| | **OR** | 17 | 5 | 13 | 5 | **1 (1)** | 0 | 5 | 1 |
| | **UT** | 0 | 0 | 0 | 0 | 0 | **0 (0)** | 0 | 0 |
| | **WV** | 49 | 12 | 27 | 25 | 5 | 0 | **0 (4)** | 2 |
| | **WY** | 6 | 1 | 2 | 2 | 1 | 0 | 2 | **0 (0)** |
| **Stand-Alone** | **CT** | **1 (3)** | 5 | 25 | 22 | 2 | 0 | 33 | 7 |
| | **HI** | 5 | **6 (6)** | 0 | 0 | 0 | 0 | 4 | 0 |
| | **MSSA** | 26 | 0 | **0 (1)** | 10 | 4 | 0 | 13 | 3 |
| | **NH** | 24 | 0 | 11 | **0 (2)** | 0 | 0 | 15 | 2 |
| | **OR** | 2 | 0 | 4 | 0 | **1 (1)** | 0 | 0 | 0 |
| | **UT** | 0 | 0 | 0 | 0 | 0 | **0 (0)** | 0 | 0 |
| | **WV** | 35 | 4 | 14 | 17 | 0 | 0 | **0 (2)** | 1 |
| | **WY** | 8 | 0 | 3 | 3 | 0 | 0 | 2 | **0 (1)** |
| **Grade Band Total** | **CT** | **4 (6)** | 14 | 61 | 50 | 18 | 0 | 82 | 13 |
| | **HI** | 15 | **6 (6)** | 7 | 8 | 5 | 0 | 16 | 1 |
| | **MSSA** | 62 | 8 | **0 (3)** | 25 | 16 | 0 | 39 | 5 |
| | **NH** | 54 | 8 | 28 | **1 (5)** | 5 | 0 | 37 | 4 |
| | **OR** | 19 | 5 | 17 | 5 | **2 (2)** | 0 | 5 | 1 |
| | **UT** | 0 | 0 | 0 | 0 | 0 | **0 (0)** | 0 | 0 |
| | **WV** | 84 | 16 | 41 | 42 | 5 | 0 | **0 (6)** | 3 |
| | **WY** | 14 | 1 | 5 | 5 | 1 | 0 | 4 | **0 (1)** |

*Table 15. Number of Common Middle School Items Administered and Calibrated, Spring 2018*

| | State | Connecticut | Hawaii | MSSA | New Hampshire | Oregon | Utah | West Virginia | Wyoming |
|---|---|---|---|---|---|---|---|---|---|
| Cluster | CT | **2 (6)** | 12 | 22 | 26 | 19 | 44 | 77 | 5 |
| | HI | 11 | **1 (0)** | 3 | 6 | 6 | 0 | 9 | 1 |
| | MSSA | 23 | 3 | **0 (1)** | 9 | 1 | 7 | 22 | 2 |
| | NH | 26 | 6 | 10 | **1 (2)** | 7 | 0 | 17 | 3 |
| | OR | 19 | 6 | 1 | 7 | **2 (2)** | 0 | 5 | 1 |
| | UT | 48 | 0 | 7 | 0 | 0 | **48 (52)** | 43 | 0 |
| | WV | 83 | 10 | 21 | 18 | 6 | 48 | **1 (9)** | 2 |
| | WY | 5 | 1 | 2 | 3 | 1 | 0 | 2 | **0 (0)** |
| Stand-Alone | CT | **2 (3)** | 6 | 27 | 25 | 3 | 0 | 33 | 12 |
| | HI | 6 | **8 (8)** | 2 | 0 | 0 | 0 | 2 | 0 |
| | MSSA | 27 | 2 | **0 (0)** | 18 | 3 | 0 | 20 | 2 |
| | NH | 25 | 0 | 18 | **0 (0)** | 0 | 0 | 21 | 3 |
| | OR | 3 | 0 | 3 | 0 | **0 (0)** | 0 | 0 | 0 |
| | State | 0 | 0 | 0 | 0 | 0 | **0 (0)** | 0 | 0 |
| | WV | 33 | 2 | 20 | 21 | 0 | 0 | **0 (0)** | 2 |
| | WY | 12 | 0 | 2 | 3 | 0 | 0 | 2 | **0 (0)** |
| Grade Band Total | CT | **4 (9)** | 18 | 49 | 51 | 22 | 44 | 110 | 17 |
| | HI | 17 | **9 (8)** | 5 | 6 | 6 | 0 | 11 | 1 |
| | MSSA | 50 | 5 | **0 (1)** | 27 | 4 | 7 | 42 | 4 |
| | NH | 51 | 6 | 28 | **1 (2)** | 7 | 0 | 38 | 6 |
| | OR | 22 | 6 | 4 | 7 | **2 (2)** | 0 | 5 | 1 |
| | UT | 48 | 0 | 7 | 0 | 0 | **48 (52)** | 43 | 0 |
| | WV | 116 | 12 | 41 | 39 | 6 | 48 | **1 (9)** | 4 |
| | WY | 17 | 1 | 4 | 6 | 1 | 0 | 4 | **0 (0)** |

*Table 16. Number of Common High School Items Administered and Calibrated, Spring 2018*

| | State | Connecticut | Hawaii | MSSA | New Hampshire | Oregon | Utah | West Virginia | Wyoming |
|---|---|---|---|---|---|---|---|---|---|
| Cluster | CT | **10 (16)** | 13 | 30 | 29 | 30 | 0 | 0 | 5 |
| | HI | 13 | **0 (0)** | 7 | 7 | 8 | 0 | 0 | 1 |
| | MSSA | 32 | 7 | **0 (2)** | 13 | 12 | 0 | 0 | 1 |
| | NH | 32 | 7 | 14 | **0 (3)** | 12 | 0 | 0 | 3 |
| | OR | 30 | 8 | 12 | 12 | **0 (0)** | 0 | 0 | 1 |
| | UT | 0 | 0 | 0 | 0 | 0 | **0 (0)** | 0 | 0 |
| | WV | 0 | 0 | 0 | 0 | 0 | 0 | **0 (0)** | 0 |
| | WY | 6 | 1 | 1 | 3 | 1 | 0 | 0 | **0 (1)** |
| Stand-Alone | CT | **4 (4)** | 9 | 40 | 27 | 8 | 0 | 0 | 8 |
| | HI | 9 | **0 (0)** | 4 | 0 | 0 | 0 | 0 | 0 |
| | MSSA | 39 | 4 | **0 (1)** | 20 | 3 | 0 | 0 | 1 |
| | NH | 25 | 0 | 20 | **0 (0)** | 0 | 0 | 0 | 1 |
| | OR | 8 | 0 | 3 | 0 | **0 (0)** | 0 | 0 | 0 |
| | UT | 0 | 0 | 0 | 0 | 0 | **0 (0)** | 0 | 0 |
| | WV | 0 | 0 | 0 | 0 | 0 | 0 | **0 (0)** | 0 |
| | WY | 7 | 0 | 1 | 1 | 0 | 0 | 0 | **0 (0)** |
| Grade Band Total | CT | **14 (20)** | 22 | 70 | 56 | 38 | 0 | 0 | 13 |
| | HI | 22 | **0 (0)** | 11 | 7 | 8 | 0 | 0 | 1 |
| | MSSA | 71 | 11 | **0 (3)** | 33 | 15 | 0 | 0 | 2 |
| | NH | 57 | 7 | 34 | **0 (3)** | 12 | 0 | 0 | 4 |
| | OR | 38 | 8 | 15 | 12 | **0 (0)** | 0 | 0 | 1 |
| | UT | 0 | 0 | 0 | 0 | 0 | **0 (0)** | 0 | 0 |
| | WV | 0 | 0 | 0 | 0 | 0 | 0 | **0 (0)** | 0 |
| | WY | 13 | 1 | 2 | 4 | 1 | 0 | 0 | **0 (1)** |

The common item design was used to calibrate all the items on a common NGSS scale. The calibration model is explained in detail in Section 5 of this volume of the technical report.

Following the (operational) field test, items went through a substantial validation process. The process begins with rubric validation. Rubric validation is a process in which a committee of state educators reviews student responses and the proposed scoring of those responses. The responses reviewed are scientifically sampled to overrepresent responses most likely to have been mis-scored. Specifically, the sample overrepresents: (1) low-scored responses from otherwise high-scoring students; and (2) high-scored responses from otherwise low-scoring students.

During rubric validation, educators recommend revisions to rubrics where necessary. AIR staff revise the rubrics and rescore the entire sample to ensure that the rubric changes have all and only the intended effects.

Following rubric validation, classical item statistics were computed for the scoring assertions, including item difficulty and item discrimination statistics, testing time, and differential item functioning (DIF) statistics. The states establish standards for the statistics. Any items violating these standards are flagged for a second educator review. Even though the scoring assertions were the basic units of analysis to compute classical item statistics, the business rules to flag items for another educator review were established at the item level because assertions cannot be reviewed in isolation. A common set of business rules was defined for all the states participating in the (operational) field test, although some states decided to include additional items for data review. The item statistics were computed on the student data of the students testing in the state that owned the item. For Rhode Island and Vermont, which share their item development, the statistics were computed on the combined data. For AIRCore items, the data from Connecticut, New Hampshire, Rhode Island, Vermont, and West Virginia (states that used AIRCore items and with either an independent or operational field test) were combined. For each state, a data review committee consisting of educators (science teachers) and supported by AIR content experts reviewed the items that were owned by the state and flagged for data review according to the established business rules. For AIRCore, cross-state review committees were established. Table 17 presents the number of items field tested in Rhode Island and Vermont, the number of items that were rejected before or during rubric validation, the number of items that were sent out for data review, and the number of items that were rejected during data review.

*Table 17. Overview of Science Administration, Rubric Validation, Item Data Review, Spring 2018*

| Grade Band and Owner | Number of Items Field Tested | | | Number of Items Rejected Before/During Rubric Validation | | | Number of Items Sent to Data Review | | | Number of Items Rejected at Data Review[b] | | | Number of Items Remaining | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Cluster | Stand-Alone | Total | Cluster | Stand-Alone | Total | Cluster | Stand-Alone | Total | Cluster | Stand-Alone | Total | Cluster | Stand-Alone | Total |
| **Elementary School** | **86** | **67** | **153** | **3** | **0** | **3** | **23** | **41** | **64** | **5** | **8** | **13** | **76** | **59** | **135** |
| AIRCore | 34 | 31 | 65 | 0 | 0 | 0 | 7 | 19 | 26 | 1 | 2 | 3 | 32 | 29 | 61 |
| Administered in MSSA | 18 | 17 | 35 | 0 | 0 | 0 | 4 | 9 | 13 | 1 | 0 | 1 | 16 | 17 | 33 |
| Other MOU States[a] | 47 | 31 | 78 | 3 | 0 | 3 | 15 | 19 | 34 | 4 | 3 | 7 | 39 | 28 | 67 |
| Administered in MSSA | 17 | 7 | 24 | 1 | 0 | 1 | 7 | 4 | 11 | 2 | 1 | 3 | 13 | 6 | 19 |
| MSSA | 5 | 5 | 10 | 0 | 0 | 0 | 1 | 3 | 4 | 0 | 3 | 3 | 5 | 2 | 7 |
| **Middle School** | **171** | **70** | **241** | **12** | **4** | **16** | **67** | **36** | **103** | **16** | **9** | **25** | **138** | **57** | **195** |
| AIRCore | 31 | 28 | 59 | 0 | 0 | 0 | 11 | 15 | 26 | 1 | 2 | 3 | 25 | 26 | 51 |
| Administered in MSSA | 11 | 21 | 32 | 0 | 0 | 0 | 2 | 11 | 13 | 0 | 1 | 1 | 8 | 20 | 28 |
| Other MOU States[a] | 137 | 40 | 177 | 12 | 4 | 16 | 56 | 21 | 77 | 15 | 7 | 22 | 110 | 29 | 139 |
| Administered in MSSA | 12 | 7 | 19 | 0 | 0 | 0 | 3 | 5 | 8 | 1 | 1 | 2 | 11 | 6 | 17 |
| MSSA | 3 | 2 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 2 | 5 |
| **High School** | **83** | **68** | **151** | **8** | **2** | **10** | **35** | **44** | **79** | **4** | **9** | **13** | **66** | **57** | **123** |
| AIRCore | 35 | 29 | 64 | 3 | 0 | 3 | 15 | 17 | 32 | 0 | 2 | 2 | 28 | 27 | 55 |
| Administered in MSSA | 14 | 22 | 36 | 1 | 0 | 1 | 6 | 13 | 19 | 0 | 2 | 2 | 12 | 20 | 32 |
| Other MOU States[a] | 43 | 32 | 75 | 3 | 2 | 5 | 19 | 23 | 42 | 4 | 4 | 8 | 36 | 26 | 62 |
| Administered in MSSA | 15 | 12 | 27 | 2 | 0 | 2 | 4 | 9 | 13 | 1 | 2 | 3 | 12 | 10 | 22 |
| MSSA | 5 | 7 | 12 | 2 | 0 | 2 | 1 | 4 | 5 | 0 | 3 | 3 | 2 | 4 | 6 |
| **Total** | **340** | **205** | **545** | **23** | **6** | **29** | **125** | **121** | **246** | **25** | **26** | **51** | **280** | **173** | **453** |

*Note.* [a]Other MOU states include Connecticut, Hawaii, Oregon, Utah, West Virginia, and Wyoming.
[b]Including three clusters rejected after item data review

Table 18 summarizes the item pool that was used in Rhode Island and Vermont for each of three science disciplines.

*Table 18. Overview of Items Field Tested in Spring 2018*

| Grade Band and Item Type | Items Field Tested in Spring 2018 | | | |
|---|---|---|---|---|
| | Total | Earth and Space Sciences | Life Sciences | Physical Sciences |
| **Elementary School** | **69** | **19** | **25** | **25** |
| Cluster | 40 | 9 | 15 | 16 |
| Stand-Alone | 29 | 10 | 10 | 9 |
| **Middle School** | **56** | **20** | **19** | **17** |
| Cluster | 26 | 9 | 9 | 8 |
| Stand-Alone | 30 | 11 | 10 | 9 |
| **High School** | **75** | **24** | **25** | **26** |
| Cluster | 34 | 9 | 13 | 12 |
| Stand-Alone | 41 | 15 | 12 | 14 |
| **Total** | **200** | **63** | **69** | **68** |

### 3.2.2 2019 Field Test

In 2019, a second wave of items was field tested in nine states. For three states (Hawaii, Idaho elementary school, and Wyoming), unscored field-test items were added as a separate segment to the operational (scored) legacy science test. An independent field test in which students were administered a full set of items was conducted for a sample of Idaho middle schools. In the remaining six states (Connecticut, New Hampshire, Oregon, Rhode Island, Vermont, and West Virginia), field-test items were administered as unscored items embedded among the operational items. In total, 123 item clusters and 224 stand-alone items were administered as field-test items in the elementary, middle, and high school grade bands. Table 19 presents the number of field-tested item clusters and stand-alone items administered in each grade for each state. The numbers in parentheses in the column representing MSSA present the number of items owned by MSSA.

*Table 19. Number of Field-Tested Items Administered in Spring 2019*

| Grade Band and Item Type | CT | HI | ID | MSSA | NH | OR | WV | WY | Entire Bank |
|---|---|---|---|---|---|---|---|---|---|
| **Elementary School** | **47** | **31** | **53** | **42 (10)** | **18** | **27** | **18** | **16** | **117** |
| Cluster | 18 | 19 | 20 | 17 (4) | 0 | 16 | 10 | 5 | 50 |
| Stand-Alone | 29 | 12 | 33 | 25 (6) | 18 | 11 | 8 | 11 | 67 |
| **Middle School** | **56** | **23** | **53** | **46 (8)** | **28** | **26** | **26** | **15** | **127** |
| Cluster | 14 | 9 | 17 | 10 (3) | 4 | 9 | 8 | 5 | 38 |
| Stand-Alone | 42 | 14 | 36 | 36 (5) | 24 | 17 | 18 | 10 | 89 |

| Grade Band and Item Type | CT | HI | ID | MSSA | NH | OR | WV | WY | Entire Bank |
|---|---|---|---|---|---|---|---|---|---|
| **High School** | **69** | **21** | **–** | **37 (6)** | **29** | **28** | **–** | **25** | **103** |
| Cluster | 25 | 14 | – | 18 (3) | 2 | 13 | – | 2 | 35 |
| Stand-Alone | 44 | 7 | – | 19 (3) | 27 | 15 | – | 23 | 68 |
| **Total** | **172** | **75** | **106** | **125 (24)** | **75** | **81** | **44** | **56** | **347** |

*Note.* MSSA-owned items are indicated in the parentheses.

For the three states with a separate field-test segment (states with a legacy science test), field-test forms were constructed using a balanced incomplete design and spiraled across students. For the independent field test, items were administered under a linear-on-the-fly (LOFT) design, where the only blueprint constraint imposed was that students received four stand-alone items and two cluster items for each of the three science disciplines. For the states with an operational test, field-test items were embedded within the operational test. Some of the states with an operational test (New Hampshire, Rhode Island, Vermont) opted for a test in which operational items were grouped by science discipline. For these three states, the field-test items were presented together in a fourth group of items. The sequence of the four sets of items (corresponding to the three disciplines and a set of field-test items) was randomized across students. Other states opted for a test design in which the items were not grouped by discipline (Connecticut, Oregon, West Virginia). In these three states, field-test items were administered at random positions throughout the test. A student received either a field-test item cluster or a set of five field-test stand-alone items. The test design for the MSSA is discussed in Section 3.3.

For any given state, a minimum sample size of 1,500 students per field-test item was targeted. Most items were administered in two or more states. Table 20 to Table 22 present the number of cluster items and stand-alone items that were shared between the field-test pools of any two states. The numbers below the diagonal represent the numbers for all the field-test items, and the numbers above the diagonal represent the number of common field-test items at the time of calibration. The shaded diagonal elements represent the number of field-test items that were administered only in the given state (in parentheses, the number of unique field-test items at the time of calibration). Table 20 presents the results for elementary schools, Table 21 presents the results for middle schools, and Table 22 presents the results for high schools. The numbers at field testing are slightly different from the numbers at calibration because some items were rejected during rubric validation.

