# New Hampshire Statewide Assessment System

# 2018–2019

# Volume 1

# Annual Technical Report

New Hampshire
Department of Education
*Serving New Hampshire's Education Community*

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

# LIST OF APPENDICES

# 1. INTRODUCTION

The New Hampshire Statewide Assessment System (NH SAS) is a series of assessments for English language arts (ELA) and mathematics in grades 3–8 and for science in grades 5, 8, and 11. The NH SAS 2018–2019 technical report volumes are 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 comprises seven separate, self-contained volumes:

1) ***Annual Technical Report***. This volume is updated each year and provides a general 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 its development process.
3) ***Setting Performance Standards***. This volume documents the methods and results of the NH SAS standard-setting process.
4) ***Reliability and Validity***. This volume provides an array of reliability and validity evidence to support the intended uses and interpretations of the test scores.
5) ***Test Administration***. This volume describes the methods used to administer all test forms, security protocols, and modifications or accommodations available.
6) ***Score Interpretation Guide***. This volume describes the score types reported as well as the appropriate inferences and uses intended for each score type.
7) ***Special Studies.*** This volume consists of any special studies conducted. It is updated each year to reflect studies relevant to the respective administration.

The New Hampshire Department of Education (NHDOE) communicates the quality of the NH SAS by making the technical report accessible to the public.

## 1.1. BACKGROUND AND HISTORICAL CONTEXT

After 10 years of testing ELA and mathematics under the New England Common Assessment Program (2004–2014) and three years of testing with the Smarter Balanced Assessment Consortium (2015–2017), in the 2017–2018 school year, New Hampshire administered the NH SAS to students in