*Table 20. Number of Common Elementary School Field-Test Items Administered and Calibrated in Spring 2019*

| | State | Connecticut | Hawaii | Idaho | MSSA | New Hampshire | Oregon | West Virginia | Wyoming |
|---|---|---|---|---|---|---|---|---|---|
| Cluster | CT | **2 (2)** | 2 | 10 | 3 | 0 | 2 | 1 | 4 |
| | HI | 2 | **0 (0)** | 3 | 8 | 0 | 14 | 2 | 0 |
| | ID | 10 | 3 | **4 (4)** | 0 | 0 | 1 | 3 | 3 |
| | MSSA | 3 | 8 | 0 | **3 (3)** | 0 | 9 | 4 | 1 |
| | NH | 0 | 0 | 0 | 0 | **0 (0)** | 0 | 0 | 0 |
| | OR | 2 | 14 | 1 | 9 | 0 | **1 (1)** | 0 | 0 |
| | WV | 1 | 2 | 3 | 4 | 0 | 0 | **1 (0)** | 1 |
| | WY | 4 | 0 | 3 | 1 | 0 | 0 | 1 | **0 (0)** |
| Stand-Alone | CT | **5 (5)** | 1 | 13 | 1 | 9 | 0 | 0 | 2 |
| | HI | 1 | **0 (0)** | 10 | 6 | 0 | 6 | 0 | 0 |
| | ID | 13 | 11 | **1 (1)** | 12 | 1 | 9 | 2 | 4 |
| | MSSA | 1 | 7 | 13 | **3 (3)** | 5 | 8 | 5 | 6 |
| | NH | 9 | 0 | 1 | 5 | **2 (3)** | 0 | 0 | 6 |
| | OR | 0 | 7 | 10 | 9 | 0 | **1 (1)** | 0 | 0 |
| | WV | 0 | 0 | 2 | 5 | 0 | 0 | **1 (1)** | 0 |
| | WY | 2 | 0 | 4 | 6 | 7 | 0 | 0 | **0 (0)** |
| Grade Band Total | CT | **7 (7)** | 3 | 23 | 4 | 9 | 2 | 1 | 6 |
| | HI | 3 | **0 (0)** | 13 | 14 | 0 | 20 | 2 | 0 |
| | ID | 23 | 14 | **5 (5)** | 12 | 1 | 10 | 5 | 7 |
| | MSSA | 4 | 15 | 13 | **6 (6)** | 5 | 17 | 9 | 7 |
| | NH | 9 | 0 | 1 | 5 | **2 (3)** | 0 | 0 | 6 |
| | OR | 2 | 21 | 11 | 18 | 0 | **2 (2)** | 0 | 0 |
| | WV | 1 | 2 | 5 | 9 | 0 | 0 | **2 (1)** | 1 |
| | WY | 6 | 0 | 7 | 7 | 7 | 0 | 1 | **0 (0)** |

*Table 21. Number of Common Middle School Field-Test Items Administered and Calibrated in Spring 2019*

| | State | Connecticut | Hawaii | Idaho | MSSA | New Hampshire | Oregon | West Virginia | Wyoming |
|---|---|---|---|---|---|---|---|---|---|
| Cluster | CT | **5 (5)** | 3 | 4 | 2 | 0 | 2 | 1 | 0 |
| | HI | 3 | **0 (0)** | 4 | 4 | 0 | 5 | 1 | 0 |
| | ID | 4 | 4 | **2 (2)** | 4 | 0 | 4 | 3 | 3 |
| | MSSA | 2 | 4 | 4 | **1 (1)** | 0 | 2 | 3 | 1 |
| | NH | 0 | 0 | 1 | 0 | **3 (0)** | 0 | 0 | 0 |
| | OR | 2 | 5 | 4 | 2 | 0 | **1 (1)** | 1 | 2 |
| | WV | 1 | 1 | 3 | 3 | 0 | 1 | **0 (0)** | 2 |
| | WY | 0 | 0 | 3 | 1 | 0 | 2 | 2 | **0 (0)** |
| Stand-Alone | CT | **10 (9)** | 2 | 13 | 9 | 10 | 3 | 6 | 0 |
| | HI | 2 | **0 (0)** | 9 | 9 | 0 | 6 | 3 | 0 |
| | ID | 13 | 9 | **2 (2)** | 11 | 1 | 12 | 6 | 5 |
| | MSSA | 9 | 9 | 11 | **1 (1)** | 6 | 11 | 9 | 7 |
| | NH | 10 | 0 | 2 | 6 | **3 (1)** | 0 | 0 | 2 |
| | OR | 3 | 6 | 12 | 11 | 0 | **0 (0)** | 2 | 7 |
| | WV | 6 | 3 | 6 | 9 | 1 | 2 | **0 (0)** | 0 |
| | WY | 0 | 0 | 5 | 7 | 2 | 7 | 0 | **0 (0)** |
| Grade Band Total | CT | **15 (14)** | 5 | 17 | 11 | 10 | 5 | 7 | 0 |
| | HI | 5 | **0 (0)** | 13 | 13 | 0 | 11 | 4 | 0 |
| | ID | 17 | 13 | **4 (4)** | 15 | 1 | 16 | 9 | 8 |
| | MSSA | 11 | 13 | 15 | **2 (2)** | 6 | 13 | 12 | 8 |
| | NH | 10 | 0 | 3 | 6 | **6 (1)** | 0 | 0 | 2 |
| | OR | 5 | 11 | 16 | 13 | 0 | **1 (1)** | 3 | 9 |
| | WV | 7 | 4 | 9 | 12 | 1 | 3 | **0 (0)** | 2 |
| | WY | 0 | 0 | 8 | 8 | 2 | 9 | 2 | **0 (0)** |

*Table 22. Number of Common High School Field-Test Items Administered and Calibrated in Spring 2019*

| | State | Connecticut | Hawaii | Idaho | MSSA | New Hampshire | Oregon | West Virginia | Wyoming |
|---|---|---|---|---|---|---|---|---|---|
| Cluster | CT | **9 (9)** | 10 | - | 11 | 0 | 8 | - | 1 |
| | HI | 11 | **0 (0)** | - | 8 | 0 | 11 | - | 0 |
| | ID | - | - | **-** | - | - | - | - | - |
| | MSSA | 12 | 9 | - | **3 (2)** | 0 | 7 | - | 2 |
| | NH | 0 | 0 | - | 0 | **1 (0)** | 1 | - | 0 |
| | OR | 8 | 11 | - | 7 | 1 | **1 (1)** | - | 0 |
| | WV | - | - | - | - | - | - | **-** | - |
| | WY | 1 | 0 | - | 2 | 0 | 0 | - | **0 (0)** |
| Stand-Alone | CT | **14 (13)** | 7 | - | 7 | 6 | 13 | - | 13 |
| | HI | 7 | **0 (0)** | - | 0 | 0 | 6 | - | 0 |
| | ID | - | - | **-** | - | - | - | - | - |
| | MSSA | 8 | 0 | - | **3 (3)** | 6 | 5 | - | 12 |
| | NH | 8 | 0 | - | 6 | **10 (10)** | 0 | - | 7 |
| | OR | 14 | 6 | - | 6 | 0 | **0 (1)** | - | 8 |
| | WV | - | - | - | - | - | - | **-** | - |
| | WY | 14 | 0 | - | 13 | 7 | 9 | - | **0 (0)** |
| Grade Band Total | CT | **23 (22)** | 17 | - | 18 | 6 | 21 | - | 14 |
| | HI | 18 | **0 (0)** | - | 8 | 0 | 17 | - | 0 |
| | ID | - | - | **-** | - | - | - | - | - |
| | MSSA | 20 | 9 | - | **6 (5)** | 6 | 12 | - | 14 |
| | NH | 8 | 0 | - | 6 | **11 (10)** | 1 | - | 7 |
| | OR | 22 | 17 | - | 13 | 1 | **1 (1)** | - | 8 |
| | WV | - | - | - | - | - | - | **-** | - |
| | WY | 15 | 0 | - | 15 | 7 | 9 | - | **0 (0)** |

The calibration and linking of the items field tested in 2019 is explained in detail in Section 5.2 of this volume of the technical report.

Following essentially the same process as explained in Section 3.2.1, items went through a substantial validation process. The modifications to the process followed in 2018 were minor. They included:

- In 2018, all the item statistics were computed on the student data of the students testing in the state that owned the item. In 2019, all the item statistics were computed on the student data of the students testing in the state that owned the item *except for the statistics related* *to differential item functioning (DIF)*. Following the recommendations of several technical advisory committees, the data of states were combined in the calculation of DIF statistics whenever possible (i.e., for states with an independent field test or an operational test for which the relevant demographic variable was available).

- In 2018, for AIRCore items, the data from Connecticut, New Hampshire, Rhode Island, Vermont, and West Virginia (states that used AIRCore items and with either an independent or operational field test) were combined. In 2019, these states were Connecticut, Idaho (only for middle school), New Hampshire, Oregon, Rhode Island, Vermont, and West Virginia.
- The business rule to flag an item cluster for DIF was slightly modified (i.e., made more liberal) following recommendations of several Technical Advisory Committees. The modification is discussed in Section 4.4 on DIF.

Table 23 presents the number of items field tested in Rhode Island and Vermont (or another state), the number of items rejected before or during rubric validation, the number of items sent out to data review, and the number of items rejected during data review. The numbers in parentheses present the number of items owned by Rhode Island and Vermont.

*Table 23. Overview of Science Administration, Rubric Validation, and Item Data Review in Spring 2019*

| Grade Band and Item Type | Number of Items Field Tested | Number of Items Rejected Before/During Rubric Validation | Number of Items Sent to Data Review | Number of Items Rejected at Data Review | Number of Items Remaining[a] |
|---|---|---|---|---|---|
| **Elementary School** | **117 (10)** | **2 (0)** | **72 (5)** | **25 (0)** | **90 (10)** |
| Clusters | 50 (4) | 1 (0) | 16 (0) | 10 (0) | 39 (4) |
| Stand-Alone | 67 (6) | 1 (0) | 56 (5) | 15 (0) | 51 (6) |
| **Middle School** | **127 (8)** | **6 (0)** | **66 (5)** | **23 (2)** | **95 (6)** |
| Clusters | 38 (3) | 1 (0) | 12 (1) | 5 (1) | 29 (2) |
| Stand-Alone | 89 (5) | 5 (0) | 54 (4) | 18 (1) | 66 (4) |
| **High School** | **103 (6)** | **6 (0)** | **52 (4)** | **16 (2)** | **79 (3)** |
| Clusters | 35 (3) | 2 (0) | 15 (1) | 5 (0) | 26 (2) |
| Stand-Alone | 68 (3) | 4 (0) | 37 (3) | 11 (2) | 53 (1) |
| **Total** | **347 (24)** | **14 (0)** | **190 (14)** | **64 (4)** | **264 (19)** |

*Note.* MSSA-owned items are indicated in the parentheses.
[a]Number of items remaining excludes five AI scoring items (four AIRCore and one MSSA-owned) field tested in spring 2019 that were not brought to item data review.

Table 24 summarizes the science item bank after adding the items that were field tested in 2019 and survived rubric validation and item data review. The numbers in parentheses present the number of items owned by Rhode Island and Vermont.

*Table 24. Overview of Combined Science Item Bank in Spring 2019*

| Grade Band and Item Type | Combined Science Item Bank | | | | |
|---|---|---|---|---|---|
| | Total | Earth and Space Sciences | Engineering and Technology | Life Sciences | Physical Sciences |
| **Elementary School** | **225 (17)** | **67 (7)** | **0 (0)** | **77 (6)** | **81 (4)** |
| Cluster | 115 (9) | 34 (3) | 0 (0) | 40 (3) | 41 (3) |
| Stand-Alone | 110 (8) | 33 (4) | 0 (0) | 37 (3) | 40 (1) |
| **Middle School** | **287 (11)** | **81 (2)** | **1 (0)** | **109 (5)** | **96 (4)** |
| Cluster | 165 (5) | 44 (1) | 1 (0) | 63 (2) | 57 (2) |
| Stand-Alone | 122 (6) | 37 (1) | 0 (0) | 46 (3) | 39 (2) |
| **High School** | **201 (9)** | **40 (4)** | **0 (0)** | **108 (2)** | **53 (3)** |
| Cluster | 92 (4) | 19 (2) | 0 (0) | 49 (1) | 24 (1) |
| Stand-Alone | 109 (5) | 21 (2) | 0 (0) | 59 (1) | 29 (2) |
| **Total** | **713 (37)** | **188 (13)** | **1 (0)** | **294 (13)** | **230 (11)** |

*Note.* MSSA-owned items are indicated in the parentheses.

## 3.3 TEST DESIGN

The science tests were assembled under a LOFT test design. Tests were assembled using AIR's adaptive testing algorithm. The adaptive item selection algorithm selects items based on their content value and information value. At any given point during the test, the content value an item is determined by its contribution to meeting the blueprint, given the content characteristics of the items that have already been administered. During the test, the content value increases for items that exhibit features that have not met their designated minimum as the end of the test approaches. Vice versa, the content value decreases for items with content features for which the minimum has been met. The information value of an item is based on the item information function evaluated at the estimated proficiency. The proficiency estimate is updated throughout the test. Under a LOFT test design, the items are selected solely based on their contributions to meeting the blueprint by assigning a weight of zero to the information value of an item with respect to the underlying proficiency. The blueprint is given in Table 25 through Table 27. Details for AIR's adaptive testing algorithm are described in Volume 2, Appendix J, of this technical report.

*Table 25. Science Test Blueprint, Grade 5*

| Grade 5 | Min Item Clusters | Max Item Clusters | Min Stand-Alone Items | Max Stand-Alone Items | Min Item Clusters + Min Stand-Alone Items | Max Item Clusters + Max Stand-Alone Items |
|---|---|---|---|---|---|---|
| **Discipline–Physical Sciences, PE Total = 17** | **2** | **2** | **4** | **4** | **6** | **6** |
| **DCI–Motion and Stability: Forces and Interactions** | **0** | **1** | **0** | **2** | **0** | **3** |
| 3-PS2-1: Forces-balanced and unbalanced forces | 0 | 1 | 0 | 1 | 0 | 1 |
| 3-PS2-2: Forces-pattern predicts future motion | 0 | 1 | 0 | 1 | 0 | 1 |
| 3-PS2-3: Forces-between objects not in contact | 0 | 1 | 0 | 1 | 0 | 1 |
| 3-PS2-4: Forces-magnets* | 0 | 1 | 0 | 1 | 0 | 1 |
| 5-PS2-1: Space Systems | 0 | 1 | 0 | 1 | 0 | 1 |
| **DCI–Energy** | **0** | **1** | **0** | **2** | **0** | **3** |
| 4-PS3-1: Energy-relationship between speed and energy of object | 0 | 1 | 0 | 1 | 0 | 1 |
| 4-PS3-2: Energy-transfer of energy | 0 | 1 | 0 | 1 | 0 | 1 |
| 4-PS3-3: Energy-changes in energy when objects collide | 0 | 1 | 0 | 1 | 0 | 1 |
| 4-PS3-4: Energy-converting energy from one form to another* | 0 | 1 | 0 | 1 | 0 | 1 |
| 5-PS3-1: Matter and Energy | 0 | 1 | 0 | 1 | 0 | 1 |
| **DCI–Waves and Their Applications in Technologies for Information Transfer** | **0** | **1** | **0** | **2** | **0** | **3** |
| 4-PS4-1: Waves-waves can cause objects to move | 0 | 1 | 0 | 1 | 0 | 1 |
| 4-PS4-2: Structure, Function, Information Processing | 0 | 1 | 0 | 1 | 0 | 1 |
| 4-PS4-3: Waves-using patterns to transfer information* | 0 | 1 | 0 | 1 | 0 | 1 |
| **DCI–Matter and Its Interactions** | **0** | **1** | **0** | **2** | **0** | **3** |
| 5-PS1-1: Structure and Properties of Matter | 0 | 1 | 0 | 1 | 0 | 1 |
| 5-PS1-2: Structure and Properties of Matter | 0 | 1 | 0 | 1 | 0 | 1 |
| 5-PS1-3: Structure and Properties of Matter | 0 | 1 | 0 | 1 | 0 | 1 |

| Grade 5 | Min Item Clusters | Max Item Clusters | Min Stand-Alone Items | Max Stand-Alone Items | Min Item Clusters + Min Stand-Alone Items | Max Item Clusters + Max Stand-Alone Items |
|---|---|---|---|---|---|---|
| 5-PS1-4: Structure and Properties of Matter | 0 | 1 | 0 | 1 | 0 | 1 |
| **Discipline–Life Sciences, PE Total = 12** | **2** | **2** | **4** | **4** | **6** | **6** |
| **DCI–From Molecules to Organisms: Structure and Function** | **0** | **1** | **0** | **2** | **0** | **3** |
| 3-LS1-1: Inheritance | 0 | 1 | 0 | 1 | 0 | 1 |
| 4-LS1-1: Structure, Function, Information Processing | 0 | 1 | 0 | 1 | 0 | 1 |
| 4-LS1-2: Structure, Function, Information Processing | 0 | 1 | 0 | 1 | 0 | 1 |
| 5-LS1-1: Matter and Energy | 0 | 1 | 0 | 1 | 0 | 1 |
| **DCI–Ecosystems: Interactions, Energy, and Dynamics** | **0** | **1** | **0** | **2** | **0** | **3** |
| 3-LS2-1: Ecosystems | 0 | 1 | 0 | 1 | 0 | 1 |
| 5-LS2-1: Matter and Energy | 0 | 1 | 0 | 1 | 0 | 1 |
| **DCI–Inheritance and Variation of Traits** | **0** | **1** | **0** | **2** | **0** | **3** |
| 3-LS3-1: Inheritance | 0 | 1 | 0 | 1 | 0 | 1 |
| 3-LS3-2: Inheritance | 0 | 1 | 0 | 1 | 0 | 1 |
| **DCI–Biological Evolution: Unity and Diversity** | **0** | **1** | **0** | **2** | **0** | **3** |
| 3-LS4-1: Ecosystems | 0 | 1 | 0 | 1 | 0 | 1 |
| 3-LS4-2: Inheritance | 0 | 1 | 0 | 1 | 0 | 1 |
| 3-LS4-3: Ecosystems | 0 | 1 | 0 | 1 | 0 | 1 |
| 3-LS4-4: Ecosystems* | 0 | 1 | 0 | 1 | 0 | 1 |
| **Discipline–Earth and Space Sciences, PE Total = 13** | **2** | **2** | **4** | **4** | **6** | **6** |
| **DCI–Earth's Systems** | **0** | **1** | **0** | **2** | **0** | **3** |
| 3-ESS2-1: Weather and Climate | 0 | 1 | 0 | 1 | 0 | 1 |
| 3-ESS2-2: Weather and Climate | 0 | 1 | 0 | 1 | 0 | 1 |
| 4-ESS2-1: Earth's Systems and Processes | 0 | 1 | 0 | 1 | 0 | 1 |

| Grade 5 | Min Item Clusters | Max Item Clusters | Min Stand-Alone Items | Max Stand-Alone Items | Min Item Clusters + Min Stand-Alone Items | Max Item Clusters + Max Stand-Alone Items |
|---|---|---|---|---|---|---|
| 4-ESS2-2: Earth's Systems and Processes | 0 | 1 | 0 | 1 | 0 | 1 |
| 5-ESS2-1: Earth's Systems | 0 | 1 | 0 | 1 | 0 | 1 |
| 5-ESS2-2: Earth's Systems | 0 | 1 | 0 | 1 | 0 | 1 |
| **DCI–Earth and Human Activity** | **0** | **1** | **0** | **2** | **0** | **3** |
| 3-ESS3-1: Weather and Climate* | 0 | 1 | 0 | 1 | 0 | 1 |
| 4-ESS3-2: Earth's Systems and Processes* | 0 | 1 | 0 | 1 | 0 | 1 |
| 4-ESS3-1: Energy | 0 | 1 | 0 | 1 | 0 | 1 |
| 5-ESS3-1: Earth's Systems | 0 | 1 | 0 | 1 | 0 | 1 |
| **DCI–Earth's Place in the Universe** | **0** | **1** | **0** | **2** | **0** | **3** |
| 4-ESS1-1: Earth's Systems and Processes | 0 | 1 | 0 | 1 | 0 | 1 |
| 5-ESS1-1: Space Systems | 0 | 1 | 0 | 1 | 0 | 1 |
| 5-ESS1-2: Space Systems | 0 | 1 | 0 | 1 | 0 | 1 |
| **PE Total = 42** | **6** | **6** | **12** | **12** | **18** | **18** |

*Note*. *These PEs have an engineering component.

*Table 26. Science Test Blueprint, Grade 8*

| Grade 8 | Min Item Clusters | Max Item Clusters | Min Stand-Alone Items | Max Stand-Alone Items | Min Item Clusters + Min Stand-Alone Items | Max Item Clusters + Max Stand-Alone Items |
|---|---|---|---|---|---|---|
| **Discipline–Physical Sciences, PE Total = 19** | **2** | **2** | **4** | **4** | **6** | **6** |
| **DCI–Matter and Its Interactions** | **0** | **1** | **0** | **2** | **0** | **3** |
| MS-PS1-1: Structure and Properties of Matter | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-PS1-2: Chemical Reactions | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-PS1-3: Structure and Properties of Matter | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-PS1-4: Structure and Properties of Matter | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-PS1-5: Chemical Reactions | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-PS1-6: Chemical Reactions* | 0 | 1 | 0 | 1 | 0 | 1 |
| **DCI–Motion and Stability: Forces and Interactions** | **0** | **1** | **0** | **2** | **0** | **3** |
| MS-PS2-1: Forces and Interactions* | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-PS2-2: Forces and Interactions | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-PS2-3: Forces and Interactions | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-PS2-4: Forces and Interactions | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-PS2-5: Forces and Interactions | 0 | 1 | 0 | 1 | 0 | 1 |
| **DCI–Energy** | **0** | **1** | **0** | **2** | **0** | **3** |
| MS-PS3-1: Energy | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-PS3-2: Energy | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-PS3-3: Energy* | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-PS3-4: Energy | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-PS3-5: Energy | 0 | 1 | 0 | 1 | 0 | 1 |
| **DCI–Waves and Their Applications in Technologies for Information Transfer** | **0** | **1** | **0** | **2** | **0** | **3** |
| MS-PS4-1: Waves and Electromagnetic Radiation | 0 | 1 | 0 | 1 | 0 | 1 |

| Grade 8 | Min Item Clusters | Max Item Clusters | Min Stand-Alone Items | Max Stand-Alone Items | Min Item Clusters + Min Stand-Alone Items | Max Item Clusters + Max Stand-Alone Items |
|---|---|---|---|---|---|---|
| MS-PS4-2: Waves and Electromagnetic Radiation | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-PS4-3: Waves and Electromagnetic Radiation | 0 | 1 | 0 | 1 | 0 | 1 |
| **Discipline–Life Sciences, PE Total = 21** | **2** | **2** | **4** | **4** | **6** | **6** |
| **DCI–From Molecules to Organisms: Structures and Processes** | **0** | **1** | **0** | **2** | **0** | **3** |
| MS-LS1-1: Structure, Function, Information Processing | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-LS1-2: Structure, Function, Information Processing | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-LS1-3: Structure, Function, Information Processing | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-LS1-4: Growth, Development, Reproduction | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-LS1-5: Growth, Development, Reproduction | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-LS1-6: Matter and Energy | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-LS1-7: Matter and Energy | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-LS1-8: Structure, Function, Information Processing | 0 | 1 | 0 | 1 | 0 | 1 |
| **DCI–Ecosystems: Interactions, Energy, and Dynamics** | **0** | **1** | **0** | **2** | **0** | **3** |
| MS-LS2-1: Matter and Energy | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-LS2-2: Interdependent Relationships in Ecosystems | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-LS2-3: Matter and Energy | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-LS2-4: Matter and Energy | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-LS2-5: Interdependent Relationships in Ecosystems* | 0 | 1 | 0 | 1 | 0 | 1 |
| **DCI–Hereditary: Inheritance and Variation of Traits** | **0** | **1** | **0** | **2** | **0** | **3** |
| MS-LS3-1: Growth, Development, Reproduction | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-LS3-2: Growth, Development, Reproduction | 0 | 1 | 0 | 1 | 0 | 1 |
| **DCI–Biological Evolution: Unity and Diversity** | **0** | **1** | **0** | **2** | **0** | **3** |
| MS-LS4-1: Natural Selection and Adaptation | 0 | 1 | 0 | 1 | 0 | 1 |

| Grade 8 | Min Item Clusters | Max Item Clusters | Min Stand-Alone Items | Max Stand-Alone Items | Min Item Clusters + Min Stand-Alone Items | Max Item Clusters + Max Stand-Alone Items |
|---|---|---|---|---|---|---|
| MS-LS4-2: Natural Selection and Adaptation | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-LS4-3: Natural Selection and Adaptation | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-LS4-4: Natural Selection and Adaptation | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-LS4-5: Growth, Development, Reproduction | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-LS4-6: Natural Selection and Adaptation | 0 | 1 | 0 | 1 | 0 | 1 |
| **Discipline–Earth and Space Sciences, PE Total = 15** | **2** | **2** | **4** | **4** | **6** | **6** |
| **DCI–Earth's Place in the Universe** | **0** | **1** | **0** | **2** | **0** | **3** |
| MS-ESS1-1: Space Systems | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-ESS1-2: Space Systems | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-ESS1-3: Space Systems | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-ESS1-4: History of Earth | 0 | 1 | 0 | 1 | 0 | 1 |
| **DCI–Earth's Systems** | **0** | **1** | **0** | **2** | **0** | **3** |
| MS-ESS2-1: Earth's Systems | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-ESS2-2: History of Earth | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-ESS2-3: History of Earth | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-ESS2-4: Earth's Systems | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-ESS2-5: Weather and Climate | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-ESS2-6: Weather and Climate | 0 | 1 | 0 | 1 | 0 | 1 |
| **DCI–Earth and Human Activity** | **0** | **1** | **0** | **2** | **0** | **3** |
| MS-ESS3-1: Earth's Systems | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-ESS3-2: Human Impacts | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-ESS3-3: Human Impacts* | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-ESS3-4: Human Impacts | 0 | 1 | 0 | 1 | 0 | 1 |

| Grade 8 | Min Item Clusters | Max Item Clusters | Min Stand-Alone Items | Max Stand-Alone Items | Min Item Clusters + Min Stand-Alone Items | Max Item Clusters + Max Stand-Alone Items |
|---|---|---|---|---|---|---|
| MS-ESS3-5: Weather and Climate | 0 | 1 | 0 | 1 | 0 | 1 |
| **PE Total = 55** | **6** | **6** | **12** | **12** | **18** | **18** |

*Note*. *These PEs have an engineering component.

*Table 27. Science Test Blueprint, Grade 11*

| Grade 11 | Min Item Clusters | Max Item Clusters | Min Stand-Alone Items | Max Stand-Alone Items | Min Item Clusters + Min Stand-Alone Items | Max Item Clusters + Max Stand-Alone Items |
|---|---|---|---|---|---|---|
| **Discipline–Physical Sciences, PE Total = 24** | **2** | **2** | **4** | **4** | **6** | **6** |
| **DCI–Matter and Its Interactions** | **0** | **1** | **0** | **2** | **0** | **3** |
| HS-PS1-1: Structure and Properties of Matter | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-PS1-2: Structure and Properties of Matter | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-PS1-3: Structure and Properties of Matter | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-PS1-4: Chemical Reactions | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-PS1-5: Chemical Reactions | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-PS1-6: Chemical Reactions* | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-PS1-7: Chemical Reactions | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-PS1-8: Nuclear Processes | 0 | 1 | 0 | 1 | 0 | 1 |
| **DCI–Motion and Stability: Forces and Interactions** | **0** | **1** | **0** | **2** | **0** | **3** |
| HS-PS2-1: Forces and Motion | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-PS2-2: Forces and Motion | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-PS2-3: Forces and Motion* | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-PS2-4: Types of Interactions | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-PS2-5: Types of Interactions | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-PS2-6: Chemical Reactions* | 0 | 1 | 0 | 1 | 0 | 1 |
| **DCI–Energy** | **0** | **1** | **0** | **2** | **0** | **3** |
| HS-PS3-1: Energy | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-PS3-2: Energy | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-PS3-3: Energy* | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-PS3-4: Energy | 0 | 1 | 0 | 1 | 0 | 1 |

| Grade 11 | Min Item Clusters | Max Item Clusters | Min Stand-Alone Items | Max Stand-Alone Items | Min Item Clusters + Min Stand-Alone Items | Max Item Clusters + Max Stand-Alone Items |
|---|---|---|---|---|---|---|
| HS-PS3-5: Energy | 0 | 1 | 0 | 1 | 0 | 1 |
| **DCI–Waves and Their Applications in Technologies for Information Transfer** | **0** | **1** | **0** | **2** | **0** | **3** |
| HS-PS4-1: Wave Properties | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-PS4-2: Wave Properties | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-PS4-3: Wave Properties/Electromagnetic Radiation | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-PS4-4: Electromagnetic Radiation | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-PS4-5: Electromagnetic Radiation* | 0 | 1 | 0 | 1 | 0 | 1 |
| **Discipline–Life Sciences, PE Total = 24** | **2** | **2** | **4** | **4** | **6** | **6** |
| **DCI–From Molecules to Organisms: Structures and Processes** | **0** | **1** | **0** | **2** | **0** | **3** |
| HS-LS1-1: Structure and Function | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-LS1-2: Structure and Function | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-LS1-3: Structure and Function | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-LS1-4: Growth and Development of Organisms | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-LS1-5: Organization for Matter and Energy Flow in Organisms | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-LS1-6: Organization for Matter and Energy Flow in Organisms | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-LS1-7: Organization for Matter and Energy Flow in Organisms | 0 | 1 | 0 | 1 | 0 | 1 |
| **DCI–Ecosystems: Interactions, Energy and Dynamics** | **0** | **1** | **0** | **2** | **0** | **3** |
| HS-LS2-1: Interdependent Relationships in Ecosystems | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-LS2-2: Interdependent Relationships in Ecosystems | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-LS2-3: Cycles of Matter and Energy Transfer in Ecosystems | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-LS2-4: Cycles of Matter and Energy Transfer in Ecosystems | 0 | 1 | 0 | 1 | 0 | 1 |

| Grade 11 | Min Item Clusters | Max Item Clusters | Min Stand-Alone Items | Max Stand-Alone Items | Min Item Clusters + Min Stand-Alone Items | Max Item Clusters + Max Stand-Alone Items |
|---|---|---|---|---|---|---|
| HS-LS2-5: Cycles of Matter and Energy Transfer in Ecosystems | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-LS2-6: Ecosystem Dynamics, Functioning, and Resilience | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-LS2-7: Ecosystem Dynamics, Functioning, and Resilience* | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-LS2-8: Social Interactions and Group Behavior | 0 | 1 | 0 | 1 | 0 | 1 |
| **DCI–Heredity: Inheritance and Variation of Traits** | **0** | **1** | **0** | **2** | **0** | **3** |
| HS-LS3-1: Structure and Function | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-LS3-2: Variation of Traits | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-LS3-3: Variation of Traits | 0 | 1 | 0 | 1 | 0 | 1 |
| **DCI–Biological Evolution: Unity and Diversity** | **0** | **1** | **0** | **2** | **0** | **3** |
| HS-LS4-1: Evidence of Common Ancestry and Diversity | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-LS4-2: Natural Selection | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-LS4-3: Natural Selection | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-LS4-4: Adaptation | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-LS4-5: Adaptation | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-LS4-6: Adaptation* | 0 | 1 | 0 | 1 | 0 | 1 |
| **Discipline–Earth and Space Sciences, PE Total = 19** | **2** | **2** | **4** | **4** | **6** | **6** |
| **DCI–Earth's Place in the Universe** | **0** | **1** | **0** | **2** | **0** | **3** |
| HS-ESS1-1: The Universe and Its Stars | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-ESS1-2: The Universe and Its Stars | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-ESS1-3: The Universe and Its Stars | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-ESS1-4: Earth and the Solar System | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-ESS1-5: The History of Planet Earth | 0 | 1 | 0 | 1 | 0 | 1 |

| Grade 11 | Min Item Clusters | Max Item Clusters | Min Stand-Alone Items | Max Stand-Alone Items | Min Item Clusters + Min Stand-Alone Items | Max Item Clusters + Max Stand-Alone Items |
|---|---|---|---|---|---|---|
| HS-ESS1-6: The History of Planet Earth | 0 | 1 | 0 | 1 | 0 | 1 |
| **DCI–Earth's Systems** | **0** | **1** | **0** | **2** | **0** | **3** |
| HS-ESS2-1: Earth Materials and Systems | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-ESS2-2: Earth Materials and Systems | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-ESS2-3: Earth Materials and Systems | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-ESS2-4: Weather and Climate | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-ESS2-5: The Roles of Water in Earth's Surface Processes | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-ESS2-6: Weather and Climate | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-ESS2-7: Weather and Climate | 0 | 1 | 0 | 1 | 0 | 1 |
| **DCI–Earth and Human Activity** | **0** | **1** | **0** | **2** | **0** | **3** |
| HS-ESS3-1: Natural Resources | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-ESS3-2: Natural Resources* | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-ESS3-3: Human Impacts on Earth Systems | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-ESS3-4: Human Impacts on Earth Systems* | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-ESS3-5: Global Climate Change | 0 | 1 | 0 | 1 | 0 | 1 |
| HS-ESS3-6: Global Climate Change* | 0 | 1 | 0 | 1 | 0 | 1 |
| **PE Total = 67** | **6** | **6** | **12** | **12** | **18** | **18** |

*Note*. *These PEs have an engineering component.

The main characteristics of the blueprint were that any performance expectation could be tested only once (indicated by the values of 0 and 1 for the Min and Max values of the individual performance expectations [PEs] in Table 25 through Table 27); in general, no more than one item cluster or two stand-alone items could be sampled from the same disciplinary core idea, and no more than three total items could be sampled from the same disciplinary core idea (as indicated by the Min and Max values in the rows representing disciplinary core ideas). For both the 2018 and 2019 test administrations, a segmented test design was used; items were administered grouped in four segments. The segments corresponded to each of the three science disciplines and a (additional) field-test segment that could contain items from all three science disciplines.

In 2018, the order of the segments corresponding to the science disciplines was randomized over students. The additional field-test segment consisted on one cluster and was always presented at the end of the test (segment 4). The primary purpose was to collect additional student responses for the item clusters that had low exposure in the first three segments.

In 2019, the scored operational part of the test consisted of the three segments corresponding to science disciplines. The embedded field-test segment consisted of two item clusters and four stand-alone items. In order to ensure that every student received exactly two item clusters and four stand-alone items as field-test items, the embedded field-test segment was split into two segments: one for field-test item clusters, and one for field-test stand-alone items. The test was taken over two days. On the first day, half of the students received two operational segments, chosen at random from the three operational segments. The other half received one randomly chosen operational segment and the embedded field-test segments. The remaining segments were administered on the second day. Within a day, the order of the segments was randomized, with the restriction that the field-test segments for item clusters and stand-alone items were always administered right after each other.

# 4. FIELD TEST CLASSICAL ANALYSIS OVERVIEW

As explained in Section 3 of this volume, science items administered as field-test items in 2018 and 2019 in Rhode Island and Vermont or any of the states that signed the memorandum of understanding (MOU) for item sharing underwent rubric validation and data review. Items were flagged for data review based on business rules defined on classical item statistics. Except for response times, the classical item statistics are computed for individual assertions, whereas the business rules for flagging are defined at the item level. In general, item statistics used to flag items for data review were computed using the student responses of the state that owned the item. However, for AIRCore items, the flagging rules were defined on the item statistics computed from the combined data of states that used AIRCore items and that administered either an independent or operational field test (Connecticut, Idaho grade 8, New Hampshire, Oregon, Rhode Island, Vermont, and West Virginia). Furthermore, for the computation of differential item functioning (DIF) statistics, the data of all states with an operational or independent field test were combined to obtain enough students for each demographic group. The criteria for flagging and reviewing items are provided in Table 28, and the statistics are described below. Items that were flagged for data review were reviewed by a committee, as explained in Section 3 of this volume.

*Table 28. Thresholds for Flagging in Classical Item Analysis*

| Analysis Type | Flagging Criteria |
|---|---|
| Item Discrimination | Average biserial correlation < 0.25 (across the assertions within an item) |
| | One or more assertions with a biserial correlation < 0 |
| Item Difficulty (Clusters) | Average *p*-value < .30 or > 0.85 (across the assertions within an item cluster) |
| Item Difficulty (Stand-Alone items) | Average *p*-value < .15 or > 0.95 (across the assertions within a stand-alone item) |
| Timing (Clusters) | Percentile 80* > 15 minutes |
| Timing (Stand-Alone items) | Percentile 80* > 3 minutes |
| Timing | Assertions per (percentile 80) minute < 0.5 |
| DIF (Clusters) | Two or more assertions show ‘C’ DIF in the same direction |
| DIF (Stand-Alone items) | One or more assertions show ‘C’ DIF in the same direction |

*Note.* *A percentile 80 of x minutes: 80% of the students spent x minutes or less on the item.

## 4.1 Item Discrimination

The item discrimination index indicates the extent to which each item differentiated between those test takers who possessed the skills being measured and those who did not. Generally, the higher the value, the better the item was able to differentiate between high- and low-achieving students.

For each assertion within an item, the discrimination index was calculated as the biserial correlation between the assertion score and the ability estimate for students. The average biserial correlation was then calculated across the assertions within an item.

## 4.2 Item Difficulty

Items that are either very difficult or very easy are flagged for review but are not necessarily removed if they are grade-level appropriate and aligned with the test specifications. For science, both the *p*-value for individual assertions and the average across all assertions of an item are calculated. Acceptable item *p*-values are summarized in Table 28.

## 4.3 Response Time

Given that the science clusters consist of multiple student interactions, they require more time for students to complete. To ensure a good balance between the amount of information an item provides, and the time students spend on the item, item response time was recorded and analyzed. Specifically, the statistic “percentile 80” was computed for each item. A percentile 80 of $x$ minutes means that 80% of the students spent $x$ minutes or fewer on the item. An item was flagged for review when

- percentile 80 > 15 minutes, if the item is an item cluster;
- percentile 80 > 3 minutes, if the item is a stand-alone item; or

- assertions per (percentile 80) minute < 0.5.

## 4.4 Differential Item Functioning Analysis

DIF refers to items that appear to function differently across identifiable groups, typically across different demographic groups. Identifying DIF is important, because it provides a statistical indicator that an item may contain cultural or other bias. DIF-flagged items are further examined by content experts who are asked to re-examine each flagged item to decide whether the item should be excluded from the pool due to bias. Not all items that exhibit DIF are biased; characteristics of the educational system may also lead to DIF.

AIR uses a generalized Mantel-Haenszel (MH) procedure to calculate DIF. The generalizations include (1) adaptation to polytomous items; and (2) improved variance estimators to render the test statistics valid under complex sample designs. With this procedure, each student's estimated theta score on the operational items on a given test is used as the ability-matching variable. That score is divided into 10 intervals to compute the MH$\chi^2$ DIF statistics for balancing the stability and sensitivity of the DIF scoring category selection. The analysis program computes the $MH\chi^2$ value, the conditional odds ratio, and the MH-delta for dichotomous items; the $GMH\chi^2$ and the standardized mean difference (SMD) are computed for polytomous items.

The MH chi-square statistic (Holland & Thayer, 1988) is calculated as:

$$MH\chi^2 = \frac{(|\sum_k n_{R1k} - \sum_k E(n_{R1k})| - 0.5)^2}{\sum_k var(n_{R1k})}$$

where $k = \{1, 2, \ldots K\}$ for the strata, $n_{R1k}$ is the number of correct responses for the reference group in stratum $k$, and 0.5 is a continuity correction. The expected value is calculated as

$$E(n_{R1k}) = \frac{n_{+1k}n_{R+k}}{n_{++k}}$$

where $n_{+1k}$ is the total number of correct responses, $n_{R+k}$ is the number of students in the reference group, and $n_{++k}$ is the number of students in stratum $k$, and the variance is calculated as

$$var(n_{R1k}) = \frac{n_{R+k}n_{F+k}n_{+1k}n_{+0k}}{n_{++k}^2(n_{++k}-1)}.$$

$n_{F+k}$ is the number of students in the focal group, $n_{+1k}$ is the number of students with correct responses, and $n_{+0k}$ is the number of students with incorrect responses in stratum $k$.

The MH conditional odds ratio is calculated as

$$\alpha_{MH} = \frac{\sum_k n_{R1k}n_{F0k}/n_{++k}}{\sum_k n_{R0k}n_{F1k}/n_{++k}}.$$

The MH-delta ($\Delta_{MH}$, Holland & Thayer, 1988) is then defined as

$$\Delta_{MH} = -2.35\ln(\alpha_{MH}).$$

The GMH statistic generalizes the MH statistic to polytomous items (Somes, 1986), and is defined as

$$GMH\chi^2 = \left(\sum_k \boldsymbol{a}_k - \sum_k E(\boldsymbol{a}_k)\right)' \left(\sum_k var(\boldsymbol{a}_k)\right)^{-1} \left(\sum_k \boldsymbol{a}_k - \sum_k E(\boldsymbol{a}_k)\right)$$

where $\boldsymbol{a}_k$ is a $(T-1)\,X\,1$ vector of item response scores, corresponding to the $T$ response categories of a polytomous item (excluding one response). $E(\boldsymbol{a}_k)$ and $var(\boldsymbol{a}_k)$, a $(T-1)\times(T-1)$ variance matrix, are calculated analogously to the corresponding elements in $MH\chi^2$ in stratum $k$.

The SMD (Dorans & Schmitt, 1991) is defined as

$$SMD = \sum_k p_{FK} m_{FK} - \sum_k p_{FK} m_{RK}$$

where

$$p_{FK} = \frac{n_{F+k}}{n_{F++}}$$

is the proportion of the focal group students in stratum $k$,

$$m_{FK} = \frac{1}{n_{F+k}}\left(\sum_t a_t n_{Ftk}\right)$$

is the mean item score for the focal group in stratum $k$, and

$$m_{RK} = \frac{1}{n_{R+k}}\left(\sum_t a_t n_{Rtk}\right)$$

is the mean item score for the reference group in stratum $k$.

DIF analysis was conducted for all field-test items with at least 200 responses per item in each subgroup (Zwick, 2012) to detect potential item bias for major demographic groups. Student responses from multiple states were combined to minimize the number of items with insufficient sample sizes for one or more demographic groups.

DIF statistics were calculated at the assertion level and were performed for the following groups (some items had insufficient sample sizes for DIF analyses in some groups):

- Male vs. Female
- American Indian/Alaskan Native vs. White
- Asian vs. White
- African American vs. White
- Hispanic vs. White
- English Language Learner (ELL) vs. Non-ELL

- Special Education (SPED) vs. Non-SPED
- Economically Disadvantaged vs. Non-Economically Disadvantaged

Just like the general MH statistic is used to classify items of traditional tests, assertions were classified into three categories (A, B, or C) for DIF, ranging from no evidence of DIF to severe DIF. The classification rules are shown in Table 29. Furthermore, assertions were categorized positively (i.e., +A, +B, or +C), signifying that an item favored the focal group (e.g., African American/Black, Hispanic, or male), or negatively (i.e., –A, –B, or–C), signifying that an item favored the reference group (e.g., white or female).

An item was flagged for data review according to the following criteria:

- **Item Clusters.** Two or more assertions showed "C" DIF in the same direction.
- **Stand-Alone Items**. One or more assertions showed "C" DIF in the same direction.

*Table 29. DIF Classification Rules*

| **Assertions** | |
|---|---|
| *Category* | *Rule* |
| C | $MH_{\chi^2}$ is significant and $\lvert SMD\rvert/\lvert SD\rvert \geq 0.25$. |
| B | $MH_{\chi^2}$ is significant and $\lvert SMD\rvert/\lvert SD\rvert \geq 0.25$. |
| A | $MH_{\chi^2}$ is not significant. |

Note that for the 2018 field test, a slightly less strict criterion was used for item clusters with 10 or more assertions (i.e., three or more assertions with C DIF in the same direction). The change was made taking into consideration the feedback received at several technical advisory committees and modified such that the rate of flagging items for DIF was similar for item clusters and stand-alone items (based in the flagging rates computed on items field tested in 2018).

## 4.5 Classical Analysis Results

This section presents a summary of results from classical item analysis of the 2018 and 2019 field-test items. The classical statistics of all field-test and operational science items are presented in Appendix A. The statistics were computed separately for each state where the item was administered. The average value of the statistics across states is also presented. Code books are provided to help with navigating the view of a specific subset of items (e.g., checking the statistics of the Rhode Island- and Vermont-owned field-test item calculated using Rhode Island and Vermont data only by applying relevant filters).

Table 30 through Table 37 provide a summary of the *p*-values and biserial correlations for the field tested items in Rhode Island and Vermont. The statistics were computed using only data from Rhode Island and Vermont. The average values across the assertions within an item was used in the computation of the percentiles and ranges.

*Table 30. Distribution of p-Values for Field-Test Items in Rhode Island, 2018*

| Grade | Total FT Items | Min | 5th Percentile | 25th Percentile | 50th Percentile | 75th Percentile | 95th Percentile | Max |
|---|---|---|---|---|---|---|---|---|
| **5** | 67 | 0.19 | 0.25 | 0.44 | 0.52 | 0.61 | 0.73 | 0.78 |
| **8** | 56 | 0.09 | 0.22 | 0.34 | 0.46 | 0.55 | 0.64 | 0.69 |
| **11** | 70 | 0.01 | 0.12 | 0.26 | 0.32 | 0.42 | 0.54 | 0.63 |

*Table 31. Distribution of Item Biserial Correlations for Field-Test Items in Rhode Island, 2018*

| Grade | Total FT Items | Min | 5th Percentile | 25th Percentile | 50th Percentile | 75th Percentile | 95th Percentile | Max |
|---|---|---|---|---|---|---|---|---|
| **5** | 67 | 0.09 | 0.27 | 0.35 | 0.48 | 0.55 | 0.69 | 0.74 |
| **8** | 56 | 0.16 | 0.23 | 0.32 | 0.45 | 0.54 | 0.62 | 0.69 |
| **11** | 70 | 0.11 | 0.20 | 0.35 | 0.42 | 0.50 | 0.55 | 0.68 |

*Table 32. Distribution of p-Values for Field-Test Items in Vermont, 2018*

| Grade | Total FT Items | Min | 5th Percentile | 25th Percentile | 50th Percentile | 75th Percentile | 95th Percentile | Max |
|---|---|---|---|---|---|---|---|---|
| **5** | 67 | 0.20 | 0.27 | 0.47 | 0.56 | 0.65 | 0.75 | 0.84 |
| **8** | 56 | 0.09 | 0.26 | 0.39 | 0.51 | 0.60 | 0.71 | 0.75 |
| **11** | 70 | 0.01 | 0.16 | 0.30 | 0.37 | 0.49 | 0.59 | 0.65 |

*Table 33. Distribution of Item Biserial Correlations for Field-Test Items in Vermont, 2018*

| Grade | Total FT Items | Min | 5th Percentile | 25th Percentile | 50th Percentile | 75th Percentile | 95th Percentile | Max |
|---|---|---|---|---|---|---|---|---|
| **5** | 67 | 0.12 | 0.23 | 0.34 | 0.47 | 0.55 | 0.66 | 0.76 |
| **8** | 56 | 0.14 | 0.21 | 0.30 | 0.46 | 0.51 | 0.61 | 0.67 |
| **11** | 70 | 0.15 | 0.26 | 0.37 | 0.43 | 0.51 | 0.59 | 0.62 |

*Table 34. Distribution of p-Values for Field-Test Items in Rhode Island, 2019*

| Grade | Total FT Items | Min | 5th Percentile | 25th Percentile | 50th Percentile | 75th Percentile | 95th Percentile | Max |
|---|---|---|---|---|---|---|---|---|
| **5** | 41 | 0.05 | 0.36 | 0.46 | 0.55 | 0.71 | 0.84 | 0.92 |
| **8** | 46 | 0.09 | 0.11 | 0.29 | 0.41 | 0.56 | 0.73 | 0.80 |
| **11** | 34 | 0.05 | 0.15 | 0.21 | 0.28 | 0.41 | 0.51 | 0.54 |

*Note.* Items rejected at rubric validation were removed from the calculation of *p*-values.

*Table 35. Distribution of Item Biserial Correlations for Field-Test Items in Rhode Island, 2019*

| Grade | Total FT Items | Min | 5th Percentile | 25th Percentile | 50th Percentile | 75th Percentile | 95th Percentile | Max |
|---|---|---|---|---|---|---|---|---|
| **5** | 41 | 0.30 | 0.33 | 0.43 | 0.51 | 0.60 | 0.65 | 0.73 |
| **8** | 46 | -0.01 | 0.09 | 0.34 | 0.46 | 0.58 | 0.66 | 0.71 |
| **11** | 34 | -0.18 | 0.13 | 0.34 | 0.45 | 0.58 | 0.67 | 0.75 |

*Note.* Items rejected at rubric validation were removed from the calculation of biserial.

*Table 36. Distribution of p-Values for Field-Test Items in Vermont, 2019*

| Grade | Total FT Items | Min | 5th Percentile | 25th Percentile | 50th Percentile | 75th Percentile | 95th Percentile | Max |
|---|---|---|---|---|---|---|---|---|
| **5** | 41 | 0.06 | 0.37 | 0.50 | 0.60 | 0.72 | 0.86 | 0.92 |
| **8** | 46 | 0.10 | 0.14 | 0.31 | 0.47 | 0.60 | 0.75 | 0.84 |
| **11** | 34 | 0.06 | 0.17 | 0.23 | 0.32 | 0.43 | 0.55 | 0.58 |

*Note.* Items rejected at rubric validation were removed from the calculation of *p*-values.

*Table 37. Distribution of Item Biserial Correlations for Field-Test Items in Vermont, 2019*

| Grade | Total FT Items | Min | 5th Percentile | 25th Percentile | 50th Percentile | 75th Percentile | 95th Percentile | Max |
|---|---|---|---|---|---|---|---|---|
| **5** | 41 | 0.25 | 0.30 | 0.42 | 0.48 | 0.55 | 0.64 | 0.65 |
| **8** | 46 | -0.04 | 0.09 | 0.32 | 0.45 | 0.56 | 0.63 | 0.69 |
| **11** | 34 | -0.18 | 0.16 | 0.37 | 0.48 | 0.54 | 0.66 | 0.70 |

*Note.* Items rejected at rubric validation were removed from the calculation of biserial.

In 2018, there were 200 items field tested, seven of which were rejected during rubric validation or due to scoring issues. Among the remaining items, 26 items were flagged for item discrimination, 15 items were flagged for *p*-value, 34 items were flagged for response time, and 34 items were flagged for DIF according to the criteria outlined in the sections above. In 2019, there were 125 items field tested, 121 of which survived rubric validation. Among these 121 items, 18 items were flagged for item discrimination, 16 items were flagged for *p*-value, 39 items were flagged for response time, and 13 items were flagged for DIF according to the criteria outlined in the preceding sections. Some items were flagged for multiple reasons. Flagged items were reviewed by educators during data review.

# 5. ITEM CALIBRATION

## 5.1 MODEL DESCRIPTION

In discussing item response theory (IRT) models for Rhode Island and Vermont, we distinguish between the underlying latent structure of a model and the parameterization of the item response function conditional on that assumed latent structure. Subsequently, we discuss how group effects are considered.

### 5.1.1 Latent Structure

Most operational assessment programs rely on a unidimensional IRT model for item calibration and computing scores for students. These models assume a single underlying trait, and they assume that items are independent given that underlying trait. In other words, the models assume that given the value of the underlying trait, knowing the response to one item provides no information about responses to other items. This assumption of conditional independence implies that the conditional probability of a pattern of *I* item responses takes the relatively simple form of a product over items for a single student:

$$P(\boldsymbol{z_j}|\theta_j) = \prod_{i=1}^{I} P(z_{ij}|\theta_j) \tag{1}$$

where $z_{ij}$ represents the scored response of student $j$ ($j$ = 1, …, $N$) to item $i$ ($I$ = 1, …, $I$), $\boldsymbol{z_j}$ represents the pattern of scored item responses for student $j$, and $\theta_j$ represents student $j$'s proficiency. Unidimensional IRT models differ with respect to the functional relation between the proficiency $\theta_j$ and the probability of obtaining a score $z_{ij}$ on item $i$.

The items of the MSSA are more complex than traditional item types. A single item may contain multiple parts, and each part may contain multiple student interactions. For example, a student may be asked to select a term from a set of terms at several places in a single item. Instead of receiving a single score for each item, multiple inferences are made about the knowledge and skills that a student has demonstrated based on specific features of the student's responses to the item. These scoring units are called *assertions* and are the basic unit of analysis in our IRT analysis. That is, they fulfill the role of items in traditional assessments. However, for the MSSA items, multiple assertions are typically developed around a single item so that assertions are clustered within items.

One approach is to apply one of the traditional IRT models to the scored assertions. However, a substantial complexity that arises from the use of this new item type is that local dependencies exist between assertions pertaining to the same stimulus (item or item cluster). The local dependencies between the assertions pertaining to the same stimulus constitute a violation of the assumption that a single latent trait can explain all dependencies between assertions. Fitting a unidimensional model in the presence of local dependencies may result in biased item parameters and standard errors of measurement (SEM). In particular, it is well documented that ignoring local item dependencies leads to an overestimation of the amount of information conveyed by a set of responses and an underestimation of the SEM (e.g., Sireci, Wainer, & Thissen, 1991; Yen, 1993).

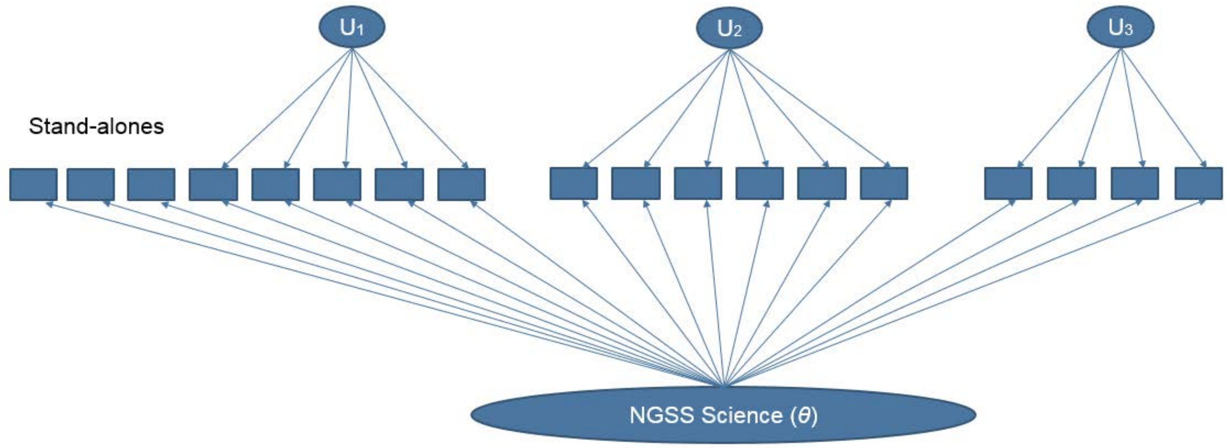
The effects of groups of assertions developed around a common stimulus can be accounted for by including additional dimensions corresponding to those groupings in the IRT model. These dimensions are considered to be nuisance dimensions. Whereas traditional unidimensional IRT models assume that all assertions (the basic units of analysis) are independent given a single underlying trait $\theta$, we now assume the conditional independence of assertions, given the underlying latent trait $\theta$ and all nuisance dimensions:

$$P(\boldsymbol{z_j}|\theta_j, \mathbf{u}_j) = \prod_{i \in \mathrm{SA}} P(z_{ij}|\theta_j) \prod_{g=1}^{G} \prod_{i \in g} P(z_{ij}|\theta_j, u_{jg}) \qquad (2)$$

where SA indicates stand-alone assertions, $u_g$ indicates the nuisance dimension for assertion group $g$ (with the position of student $j$ on that dimension denoted as $u_{jg}$), and $u$ is the vector of all $G$ nuisance dimensions. It can be seen that the conditional probability $P(z_{ij}|\theta_j, u_{jg})$ now becomes a function of two latent variables: the latent trait $\theta$, representing a student's proficiency in science (the underlying trait of interest), and the nuisance dimension $u_g$, accounting for the conditional dependencies between assertions of the same group. Furthermore, we assume that the nuisance dimensions are all uncorrelated with one another and with the general dimension. It is important to point out that even though every group of assertions introduces an additional dimension, models with this latent structure do not suffer from the curse of dimensionality like other multidimensional IRT models because one can take advantage of this special structure during model calibration (Gibbons & Hedeker, 1992). In this regard, Rijmen (2010) showed that it is unnecessary to assume that all nuisance dimensions are uncorrelated; rather, it is sufficient that they are independent, given the general dimension $\theta$.

The model structure of the IRT model for science is illustrated in Figure 1. Note that stand-alone items can be scored with more than one assertion. The assertions of stand-alone items with more than one assertion but fewer than four assertions were also modeled as stand-alone assertions. Even though these assertions are likely to exhibit conditional dependencies, the variance of the nuisance dimension cannot be reliably estimated if it is based on a very small number of assertions. The few stand-alone items with four or more assertions were treated as item clusters to take into account the conditional dependencies.

*Figure 1. Directed Graph of the Science IRT Model*




### 5.1.2 Item Response Function

The item response functions of the stand-alone assertions are modeled with a unidimensional model. For the grouped assertions, like in unidimensional models, different parametric forms can be assumed for the conditional probability of obtaining a score of $z_{ij}$. For binary data, the Rasch testlet model (Wang & Wilson, 2005) is defined as

$$P(z_{ij}|\theta_j, u_{jg}; b_i) = \frac{\exp(\theta_j + u_{jg} - b_i)}{1 + \exp(\theta_j + u_{jg} - b_i)} \quad (3)$$

The IRT model for science does not include item discrimination parameters. However, the same model structure as presented in Figure 1 could be employed with discrimination parameters included in Equations 2 and 3. Furthermore, only models for binary data are considered. Assertions are always binary because they are either true or false. Nevertheless, the model could easily accommodate polytomous responses by using the same response function that is incorporated in unidimensional models for polytomous data.

### 5.1.3 Multigroup Model

The item bank for science was calibrated concurrently using all the items administered in any of the states that collaborate with AIR on their new science assessments. In the calibration, each state was treated as a population of students or group. Overall group differences were taken into account by allowing a group-specific distribution of the overall proficiency variable $\theta$. Specifically, for every student *j* belonging to group *k*, *k* = 1, …, *K*, a normal distribution was assumed,

$$\theta_j \sim N\left(\mu_k, \sigma_k^2\right),$$

where $\mu_k$ and $\sigma_k^2$ are the mean and variance of a normal distribution. The mean of the reference distribution (*k* = 1) was set to 1 to identify the model. For each of the nuisance variables $u_g$, a common variance parameter across groups was assumed, and the means were set to 0 in order to identify the model,

$$u_{jg} \sim N\left(0, \sigma_{u_g}^2\right).$$

## 5.2 ITEM CALIBRATION

### 5.2.1 Estimation

A separate IRT model was fit for each grade band. The parameters of the IRT model were estimated using the marginal maximum likelihood (MML) method. In the MML method, the latent proficiency variable $\theta_j$ and the vector of nuisance parameters $\boldsymbol{u}_j$ for each student $j$ are treated as random effects and integrated out to obtain the marginal log likelihood corresponding to the observed response pattern $\boldsymbol{z}_j$ for student $j$,

$$\ell_j = \log \int \int P\left(\boldsymbol{z}_j | \theta_j, \boldsymbol{u}_j\right) N\left(\theta_j | \mu_k, \sigma_k^2\right) N\left(\boldsymbol{u}_j | \boldsymbol{0}, \boldsymbol{\Sigma}\right) d\boldsymbol{u}_j d\theta_j,$$

where $\boldsymbol{\Sigma}$ is a diagonal matrix with diagonal elements $\sigma_{u_k}^2$. Across all students and groups, the overall log likelihood to be maximized with respect to the vector $\boldsymbol{\gamma}$ of all model parameters (item difficulty parameters, and the mean and variance parameters of the latent variables) is

$$\ell(\boldsymbol{\gamma}) = \sum_k \sum_{j \in k} \ell_j.$$

Even though the number of latent variables in the equation above is very high, the curse of dimensionality can be avoided because the integration over the high-dimensional latent $(\theta, \boldsymbol{u})$ space can be carried out as a sequence of computations in two-dimensional space $(\theta, \boldsymbol{u_g})$ (Gibbons & Hedeker, 1992; Rijmen, 2010).

The item bank was calibrated in 2018 after the 2018 science test administrations concluded, and it was recalibrated in 2019 following the 2019 test administrations. The scores reported in 2019 were computed using the 2019 parameters because Rhode Island and Vermont report scores after the testing window closes (with no immediate score reporting). The 2019 parameters will also be used for the 2020 test administration. Because the calibration sequence was somewhat different between 2018 and 2019, the calibration sequence for both years is presented in detail here.

The IRT models were fitted using the BNL (Bayesian networks with logistic regression) suite of Matlab functions (Rijmen, 2006) and flexMIRT (Cai, 2017). The resulting parameters from BNL were used as starting values for flexMIRT, to reduce the estimation time for flexMIRT. The flexMIRT estimates were taken to be the operational parameters, except for the middle school items calibrated in 2018 during the core calibration (see the following section on the 2018 calibration sequence). For the 2018 core calibration of middle school items, flexMIRT did not converge after several weeks, and the estimates obtained from BNL were used as operational parameters. Note that the parameter estimates were very similar across software packages.

*Table 38. Groups per Grade for the Core Calibration*

| Group | Elementary School | Middle School | High School |
|---|---|---|---|
| **Connecticut** | X | X | X |
| **Hawaii** | X | X | X |
| **New Hampshire** | X | X | X |
| **Rhode Island** | X | X | X |
| **Utah Grade 6** | | X | |
| **Utah Grade 7** | | X | |
| **Utah Grade 8** | | X | |
| **Vermont** | X | X | X |
| **West Virginia** | X | X | |

### 5.2.2 2018 Calibration Sequence

Table 38 provides an overview of the groups per grade for the 2018 calibration.

Items were calibrated in three steps for two reasons. First, the rubric validations for some states took place at a later date, and the student responses for the items owned by those states could not be included in the first round of calibrations without jeopardizing the reporting schedule of the two states with operational field tests. (Those two states did not have any of the items with late rubric validation in their item pool.) Second, to divide the very large set of items (and assertions) into more manageable pieces, a separate calibration was carried out for two states with many items administered only in those states. Specifically, the following sequence of calibrations was carried out:

1. Core calibration. The core calibration was performed on the following:
   a. All the item responses of New Hampshire and West Virginia. These states administered items from the following (as described in the bank sharing matrix in Table 39):
      i. AIRCore
      ii. Connecticut
      iii. Hawaii
      iv. Rhode Island
      v. Vermont
      vi. Utah
      vii. West Virginia

A more detailed overlap of the common items at the time of the 2018 calibration was given in Section 3.2.1 (see Table 14 through Table 16).

b. All the item responses of Connecticut, Rhode Island, and Vermont, except for the responses to Oregon and Wyoming items. These states administered items from the following:

   i. AIRCore

   ii. Connecticut

   iii. Hawaii

   iv. Rhode Island

   v. Vermont

   vi. Utah

   vii. West Virginia

   viii. Wyoming (items were treated as not administered; responses were replaced by missing code)

   ix. Oregon (items were treated as not administered; responses were replaced by missing code)

c. Item responses from Hawaii to items also administered in another state (Hawaii items were used in Hawaii, Connecticut, Rhode Island, Vermont, and West Virginia).

d. Item responses from Utah to items also administered in another state (Utah items were used in Utah, Connecticut, Rhode Island, Vermont, and West Virginia). Utah tested middle school students only but included every grade in middle school. One-third of students were selected at random to balance the large population size for Utah.

*Table 39. State Sharing Matrix*

| Source Bank | CT | HI | MSSA | NH | OR | UT | WV | WY |
|---|---|---|---|---|---|---|---|---|
| **AIRCore** | X | X | X | X | X | | X | X |
| **Connecticut** | X | | X | | | | X | |
| **Hawaii** | X | X | X | | | | X | |
| **MSSA** | X | | X | | | | X | |
| **Oregon** | X | | X | | X | | | |
| **Utah** | X | | X | | | X | X | |
| **West Virginia** | X | | X | | | | X | |
| **Wyoming** | X | | X | | | | | X |

*Note.* The core calibration provided parameters for all items used in New Hampshire and West Virginia.

2. Calibration of state-specific items.

   Both Hawaii and Utah had a substantial proportion of items that were only administered in Utah and Hawaii, respectively. Hawaii has both Hawaii and AIRCore items in common with the states of the core calibration (Hawaii only administered Hawaii and AIRCore items); Utah has only Utah items in common (Utah only administered Utah items). The parameters for the unique Hawaii items depend only on responses from Hawaii students, and the parameters for the unique Utah items depend only on responses from Utah students. For both states, the state-specific items were calibrated through a separate calibration based on the state data only, with the items in common with the core states mentioned in step 1 anchored to the estimates from step 1. These calibrations were done separately for each group, under a single-group IRT model. The mean and variance of the groups were fixed to the estimated mean and variance from core calibration 1.

3. Calibration of states with late rubric validation.

   Oregon and Wyoming items were administered in some of the states from the core calibration (Connecticut, Rhode Island, and Vermont) but could not be calibrated in step 1 because of their late rubric validation dates. In a later stage, items from Oregon and Wyoming were calibrated by:

   a. adding Oregon and Wyoming student responses to the core calibration;
   b. keeping the responses from Connecticut, Rhode Island, and Vermont to Wyoming and Oregon items (as opposed to treating them as missing in step 1);
   c. removing the responses from the states that did not administer Oregon or Wyoming items (as the item parameters for the Oregon and Wyoming items did not depend on the students from these states) (The removed states were Hawaii, New Hampshire, Utah, and West Virginia.); and
   d. fixing the parameters of all other items to the values obtained in step 1, as well as the group means and standard deviations that were estimated in step 1.

### 5.2.3 2019 Calibration Sequence

The calibration was performed in two steps. First, all items in operational use in 2019 for which 1,000 or more student responses were observed were calibrated (for all but three items, there were 1,500 or more student responses). In this step, the data of states with an operational test only were included. Table 40 provides an overview of the groups per grade for this first calibration. All students who attempted the test were included in the calibration. The assertions of skipped items were scored as incorrect. Note that only Rhode Island allowed students to skip items. There were nine items administered as operational items in 2019 for which the sample size was smaller than 1,000 students, out of a total of 438 items.

Table 41 through Table 43 present the number of operational item clusters and stand-alone items that were shared between the item pools of any two states. The numbers below the diagonal represent the numbers for all the operational items administered, and the numbers above the diagonal represent the number of common operational items at the time of the 2019 calibration. The shaded diagonal elements represent the number of operational items that were administered only in the given state (in parentheses, the number of unique operational items at the time of calibration).

Table 41 presents the results for elementary schools, Table 42 presents the results for middle schools, and Table 43 presents the results for high schools. The numbers at operational administration are slightly different from the numbers at calibration because items with a sample size smaller than 1,000 students were excluded from the calibration.

*Table 40. Groups per Grade for the Calibration of Operational Items*

| Group | Elementary School | Middle School | High School |
|---|---|---|---|
| **Connecticut** | X | X | X |
| **New Hampshire** | X | X | X |
| **Oregon** | X | X | X |
| **Rhode Island** | X | X | X |
| **Vermont** | X | X | X |
| **West Virginia** | X | X | |

*Table 41. Number of Common Elementary School Operational Items Administered and Calibrated in Spring 2019*

| | State | Connecticut | MSSA | New Hampshire | Oregon | West Virginia |
|---|---|---|---|---|---|---|
| Cluster | **CT** | **1 (1)** | 44 | 24 | 42 | 55 |
| | **MSSA** | 44 | **0 (0)** | 17 | 37 | 41 |
| | **NH** | 24 | 17 | **0 (0)** | 14 | 27 |
| | **OR** | 42 | 37 | 14 | **0 (0)** | 41 |
| | **WV** | 55 | 41 | 27 | 41 | **1 (1)** |
| Stand-Alone | **CT** | **3 (3)** | 34 | 26 | 30 | 47 |
| | **MSSA** | 34 | **0 (0)** | 20 | 23 | 32 |
| | **NH** | 26 | 20 | **0 (0)** | 14 | 25 |
| | **OR** | 30 | 23 | 14 | **0 (0)** | 25 |
| | **WV** | 47 | 32 | 25 | 25 | **1 (1)** |
| Grade Band Total | **CT** | **4 (4)** | 78 | 50 | 72 | 102 |
| | **MSSA** | 78 | **0 (0)** | 37 | 60 | 73 |
| | **NH** | 50 | 37 | **0 (0)** | 28 | 52 |
| | **OR** | 72 | 60 | 28 | **0 (0)** | 66 |
| | **WV** | 102 | 73 | 52 | 66 | **2 (2)** |

*Table 42. Number of Common Middle School Operational Items Administered and Calibrated in Spring 2019*

| | State | Connecticut | MSSA | New Hampshire | Oregon | West Virginia |
|---|---|---|---|---|---|---|
| Cluster | **CT** | **3 (3)** | 26 | 24 | 54 | 92 |
| | **MSSA** | 26 | **0 (0)** | 11 | 14 | 21 |
| | **NH** | 24 | 11 | **1 (1)** | 9 | 18 |
| | **OR** | 54 | 14 | 9 | **2 (2)** | 56 |
| | **WV** | 92 | 21 | 18 | 56 | **12 (4)** |
| Stand-Alone | **CT** | **0 (0)** | 42 | 26 | 34 | 50 |
| | **MSSA** | 42 | **0 (0)** | 25 | 30 | 37 |
| | **NH** | 26 | 25 | **0 (0)** | 16 | 21 |
| | **OR** | 34 | 30 | 16 | **1 (0)** | 29 |
| | **WV** | 50 | 37 | 21 | 29 | **0 (0)** |
| Grade Band Total | **CT** | **3 (3)** | 68 | 50 | 88 | 142 |
| | **MSSA** | 68 | **0 (0)** | 36 | 44 | 58 |
| | **NH** | 50 | 36 | **1 (1)** | 25 | 39 |
| | **OR** | 88 | 44 | 25 | **3 (2)** | 85 |
| | **WV** | 142 | 58 | 39 | 85 | **12 (4)** |

*Table 43. Number of Common High School Operational Items Administered and Calibrated in Spring 2019*

| | State | Connecticut | MSSA | New Hampshire | Oregon | West Virginia |
|---|---|---|---|---|---|---|
| **Cluster** | **CT** | **5 (5)** | 33 | 22 | 30 | 0 |
| | **MSSA** | 33 | **0 (0)** | 20 | 31 | 0 |
| | **NH** | 22 | 20 | **2 (2)** | 15 | 0 |
| | **OR** | 30 | 31 | 15 | **1 (1)** | 0 |
| | **WV** | 0 | 0 | 0 | 0 | **0 (0)** |
| **Stand-alone** | **CT** | **0 (0)** | 39 | 27 | 40 | 0 |
| | **MSSA** | 39 | **2 (2)** | 23 | 32 | 0 |
| | **NH** | 27 | 23 | **0 (0)** | 20 | 0 |
| | **OR** | 40 | 32 | 20 | **4 (4)** | 0 |
| | **WV** | 0 | 0 | 0 | 0 | **0 (0)** |
| **Grade Band Total** | **CT** | **5 (5)** | 72 | 49 | 70 | 0 |
| | **MSSA** | 72 | **2 (2)** | 43 | 63 | 0 |
| | **NH** | 49 | 43 | **2 (2)** | 35 | 0 |
| | **OR** | 70 | 63 | 35 | **5 (5)** | 0 |
| | **WV** | 0 | 0 | 0 | 0 | **0 (0)** |

In the second step, the field-test items were calibrated. The calibration included the operational items that were calibrated in step 1, and the field-test items across all states that administered field-test items. All students who attempted at least one field-test item were included in the calibration. Table 44 provides an overview of the groups per grade for calibration of the field-test items.

*Table 44. Groups per Grade for the Calibration of Field-Test Items*

| Group | Elementary School | Middle School | High School |
|---|---|---|---|
| **Connecticut** | X | X | X |
| **Hawaii** | X | X | X |
| **Idaho** | X | X | |
| **New Hampshire** | X | X | X |
| **Oregon** | X | X | X |
| **Rhode Island** | X | X | X |
| **Vermont** | X | X | X |
| **West Virginia** | X | X | |
| **Wyoming** | X | X | X |

## 5.3 Linking the 2018 Scale to the 2019 Scale

The item parameter estimates obtained from the 2018 student responses were highly correlated with the item parameters obtained from the 2019 student responses. For the item difficulties, the correlation between the 2018 and 2019 estimates was 0.993 for elementary school, 0.986 for middle school, and 0.994 for high school. For the standard deviations of the item clusters, these correlations were 0.971, 0.972, and 0.964, respectively. These high correlations indicate that items functioned similarly in 2018 and 2019. Nevertheless, item parameters from separate calibrations cannot be directly compared because the scale of an item response theory (IRT) model is not determined. In the multigroup Rasch testlet model, the only scale indeterminacy is the origin of the scale. The models can be identified by setting the mean of the overall proficiency variable $\theta$ to 0 for the reference distribution. As a result, the 2018 and 2019 $\theta$ and item parameters are on the same scale except for an overall shift parameter *B*. Specifically, the 2018 scale can be linked to the 2019 scale as follows:

$$P\left(z_{ij}|\theta_{j\ 2018}, u_{jg}; b_{i\ 2018}\right) = \frac{\exp\left(\theta_{j\ 2018} + u_{jg} - b_{i\ 2018}\right)}{1 + \exp\left(\theta_{j\ 2018} + u_{jg} - b_{i\ 2018}\right)}$$

$$= \frac{exp(\theta_{j\ 2018} + B + u_{jg} - b_{i\ 2018} - B)}{1 + exp(\theta_{j\ 2018} + B + u_{jg} - b_{i\ 2018} - B)}$$

$$= \frac{exp(\theta_{j\ 2019} + u_{jg} - b_{i\ 2019})}{1 + exp(\theta_{j\ 2019} + u_{jg} - b_{i\ 2019})}.$$

Because $\theta_{j\ 2019} = \theta_{j\ 2018} + B$, the population means of $\theta$ must be transformed accordingly,

$$\theta_{j\ 2019} \sim N\ (\mu_{k\ 2018} + B, \sigma_k^2)$$

$$\theta_{j\ 2018} \sim N\ (\mu_{k\ 2018}, \sigma_k^2).$$

Item parameters based on 2018 student responses can be expressed on the 2019 scale by adding the constant *B* to the 2018 item parameter. The 2018 parameters were expressed on the 2019 scale for items that were part of the pool in both 2018 and 2019 but not administered in any states in 2019 (13 items), and for items that were administered in 2019, but the number of student responses from the 2019 assessments was lower than 1,000 (nine items).

All items that were operational in 2019 were also administered in 2018. Therefore, the shift parameter *B* can be estimated from a separate calibration of the items operational in 2019 using the 2019 student responses (of the six operational states) but with the item parameters fixed to the estimates obtained from the 2018 calibrations. By fixing (a subset of) the item parameters, the model is identified so that the means and variances of $\theta$ can be estimated for all groups. *B* can be obtained by equating the overall mean of $\theta$ across all groups for the 2019 student response data from the free calibration (2019 overall mean expressed on the 2019 scale) to the overall mean of $\theta$ across all groups for the 2019 student response data from the calibration with items anchored to their 2018 parameters values (2019 overall mean expressed on the 2018 scale):

$$\frac{1}{K}\sum_{k=1}^{K}\mu_{k\ 2019} = \frac{1}{K}\sum_{k=1}^{K}(\mu_{k\ 2018} + B),$$

Therefore, an estimate of *B* can be obtained as

$$\hat{B} = \frac{1}{K}\sum_{k=1}^{K}(\hat{\mu}_{k\ 2019} - \hat{\mu}_{k\ 2018}).$$

The estimated means of $\theta$ under both the free and anchored calibrations, as well as the number of students per state, are presented in Table 45. The table also presents the overall means and estimated shift parameter $B$. Note that the parameters for three items were not anchored but freely estimated together with the means and variances in the anchored calibration. The reason for not treating these items as common items across the 2018 and 2019 administrations was that they had an omit rate of 4% or higher for the last item interaction in the 2018 administration in at least one state; in 2019, these interactions could no longer be omitted because all interactions of an item needed to be responded to in states where skipping was not allowed (these were all states except Rhode Island). So, out of an abundance of caution, these three items were not anchored to their 2018 parameter values.

*Table 45. Estimated Latent Means and Number of Students per State*

| Group | Elementary School | | | Middle School | | | High School | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\hat{\mu}_{k\ 2019}$ | $\hat{\mu}_{k\ 2018}$ | *N* | $\hat{\mu}_{k\ 2019}$ | $\hat{\mu}_{k\ 2018}$ | *N* | $\hat{\mu}_{k\ 2019}$ | $\hat{\mu}_{k\ 2018}$ | *N* |
| **Connecticut** | 0.0000 | 0.0518 | 38,549 | 0.0000 | 0.0234 | 39,347 | 0.0000 | 0.1443 | 37,616 |
| **New Hampshire** | 0.0631 | 0.1083 | 13,187 | 0.0940 | 0.1108 | 12,060 | 0.0798 | 0.2278 | 11,385 |
| **Oregon** | -0.0101 | 0.0096 | 44,989 | 0.0028 | 0.0156 | 42,043 | -0.0383 | 0.1030 | 41,630 |
| **Rhode Island** | -0.0312 | 0.0142 | 10,751 | -0.1044 | -0.0692 | 10,306 | -0.2261 | -0.0879 | 9,612 |
| **Vermont** | 0.1069 | 0.1504 | 6,017 | 0.0781 | 0.1133 | 5,894 | 0.0179 | 0.1545 | 5,332 |
| **West Virginia** | -0.1970 | -0.1529 | 19,540 | -0.3012 | -0.2783 | 19,043 | – | – | – |
| | $\frac{1}{K}\sum_{k=1}^{K}\hat{\mu}_{k\ 2019}$ | $\frac{1}{K}\sum_{k=1}^{K}\hat{\mu}_{k\ 2018}$ | $\hat{B}$ | $\frac{1}{K}\sum_{k=1}^{K}\hat{\mu}_{k\ 2019}$ | $\frac{1}{K}\sum_{k=1}^{K}\hat{\mu}_{k\ 2018}$ | $\hat{B}$ | $\frac{1}{K}\sum_{k=1}^{K}\hat{\mu}_{k\ 2019}$ | $\frac{1}{K}\sum_{k=1}^{K}\hat{\mu}_{k\ 2018}$ | $\hat{B}$ |
| **Overall** | -0.0114 | 0.0303 | -0.0416 | -0.0385 | -0.0141 | -0.0244 | -0.0333 | 0.1083 | -0.1417 |

The estimated parameters of all science items (AIRCore and MOU), as well as the estimated group means and variances, are presented in Appendix B. The appendix contains the results for both the 2018 and 2019 calibrations. For the 2018 calibrations, the items parameters are presented for both the original 2018 scale and after linking the 2018 parameters to the 2019 scale. Figures in Appendix B display the histogram of the difficulty parameters for elementary school, middle school, and high school for all items that are part of the Rhode Island and Vermont operational pool. The figures also display the student proficiency distributions. The distributions of the difficulty parameter are slightly less difficult than the student proficiency distribution in grade 5 and grade 8. The grade 11 items are slightly more difficult than the student proficiency in general.

# 6. SCORING

## 6.1 MAXIMUM LIKELIHOOD FUNCTION

Student scores are obtained by marginalizing out the nuisance dimensions $u_j$ from the likelihood of the observed response pattern $z_j$ for student $j$,

$$\ell_i(\theta_j) = log \int_{u_j} P(z_j|\theta_j, u_j)\, N(u_j|0, \Sigma) du_j,$$

and maximizing this marginalized likelihood function for $\theta_j$. The marginal maximum likelihood estimation (MMLE) estimator is a hybrid between the expected a posteriori (EAP) estimator (by marginalizing out the nuisance dimensions) and the MLE estimator (by maximizing the resulting marginal likelihood for $\theta$). The marginal likelihood is maximized with respect to $\theta$ using the Newton Raphson method.

The proposed model reduces to the unidimensional Rasch model when the nuisance variances are zero for all $g$. Likewise, the proposed MMLE is equivalent to the MLE of the unidimensional Rasch model when all the nuisance variances are zero. This can be shown by using the variable transformation $v = \mathbf{\Sigma}^{-\frac{1}{2}} u$. Then we have

$$\int_{u_j} P(z_j|\theta_j, u_j)\, N(u_j|0, \Sigma) du_j = \int_{v_j} P\left(z_j \middle| \theta_j, \Sigma^{\frac{1}{2}} v_j\right) N(v_j|0, I) dv_j.$$

If $\sigma^2_{u_g} = 0$ for all g, then

$$\int_{u_j} P(z_j|\theta_j, u_j)\, N(u_j|0, \Sigma) du_j = P(z_j|\theta_j),$$

which is the likelihood under the unidimensional Rasch model.

## 6.2 DERIVATIVE

The marginal log likelihood function based on the IRT model with one overall dimension and one nuisance dimension for each grouping of assertions can be written as

$$l(\theta) = \sum_{i \in SA} log\big(P(z_i|\theta)\big) + \sum_{g=1}^{G} log\left\{\int Exp\left[\sum_{i \in g} log\left(P(z_{ig}|\theta, u_g)\right)\right] N\left(u_g \middle| 0, \sigma^2_{u_g}\right) du_g\right\}.$$

The first derivative of the marginal log likelihood function with respect to $\theta$ is

$$\frac{dl(\theta)}{d\theta} = \sum_{i\in SA} \frac{\frac{dP(z_i|\theta)}{d\theta}}{P(z_i|\theta)} + \sum_{g=1}^{G} \frac{\int \left\{ \text{Exp}\left[\sum_{i\in g} \log\left(P\left(z_{ig}|\theta,u_g\right)\right)\right] \left( \sum_{i\in g} \frac{\frac{dP\left(z_{ig}|\theta,u_g\right)}{d\theta}}{P\left(z_{ig}|\theta,u_g\right)} \right) N\left(u_g \middle| 0,\sigma_{u_g}^2\right) \right\} du_g}{\int \left\{ \text{Exp}\left[\sum_{i\in g} \log\left(P\left(z_{ig}|\theta,u_g\right)\right)\right] N\left(u_g \middle| 0,\sigma_{u_g}^2\right) \right\} du_g}$$

and the second derivative of the marginal log likelihood function with respect to $\theta$ is

$$\frac{d^2 l(\theta)}{d\theta^2} = \sum_{i\in SA} \left[ \frac{\frac{d^2 P(z_i|\theta)}{d\theta^2}}{P(z_i|\theta)} - \left( \frac{\frac{d\,P(z_i|\theta)}{d\theta}}{P(z_i|\theta)} \right)^2 \right]$$
$$+ \sum_{g=1}^{G} \frac{\int \text{Exp}\left[\sum_{i\in g} \log\left(P\left(z_{ig}|\theta,u_g\right)\right)\right] \left( \sum_{i\in g} \frac{\frac{d\,P\left(z_{ig}|\theta,u_g\right)}{d\theta}}{P\left(z_{ig}|\theta,u_g\right)} \right)^2 N\left(u_g \middle| 0,\sigma_{u_g}^2\right) du_g}{\int \left\{ \text{Exp}\left[\sum_{i\in g} \log\left(P\left(z_{ig}|\theta,u_g\right)\right)\right] N\left(u_g \middle| 0,\sigma_{u_g}^2\right) \right\} du_g}$$
$$+ \sum_{g=1}^{G} \frac{\int \text{Exp}\left[\sum_{i\in g} \log\left(P\left(z_{ig}|\theta,u_g\right)\right)\right] \left( \sum_{i\in g} \left[ \frac{\frac{d^2\,P\left(z_{ig}|\theta,u_g\right)}{d\theta^2}}{P\left(z_{ig}|\theta,u_g\right)} - \left( \frac{\frac{d\,P\left(z_{ig}|\theta,u_g\right)}{d\theta}}{P\left(z_{ig}|\theta,u_g\right)} \right)^2 \right] \right) N\left(u_g \middle| 0,\sigma_{u_g}^2\right) du_g}{\int \left\{ \text{Exp}\left[\sum_{i\in g} \log\left(P\left(z_{ig}|\theta,u_g\right)\right)\right] N\left(u_g \middle| 0,\sigma_{u_g}^2\right) \right\} du_g}$$
$$- \sum_{g=1}^{G} \left\{ \frac{\int \text{Exp}\left[\sum_{i\in g} \log\left(P\left(z_{ig}|\theta,u_g\right)\right)\right] \left( \sum_{i\in g} \frac{\frac{d\,P\left(z_{ig}|\theta,u_g\right)}{d\theta}}{P\left(z_{ig}|\theta,u_g\right)} \right) N\left(u_g \middle| 0,\sigma_{u_g}^2\right) du_g}{\int \left\{ \text{Exp}\left[\sum_{i\in g} \log\left(P\left(z_{ig}|\theta,u_g\right)\right)\right] N\left(u_g \middle| 0,\sigma_{u_g}^2\right) \right\} du_g} \right\}^2$$

Based on the above equations, we need only to define the ratios of the first and second derivatives of the item response probabilities with respect to $\theta$ to the response probabilities. For the Rasch testlet model, these are obtained as

$$p_i = P(z_i = 1|\theta) = \frac{Exp(\theta - b_i)}{1 + Exp(\theta - b_i)},\ q_i = P(z_i = 0|\theta) = 1 - p_i,$$

and

$$p_{ig} = P(z_{ig} = 1|\theta, u_g) = \frac{\mathrm{Exp}(\theta + u_g - b_i)}{1 + \mathrm{Exp}(\theta + u_g - b_i)},\ q_{ig} = P(z_{ig} = 0|\theta, u_g) = 1 - p_{ig}.$$

Therefore, we have,

$$\frac{\frac{dp_i}{d\theta}}{p_i} = q_i\ ,\ \frac{\frac{dq_i}{d\theta}}{q_i} = -p_i,$$

$$\frac{\frac{dp_{ig}}{d\theta}}{p_{ig}} = q_{ig}\ ,\ \frac{\frac{dq_{ig}}{d\theta}}{q_{ig}} = -p_{ig},$$

$$\frac{\frac{d^2 p_i}{d\theta^2}}{p_i} - \left(\frac{\frac{d\,p_i}{d\theta}}{p_i}\right)^2 = -p_i q_i,$$

$$\frac{\frac{d^2 q_i}{d\theta^2}}{q_i} - \left(\frac{\frac{d\,q_i}{d\theta}}{q_i}\right)^2 = -p_i q_i,$$

$$\frac{\frac{d^2 p_{ig}}{d\theta^2}}{p_{ig}} - \left(\frac{\frac{d\,p_{ig}}{d\theta}}{p_{ig}}\right)^2 = -p_{ig} q_{ig}, \text{ and}$$

$$\frac{\frac{d^2 q_{ig}}{d\theta^2}}{q_{ig}} - \left(\frac{\frac{d\,q_{ig}}{d\theta}}{q_{ig}}\right)^2 = -p_{ig} q_{ig}.$$

## 6.3 EXTREME CASE HANDLING

As with the MLE, the MMLE is not defined for zero and perfect scores. These cases are handled by assigning the lowest obtainable theta (LOT) scores and highest obtainable theta (HOT) scores, respectively. Table 46 contains the LOT and HOT values for each grade.

## 6.4 STANDARD ERRORS OF ESTIMATE

The standard error of measurement (SEM) of the MMLE score estimate is:

$$SEM(\hat{\theta}_{MMLE}) = \frac{1}{\sqrt{I(\hat{\theta}_{MMLE})}}$$

where $I(\hat{\theta}_{MMLE})$ is the observed information evaluated at $\hat{\theta}_{MMLE}$. The observed information is calculated as $I(\theta^2) = -\frac{d^2 l(\theta)}{d\theta^2}$ where $\frac{d^2 l(\theta)}{d\theta^2}$ is defined in the previous section on derivatives. Note that the calculation of the standard error of estimate depends on the unique set of items that each student answers and their estimate of $\theta$. Different students have different standard errors of measurement, even if they have the same raw score and/or theta estimate. Standard errors are truncated at 1 for the overall science scores and truncated at 1.4 for the discipline scores.

Standard errors for MMLE estimates truncated at the LOT (HOT) are computed by evaluating the observed information at the MMLE before truncation. For all incorrect or all correct answers, the reported standard are set at the truncation value for the standard error.

## 6.5 STUDENT-LEVEL SCALE SCORE

At the student level, scale scores are computed for

1. Overall Science;
2. Life Sciences;
3. Physical Sciences; and
4. Earth and Space Sciences.

Scores are computed using the MMLE method outlined in this report, with all items for overall science or only items within the given discipline. Scores are truncated on the "theta" scale at the LOT and HOT values specified in Table 46, which correspond to values of the estimated mean minus/plus four times the estimated standard deviation of $\theta$.

The reporting scales will be a linear transformation of the theta scales:

$$SS = a * \hat{\theta}_{MMLE} + b$$

Where $a$ and $b$ are the slope and intercept of the linear transformation that transforms $\hat{\theta}_{MMLE}$ to the reporting scale (see Table 46). The standard error of estimate for the estimated scale score is obtained as:

$$SEM_{SS} = a * SEM_{\hat{\theta}_{MMLE}}.$$

In 2019, the reporting scale has a range of 120 points, from 1 to 120. The slope $a$ and intercept $b$ were chosen so that the center of the reporting scale of each grade (SS = 60) is centered at the proficiency cut and has a standard deviation of 15. Because a scale was required during standard setting, before the proficiency cut was known, the scale is established in two steps. In the first step, the scale was established based on a tentative cut where 40% of the population would be proficient, corresponding to how proficiency cuts were set in New Hampshire and West Virginia across grades in 2018. Specifically, for grade 5, the slope $a$ is obtained as:

$$\begin{aligned} SS &= 15\theta^* + b \\ &= 15\frac{\theta}{\hat{\sigma}_\theta} + b \\ &= a\theta + b, \end{aligned}$$

where the second line stems from transforming theta into a variable with a standard deviation of 1, $\theta^* = \frac{\theta}{\hat{\sigma}_\theta}$. Subsequently, the intercept $b$ is obtained by equating the center of the scale (SS = 60) to the linear transformation of the tentative cut score on the theta scale,

$$SS = 60 = a\hat{\theta}_{tentative\_cut} + b$$

$$B = 60 - a\hat{\theta}_{tentative\_cut}$$

For grades 8 and 11, the slope and intercept can also be derived in a similar fashion.

After the 2019 standard setting, the final proficiency cut was set at 63 on the proposed scale for all three grades (detailed standard-setting results are presented in Volume 3 of this technical report). In order to center the reporting scale around the final cut, the scale was translated by minus 3, the difference between the tentative and final cuts expressed on the reporting scale. Table 46 presents the intercept and slope, as well as the LOT, HOT, Lowest of Scale Score (LOSS), and Highest of Scale Score (HOSS) values that were used for the final 2019 reporting scale. The scale score distribution is reported in Appendix C for overall science, and in Appendix D for the disciplines.

*Table 46. Reporting Scale Linear Transformation Constants and Theta and Corresponding Scaled-Score Limits for Extreme Ability Estimates (for 2019 $\theta$ scale)*

| Grade | Slope | Intercept | Lowest of Theta (LOT) | Highest of Theta (HOT) | Lowest of Scale Score (LOSS) | Highest of Scale Score (HOSS) |
|---|---|---|---|---|---|---|
| **5** | 16.677 | 52.196 | -3.06 | 4.06 | 1 | 120 |
| **8** | 17.001 | 53.266 | -3.07 | 3.92 | 1 | 120 |
| **11** | 18.084 | 57.041 | -3.09 | 3.48 | 1 | 120 |

## 6.6 RULES FOR CALCULATING ACHIEVEMENT LEVELS

Achievement levels and corresponding cut scores were set during standard setting in summer 2019. Students are classified into one of four achievement levels, based on their total score. The distribution of achievement levels is summarized in Appendix C. Further, the distribution of scale scores and achievement levels for subgroups described in Section 4.4 are presented in Appendix E.

Table 47 lists the cut scores on the reporting scale metrics for each grade.

*Table 47. Achievement-Level Cut Scores*

| Grade | Cut 1 | Cut 2 | Cut 3 |
|---|---|---|---|
| **5** | 37 | 60 | 72 |
| **8** | 38 | 60 | 74 |
| **11** | 36 | 60 | 71 |

### 6.6.1 Strengths and Weaknesses for Disciplines Relative to Proficiency Cut Score

Discipline-level classifications are computed to classify student achievement levels for each of the science disciplines. The classification rules are:

- if $(\hat{\theta}_{discipline} < \theta_{proficient} - 1.5 * SEM\left(\hat{\theta}_{\text{discipline}}\right))$, then achievement is classified as *Below Mastery*;
- if $(\theta_{proficient} - 1.5 * SEM\left(\hat{\theta}_{\text{discipline}}\right) \leq \hat{\theta}_{discipline} < \theta_{proficient} + 1.5 * SEM\left(\hat{\theta}_{\text{discipline}}\right))$, then achievement is classified as *At/Near Mastery*; and
- if $(\hat{\theta}_{discipline} \geq \theta_{proficient} + 1.5 * SEM\left(\hat{\theta}_{\text{discipline}}\right))$, then achievement is classified as *Above Mastery*,

where $\theta_{Proficient}$ is the proficiency cut score of the overall test. Standard errors are truncated at 1.4. The LOT is always classified as *Below Mastery*, and the HOT is always classified as *Above Mastery*.

## 6.7 Disciplinary Core Ideas and Performance Expectation-Level Reporting

### 6.7.1 Relative to Overall Achievement

For aggregated units (classrooms, schools, districts), there is reporting at levels below the science discipline level. In 2018–2019, reports were provided at the level of disciplinary core ideas (DCI). The method for reporting at levels below the science discipline level is based on the use of residuals. The equations are presented first for DCIs.

For each assertion *i*, the residual between observed and expected score for each student *j* is defined as

$$\delta_{ij} = z_{ij} - E(z_{ij}).$$

The expected score is computed for a student's estimated overall ability. For the assertions clustered within an item, the expected score is marginalized over the nuisance dimensions for the assertions clustered within an item,

$$E(z_{ijg} = 1; \theta_{j,overall}, \tau_i) = \int P(z_{ijg} = 1 | u_{jg};\ \theta_{j,overall}, \tau_i) N(u_{jg}) du_{jg},$$

where $\tau_i$ is the vector of parameters for assertion *i* (e.g., for the Rasch testlet model, $\tau_i = b_i$), and $P(z_{ijg} = 1 | u_{jg};\ \theta_{j,overall}, \tau_i)$ is defined in Section 6.2. Next, residuals are aggregated over assertions within students,

$$\delta_{jDCI} = \frac{\sum_{i \in DCI} \delta_{ij}}{n_{jDCI}}.$$

and over students of the group on which is reported,

$$\bar{\delta}_{DCIg} = \frac{1}{n_g}\sum_{j\in g}\delta_{jDCI}\ ,$$

where $n_{jDCI}$ is the number of assertions related to the DCI for student $j$, and $n_g$ is the number of students in a group assessed on the DCI. If a student did not see any items on a DCI, the student is not included in the $n_g$ count for the aggregate. The standard error of the average residual is computed as

$$SEM(\bar{\delta}_{DCIg}) = \sqrt{\frac{1}{n_g(n_g-1)}\sum_{j\in g}\left(\delta_{jDCI} - \bar{\delta}_{PDCIg}\right)^2}.$$

A statistically significant difference from zero in these aggregates is evidence that a class, teacher, school, or district is more effective (if $\bar{\delta}_{DCIg}$ is positive) or less effective (negative $\bar{\delta}_{DCIg}$) in teaching a given DCI.

We do not suggest the direct reporting of the statistic $\bar{\delta}_{DCIg}$; instead, we recommend reporting whether, in the aggregate, a group of students performs better, worse, or as expected on this DCI. In some cases, sufficient information is not available, and that will be indicated, as well.

For target-level strengths/weakness, the following is reported:

- If $\bar{\delta}_{DCIg} \leq -1.5 * SEM\left(\bar{\delta}_{DCIg}\right)$, then achievement is *worse than* on the overall test.
- If $\bar{\delta}_{DCIg} \geq 1.5 * SEM(\bar{\delta}_{DCIg})$, then achievement is *better than* on the overall test.
- Otherwise, achievement is *similar to* the overall test.
- If $SEM(\bar{\delta}_{DCIg}) > 0.2$, data are insufficient.

### 6.7.2 Relative to Proficiency Cut Score

DCI level scores for aggregated units can be computed using the same method as outlined in Section 6.7.1, but with the expected score computed at the theta value corresponding to the proficiency cut score:

$$E\left(z_{ijg} = 1; \theta_{proficiency}, \tau_i\right) = \int P\left(z_{ijg} = 1 | u_{jg};\ \theta_{proficiency}, \tau_i\right) N\left(u_{jg}\right) du_{jg}.$$

The following is reported for DCIs for aggregate units:

- If $\bar{\delta}_{DCIg} \leq -1.5 * SEM\left(\bar{\delta}_{DCIg}\right)$, then achievement is *below* the proficiency cut score.
- If $\bar{\delta}_{DCIg} \geq 1.5 * SEM\left(\bar{\delta}_{DCIg}\right)$, then achievement is *above* the proficiency cut score.
- Otherwise, achievement is *near* the proficiency cut score.
- If $SEM\left(\bar{\delta}_{DCIg}\right) > 0.2$, data are insufficient.

# 7. QUALITY CONTROL PROCEDURES

AIR's quality assurance (QA) procedures are built on two key principles: automation and replication. Certain procedures can be automated, which removes the potential for human error. Procedures that cannot be reasonably automated are replicated by two independent analysts at AIR.

Although the quality of any test is monitored as an ongoing activity, several sources of AIR's quality control system are described here. First, QA reports are routinely generated and evaluated throughout the testing window to ensure that each test is performing as anticipated. Second, the quality of scores is ensured by employing a second independent scoring verification system.

## 7.1 QUALITY ASSURANCE REPORTS

Test monitoring occurs while tests are administered in a live environment to ensure that item behavior is consistent with expectations. This is accomplished using AIR's quality monitoring system that yields item statistics, blueprint match rates, and item exposure rate reports.

### 7.1.1 Item Analysis

The item analysis report is a key check for the early detection of potential problems with item scoring, including incorrect designation of a keyed response or other scoring errors, as well as potential breaches of test security that may be indicated by changes in the difficulty of test items. To examine the performance of test items, this report generates classical item analysis indicators of difficulty and discrimination, including proportion correct and biserial/polyserial correlation, as well as item fit statistics based on the IRT. The report is configurable and can be produced to flag only items with statistics falling outside a specified range or to generate reports based on all items in the pool. For science, statistics reports at the assertion level (which are the units of analysis for science) are currently not yet available. However, our psychometricians compute and monitor classical item statistics at the end of the testing window. As described in Section 4.5, the classical statistics of all operational and field-test science items are presented in Appendix A.

### 7.1.2 Blueprint Match

The QA system generates blueprint match reports at the content standards level and for other content requirements such as strand and affinity groups for science. For each blueprint element, the report indicates the minimum and maximum number of items specified in the blueprint, the number of test administrations in which those specifications were met, the number of administrations in which the blueprint requirements were not met, and, for administrations in which specifications were not met, the number of items by which the requirement was not met.

For all three grades, every test met the blueprint specifications at the level of the science disciplines, which is the lowest content level at which scores for individual students are reported. Some violations did occur at lower content levels, primarily for the Spanish tests due to the limited number of items for which a Spanish version is available. Blueprint match is discussed in detail in Volume 2 of this technical report for both simulated and operational test administrations.

### 7.1.3 Item Exposure Rates

The QA system also generates item exposure reports that allow test items to be monitored for unexpectedly large exposure rates or unusually low item-pool usage throughout the testing window. As with other reports, it is possible to examine the exposure rate for all items or flag items with exposure rates that exceed an acceptable range. Often, item overexposure indicates a blueprint element or combination of blueprint elements that are underrepresented in the item pool and should be targeted for future item development. Such item overexposure is also usually anticipated in the simulation studies used to configure the adaptive algorithm. Details about item exposure rates are discussed in Volume 2 of this technical report.

## 7.2 SCORING QUALITY CHECK

All student test scores are produced using AIR's scoring engine. Before releasing any scores, a second score verification system is used to verify that all test scores match with 100% agreement in all tested grades. This second system is independently constructed and maintained from the main scoring engine and separately estimates scores using the procedures described within this report.

# 8. References

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

Cai, L. (2017). flexMIRT version 3.51: Flexible multilevel multidimensional item analysis and test scoring (computer software). Chapel Hill, NC: Vector Psychometric Group.

Dorans, N.J., & Schmitt, A.P. (1991). Constructed response and differential item functioning: A pragmatic approach (ETS Research Report No. 91–47). Princeton, NJ: Educational Testing Service.

Gibbons, R.D., & Hedeker, D.R. (1992). Full-information item bi-factor analysis. *Psychometrika, 57,* 423–436. doi:10.1007/BF02295430.

Holland, P.W., & Thayer, D.T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.

Koretz, D., & Hamilton, L.S. (2006). Testing for accountability in K–12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531–578). Westport, CT: American Council on Education/Praeger.

Rijmen, F. (2006*). BNL: A Matlab toolbox for Bayesian networks with logistic regression nodes*. (Technical Report). Amsterdam: VU University Medical Center.

Rijmen, F. (2010). Formal relations and empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement, 47,* 361–372. doi:10.1111/j.1745-3984.2010.00118.

Sireci, S.G., Wainer, H., & Thissen, D. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28,* 237–247.

Somes, G.W. (1986). The generalized Mantel Haenszel statistic. *The American Statistician*, 40:106–108.

Wang, W.C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, *29*(2), 126–149. doi:10.1177/0146621604271053.

Yen, W. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187–213.

Zwick, R. (2012). *A Review of ETS Differential Item Functioning Assessment Procedures: Flagging Rules, Minimum Sample Size Requirements, and Criterion Refinement (*ETS Research Report No. 12–08). Princeton, NJ: Educational Testing Service.

# Appendix C

# Distribution of Scale Scores and Achievement Levels

# Distribution of Scale Scores and Achievement Levels

*Table C-1. Combined Scale Score Mean and Standard Deviation by Grade*

| | Grade | | |
|---|---|---|---|
| | **5** | **8** | **11** |
| *Number of Students* | 16,867 | 16,482 | 15,156 |
| Mean Scale Score | 52.37 | 52.12 | 54.01 |
| SD of Scale Score | 16.56 | 16.61 | 16.63 |

*Table C-2. Combined Proportion of Students in Each Achievement Level by Grade*[1]

| **Achievement Level** | Grade | | |
|---|---|---|---|
| | **5** | **8** | **11** |
| *Number of Students* | 16,867 | 16,482 | 15,156 |
| Level 1 | 0.18 | 0.21 | 0.11 |
| Level 2 | 0.48 | 0.45 | 0.55 |
| Level 3 | 0.22 | 0.24 | 0.18 |
| Level 4 | 0.12 | 0.10 | 0.16 |

[1] These percentages differ slightly from those in Volume 3 due to this data also including students with a scoring status of "Pending." These students accounted for 1% of the population in grade 5 and grade 8, and for 2% of the population in grade 11.

*Table C-3. Scale Score Mean and Standard Deviation by Grade, Rhode Island*

| | Grade | | |
|---|---|---|---|
| | **5** | **8** | **11** |
| *Number of Students* | 10,798 | 10,546 | 9,733 |
| Mean Scale Score | 51.64 | 50.83 | 52.51 |
| SD of Scale Score | 16.68 | 16.71 | 16.31 |

*Table C-4. Proportion of Students in Each Achievement Level by Grade, Rhode Island[1]*

| Achievement Level | Grade | | |
|---|---|---|---|
| | **5** | **8** | **11** |
| *Number of Students* | 10,798 | 10,546 | 9,733 |
| Level 1 | 0.19 | 0.23 | 0.12 |
| Level 2 | 0.48 | 0.46 | 0.57 |
| Level 3 | 0.20 | 0.22 | 0.17 |
| Level 4 | 0.12 | 0.09 | 0.14 |

*Table C-5. Scale Score Mean and Standard Deviation by Grade, Vermont*

| | Grade | | |
|---|---|---|---|
| | **5** | **8** | **11** |
| *Number of Students* | 6,069 | 5,936 | 5,423 |
| Mean Scale Score | 53.67 | 54.40 | 56.72 |
| SD of Scale Score | 16.27 | 16.19 | 16.85 |

*Table C-6. Proportion of Students in Each Achievement Level by Grade, Vermont*[1]

| **Achievement Level** | **Grade** | | |
|---|---|---|---|
| | **5** | **8** | **11** |
| *Number of Students* | 6,069 | 5,936 | 5,423 |
| Level 1 | 0.15 | 0.17 | 0.08 |
| Level 2 | 0.48 | 0.44 | 0.51 |
| Level 3 | 0.24 | 0.27 | 0.21 |
| Level 4 | 0.13 | 0.12 | 0.20 |

# Appendix D

# Distribution of Scale Scores by Science Discipline

# Distribution of Scale Scores by Science Discipline

*Table D-1. Science Disciplines*

| Grade | Discipline |
|---|---|
| 5, 8, 11 | Physical Sciences (PS)<br>Life Sciences (LS)<br>Earth & Space Sciences (ESS) |

*Table D-2. Combined Overall Discipline Score Mean and Standard Deviation, Grade 5 Science*

| N | Scale Score | Discipline | | |
|---|---|---|---|---|
| | | PS | LS | ESS |
| 16,867 | Mean | 53.20 | 52.35 | 52.93 |
| | SD | 19.28 | 19.74 | 20.21 |

*Table D-3. Combined Overall Discipline Score Mean and Standard Deviation, Grade 8 Science*

| N | Scale Score | Discipline | | |
|---|---|---|---|---|
| | | PS | LS | ESS |
| 16,482 | Mean | 52.32 | 52.35 | 52.19 |
| | SD | 18.63 | 19.96 | 19.78 |

*Table D-4. Combined Overall Discipline Score Mean and Standard Deviation, Grade 11 Science*

| N | Scale Score | Discipline | | |
|---|---|---|---|---|
| | | PS | LS | ESS |
| 15,156 | Mean | 54.06 | 52.54 | 54.03 |
| | SD | 16.84 | 22.99 | 20.60 |

*Table D-5. Overall Discipline Score Mean and Standard Deviation, Grade 5 Science, Rhode Island*

| N | Scale Score | Discipline | | |
|---|---|---|---|---|
| | | PS | LS | ESS |
| 10,798 | Mean | 52.80 | 51.17 | 51.92 |
| | SD | 19.74 | 19.85 | 20.09 |

*Table D-6. Overall Discipline Score Mean and Standard Deviation, Grade 8 Science, Rhode Island*

| N | Scale Score | Discipline | | |
|---|---|---|---|---|
| | | PS | LS | ESS |
| 10,546 | Mean | 51.18 | 50.70 | 51.13 |
| | SD | 18.69 | 19.97 | 19.72 |

*Table D-7. Overall Discipline Score Mean and Standard Deviation, Grade 11 Science, Rhode Island*

| N | Scale Score | Discipline | | |
|---|---|---|---|---|
| | | PS | LS | ESS |
| 9,733 | Mean | 53.07 | 50.77 | 52.01 |
| | SD | 16.60 | 22.84 | 20.09 |

*Table D-8. Overall Discipline Score Mean and Standard Deviation, Grade 5 Science, Vermont*

| N | Scale Score | Discipline | | |
|---|---|---|---|---|
| | | PS | LS | ESS |
| 6,069 | Mean | 53.91 | 54.44 | 54.72 |
| | SD | 18.41 | 19.37 | 20.29 |

*Table D-9. Overall Discipline Score Mean and Standard Deviation, Grade 8 Science, Vermont*

| N | Scale Score | Discipline | | |
|---|---|---|---|---|
| | | PS | LS | ESS |
| 5,936 | Mean | 54.33 | 55.27 | 54.08 |
| | SD | 18.34 | 19.60 | 19.75 |

*Table D-10. Overall Discipline Score Mean and Standard Deviation, Grade 11 Science, Vermont*

| N | Scale Score | Discipline | | |
|---|---|---|---|---|
| | | PS | LS | ESS |
| 5,423 | Mean | 55.86 | 55.71 | 57.66 |
| | SD | 17.11 | 22.91 | 21.00 |

# Appendix E

# Distribution of Scale Scores and Achievement Levels by Subgroup

# Distribution of Scale Scores and Achievement Levels by Subgroup

*Table E-1. Combined Mean and Standard Deviation of Scale Scores by Subgroup*

| Group | Grade 5 | | | Grade 8 | | | Grade 11 | | |
|---|---|---|---|---|---|---|---|---|---|
| | *N* | *Mean* | *SD* | *N* | *Mean* | *SD* | *N* | *Mean* | *SD* |
| **All Students** | 16,867 | 52.37 | 16.56 | 16,482 | 52.12 | 16.61 | 15,156 | 54.01 | 16.63 |
| **Female** | 8,317 | 52.67 | 16.05 | 8,061 | 52.68 | 15.73 | 7,400 | 54.65 | 15.45 |
| **Male** | 8,547 | 52.07 | 17.04 | 8,420 | 51.58 | 17.40 | 7,753 | 53.40 | 17.66 |
| **African American** | 1,102 | 43.85 | 14.54 | 1,074 | 41.75 | 15.09 | 960 | 44.78 | 13.81 |
| **American Indian/Alaskan Native** | 90 | 45.25 | 14.33 | 100 | 43.29 | 17.29 | 95 | 45.28 | 14.41 |
| **Asian** | 496 | 55.15 | 16.95 | 425 | 56.13 | 16.74 | 464 | 57.95 | 18.83 |
| **Hispanic** | 3,128 | 45.08 | 15.14 | 2,842 | 42.82 | 14.67 | 2,458 | 45.29 | 13.22 |
| **Multi-Racial** | 663 | 50.57 | 15.99 | 593 | 49.38 | 16.60 | 402 | 51.09 | 15.65 |
| **Pacific Islander** | 23 | 50.82 | 12.12 | 25 | 51.18 | 16.62 | 21 | 50.85 | 11.57 |
| **White** | 11,365 | 55.24 | 16.23 | 11,423 | 55.48 | 15.84 | 10,756 | 56.85 | 16.47 |
| **Limited English Proficiency** | 1,384 | 37.19 | 12.22 | 898 | 34.07 | 11.19 | 868 | 38.00 | 10.88 |
| **Special Education** | 2,654 | 38.77 | 14.50 | 2,549 | 38.99 | 12.80 | 1,791 | 42.26 | 11.59 |
| **Economically Disadvantaged** | 2,460 | 48.38 | 15.54 | 2,135 | 49.48 | 15.49 | 1,618 | 51.28 | 14.86 |

*Table E-2. Combined Percentage of Achievement Level by Subgroup*

| Group | Grade 5 | | | | | Grade 8 | | | | | Grade 11 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *N* | *L1* | *L2* | *L3* | *L4* | *N* | *L1* | *L2* | *L3* | *L4* | *N* | *L1* | *L2* | *L3* | *L4* |
| **All Students** | 16,867 | 0.18 | 0.48 | 0.22 | 0.12 | 16,482 | 0.21 | 0.45 | 0.24 | 0.10 | 15,156 | 0.11 | 0.55 | 0.18 | 0.16 |
| **Female** | 8,317 | 0.16 | 0.50 | 0.22 | 0.12 | 8,061 | 0.18 | 0.48 | 0.25 | 0.09 | 7,400 | 0.09 | 0.56 | 0.20 | 0.16 |
| **Male** | 8,547 | 0.19 | 0.47 | 0.21 | 0.13 | 8,420 | 0.24 | 0.43 | 0.23 | 0.11 | 7,753 | 0.13 | 0.54 | 0.16 | 0.17 |
| **African American** | 1,102 | 0.33 | 0.53 | 0.12 | 0.03 | 1,074 | 0.42 | 0.45 | 0.11 | 0.02 | 960 | 0.21 | 0.65 | 0.09 | 0.04 |
| **American Indian/Alaskan Native** | 90 | 0.27 | 0.58 | 0.12 | 0.03 | 100 | 0.40 | 0.45 | 0.11 | 0.04 | 95 | 0.22 | 0.69 | 0.03 | 0.05 |
| **Asian** | 496 | 0.14 | 0.46 | 0.25 | 0.15 | 425 | 0.14 | 0.41 | 0.31 | 0.13 | 464 | 0.06 | 0.54 | 0.15 | 0.25 |
| **Hispanic** | 3,128 | 0.30 | 0.52 | 0.13 | 0.04 | 2,842 | 0.40 | 0.47 | 0.11 | 0.03 | 2,458 | 0.20 | 0.67 | 0.09 | 0.04 |
| **Multi-Racial** | 663 | 0.20 | 0.51 | 0.21 | 0.08 | 593 | 0.26 | 0.46 | 0.20 | 0.08 | 402 | 0.11 | 0.63 | 0.15 | 0.10 |
| **Pacific Islander** | 23 | 0.13 | 0.65 | 0.13 | 0.09 | 25 | 0.12 | 0.6 | 0.20 | 0.08 | 21 | 0.10 | 0.67 | 0.19 | 0.05 |
| **White** | 11,365 | 0.13 | 0.47 | 0.25 | 0.15 | 11,423 | 0.14 | 0.45 | 0.29 | 0.12 | 10,756 | 0.08 | 0.51 | 0.21 | 0.20 |
| **Limited English Proficiency** | 1,384 | 0.49 | 0.47 | 0.03 | 0 | 898 | 0.67 | 0.31 | 0.02 | 0 | 868 | 0.37 | 0.61 | 0.01 | 0.01 |
| **Special Education** | 2,654 | 0.48 | 0.43 | 0.07 | 0.02 | 2,549 | 0.50 | 0.43 | 0.06 | 0.01 | 1,791 | 0.24 | 0.69 | 0.05 | 0.02 |
| **Economically Disadvantaged** | 2,460 | 0.23 | 0.53 | 0.18 | 0.06 | 2,135 | 0.24 | 0.49 | 0.20 | 0.07 | 1,618 | 0.12 | 0.62 | 0.15 | 0.10 |

*Table E-3. Mean and Standard Deviation of Scale Scores by Subgroup, Rhode Island*

| Group | Grade 5 | | | Grade 8 | | | Grade 11 | | |
|---|---|---|---|---|---|---|---|---|---|
| | *N* | *Mean* | *SD* | *N* | *Mean* | *SD* | *N* | *Mean* | *SD* |
| **All Students** | 10,798 | 51.64 | 16.68 | 10,546 | 50.83 | 16.71 | 9,733 | 52.51 | 16.31 |
| **Female** | 5,349 | 52.00 | 16.06 | 5,215 | 51.34 | 15.82 | 4,748 | 53.10 | 15.21 |
| **Male** | 5,449 | 51.28 | 17.26 | 5,331 | 50.34 | 17.53 | 4,985 | 51.94 | 17.28 |
| **African American** | 938 | 43.61 | 14.21 | 920 | 41.60 | 14.84 | 832 | 44.13 | 13.45 |
| **American Indian/Alaskan Native** | 78 | 43.47 | 13.92 | 78 | 41.38 | 16.59 | 79 | 44.06 | 14.00 |
| **Asian** | 397 | 56.29 | 16.99 | 291 | 55.77 | 17.35 | 317 | 57.67 | 18.19 |
| **Hispanic** | 2,986 | 44.74 | 14.96 | 2,721 | 42.41 | 14.44 | 2,328 | 44.71 | 12.89 |
| **Multi-Racial** | 494 | 49.65 | 15.71 | 422 | 47.95 | 16.57 | 310 | 50.14 | 15.86 |
| **Pacific Islander** | 8 | 46.44 | 7.80 | 14 | 41.01 | 10.48 | 15 | 52.24 | 12.57 |
| **White** | 5,897 | 56.38 | 16.19 | 6,100 | 56.09 | 15.64 | 5,852 | 56.76 | 16.17 |
| **Limited English Proficiency** | 1,239 | 36.78 | 11.88 | 826 | 33.89 | 10.90 | 792 | 37.67 | 11.00 |
| **Special Education** | 1,457 | 37.17 | 13.87 | 1,491 | 38.01 | 12.82 | 1,071 | 41.84 | 11.18 |
| **Economically Disadvantaged** | – | – | – | – | – | – | – | – | – |

*Table E-4. Percentage of Achievement Level by Subgroup, Rhode Island*

| Group | Grade 5 | | | | | Grade 8 | | | | | Grade 11 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *N* | *L1* | *L2* | *L3* | *L4* | *N* | *L1* | *L2* | *L3* | *L4* | *N* | *L1* | *L2* | *L3* | *L4* |
| **All Students** | 10,798 | 0.19 | 0.48 | 0.20 | 0.12 | 10,546 | 0.23 | 0.46 | 0.22 | 0.09 | 9,733 | 0.12 | 0.57 | 0.17 | 0.14 |
| **Female** | 5,349 | 0.17 | 0.51 | 0.21 | 0.11 | 5,215 | 0.2 | 0.49 | 0.23 | 0.08 | 4,748 | 0.10 | 0.59 | 0.18 | 0.13 |
| **Male** | 5,449 | 0.22 | 0.46 | 0.20 | 0.12 | 5,331 | 0.26 | 0.43 | 0.21 | 0.10 | 4,985 | 0.14 | 0.56 | 0.15 | 0.15 |
| **African American** | 938 | 0.33 | 0.53 | 0.11 | 0.03 | 920 | 0.43 | 0.45 | 0.10 | 0.02 | 832 | 0.22 | 0.66 | 0.08 | 0.03 |
| **American Indian/Alaskan Native** | 78 | 0.31 | 0.55 | 0.13 | 0.01 | 78 | 0.45 | 0.41 | 0.12 | 0.03 | 79 | 0.25 | 0.67 | 0.04 | 0.04 |
| **Asian** | 397 | 0.13 | 0.45 | 0.26 | 0.17 | 291 | 0.15 | 0.41 | 0.29 | 0.14 | 317 | 0.04 | 0.58 | 0.13 | 0.24 |
| **Hispanic** | 2,986 | 0.31 | 0.52 | 0.13 | 0.04 | 2,721 | 0.40 | 0.47 | 0.10 | 0.02 | 2,328 | 0.21 | 0.67 | 0.08 | 0.03 |
| **Multi-Racial** | 494 | 0.20 | 0.53 | 0.19 | 0.07 | 422 | 0.29 | 0.45 | 0.19 | 0.07 | 310 | 0.12 | 0.64 | 0.15 | 0.08 |
| **Pacific Islander** | 8 | 0.13 | 0.75 | 0.13 | – | 14 | 0.21 | 0.79 | – | – | 15 | 0.07 | 0.60 | 0.27 | 0.07 |
| **White** | 5,897 | 0.12 | 0.46 | 0.25 | 0.17 | 6,100 | 0.13 | 0.45 | 0.30 | 0.13 | 5,852 | 0.07 | 0.52 | 0.22 | 0.20 |
| **Limited English Proficiency** | 1,239 | 0.50 | 0.47 | 0.03 | 0 | 826 | 0.67 | 0.31 | 0.02 | 0 | 792 | 0.39 | 0.59 | 0.01 | 0.01 |
| **Special Education** | 1,457 | 0.53 | 0.40 | 0.05 | 0.02 | 1,491 | 0.54 | 0.40 | 0.05 | 0.01 | 1,071 | 0.25 | 0.69 | 0.04 | 0.01 |
| **Economically Disadvantaged** | – | – | – | – | – | – | – | – | – | – | – | – | – | – | – |

*Table E-5. Mean and Standard Deviation of Scale Scores by Subgroup, Vermont*

| Group | Grade 5 | | | Grade 8 | | | Grade 11 | | |
|---|---|---|---|---|---|---|---|---|---|
| | *N* | *Mean* | *SD* | *N* | *Mean* | *SD* | *N* | *Mean* | *SD* |
| **All Students** | 6,069 | 53.67 | 16.27 | 5,936 | 54.40 | 16.19 | 5,423 | 56.72 | 16.85 |
| **Female** | 2,968 | 53.88 | 15.95 | 2,846 | 55.14 | 15.25 | 2,652 | 57.43 | 15.49 |
| **Male** | 3,098 | 53.46 | 16.57 | 3,089 | 53.72 | 16.98 | 2,768 | 56.03 | 18.03 |
| **African American** | 164 | 45.24 | 16.27 | 154 | 42.69 | 16.51 | 128 | 48.97 | 15.37 |
| **American Indian/Alaskan Native** | 12 | 56.87 | 11.62 | 22 | 50.06 | 18.4 | 16 | 51.33 | 15.35 |
| **Asian** | 99 | 50.6 | 16.09 | 134 | 56.91 | 15.36 | 147 | 58.56 | 20.18 |
| **Hispanic** | 142 | 52.29 | 17.07 | 121 | 52.04 | 16.59 | 130 | 55.64 | 14.69 |
| **Multi-Racial** | 169 | 53.27 | 16.54 | 171 | 52.91 | 16.19 | 92 | 54.29 | 14.53 |
| **Pacific Islander** | 15 | 53.15 | 13.55 | 11 | 64.12 | 13.84 | 6 | 47.37 | 8.57 |
| **White** | 5,468 | 54.02 | 16.19 | 5,323 | 54.77 | 16.05 | 4,904 | 56.97 | 16.83 |
| **Limited English Proficiency** | 145 | 40.74 | 14.36 | 72 | 36.15 | 14.03 | 76 | 41.53 | 8.95 |
| **Special Education** | 1,197 | 40.72 | 15.01 | 1,058 | 40.37 | 12.65 | 720 | 42.88 | 12.16 |
| **Economically Disadvantaged** | 2,460 | 48.38 | 15.54 | 2,135 | 49.48 | 15.49 | 1,618 | 51.28 | 14.86 |

*Table E-6. Percentage of Achievement Level by Subgroup, Vermont*

| Group | Grade 5 | | | | | Grade 8 | | | | | Grade 11 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *N* | *L1* | *L2* | *L3* | *L4* | *N* | *L1* | *L2* | *L3* | *L4* | *N* | *L1* | *L2* | *L3* | *L4* |
| **All Students** | 6,069 | 0.15 | 0.48 | 0.24 | 0.13 | 5,936 | 0.17 | 0.44 | 0.27 | 0.12 | 5,423 | 0.08 | 0.51 | 0.21 | 0.20 |
| **Female** | 2,968 | 0.14 | 0.48 | 0.25 | 0.12 | 2,846 | 0.14 | 0.47 | 0.29 | 0.11 | 2,652 | 0.06 | 0.51 | 0.23 | 0.20 |
| **Male** | 3,098 | 0.16 | 0.47 | 0.24 | 0.13 | 3,089 | 0.19 | 0.42 | 0.25 | 0.13 | 2,768 | 0.10 | 0.50 | 0.19 | 0.21 |
| **African American** | 164 | 0.30 | 0.51 | 0.13 | 0.05 | 154 | 0.42 | 0.42 | 0.12 | 0.05 | 128 | 0.16 | 0.57 | 0.17 | 0.09 |
| **American Indian/ Alaskan Native** | 12 | – | 0.75 | 0.08 | 0.17 | 22 | 0.23 | 0.59 | 0.09 | 0.09 | 16 | 0.06 | 0.81 | – | 0.13 |
| **Asian** | 99 | 0.20 | 0.48 | 0.22 | 0.09 | 134 | 0.11 | 0.42 | 0.36 | 0.11 | 147 | 0.11 | 0.44 | 0.19 | 0.27 |
| **Hispanic** | 142 | 0.17 | 0.46 | 0.27 | 0.11 | 121 | 0.21 | 0.45 | 0.24 | 0.10 | 130 | 0.03 | 0.60 | 0.20 | 0.17 |
| **Multi-Racial** | 169 | 0.18 | 0.45 | 0.27 | 0.11 | 171 | 0.19 | 0.47 | 0.23 | 0.11 | 92 | 0.07 | 0.61 | 0.15 | 0.17 |
| **Pacific Islander** | 15 | 0.13 | 0.60 | 0.13 | 0.13 | 11 | – | 0.36 | 0.45 | 0.18 | 6 | 0.17 | 0.83 | – | – |
| **White** | 5,468 | 0.14 | 0.48 | 0.25 | 0.13 | 5,323 | 0.16 | 0.44 | 0.28 | 0.12 | 4,904 | 0.08 | 0.50 | 0.21 | 0.21 |
| **Limited English Proficiency** | 145 | 0.41 | 0.48 | 0.09 | 0.01 | 72 | 0.61 | 0.29 | 0.08 | 0.01 | 76 | 0.25 | 0.72 | 0.03 | – |
| **Special Education** | 1,197 | 0.42 | 0.47 | 0.09 | 0.02 | 1,058 | 0.44 | 0.47 | 0.07 | 0.02 | 720 | 0.23 | 0.69 | 0.06 | 0.03 |
| **Economically Disadvantaged** | 2,460 | 0.23 | 0.53 | 0.18 | 0.06 | 2,135 | 0.24 | 0.49 | 0.20 | 0.07 | 1,618 | 0.12 | 0.62 | 0.15 | 0.10 |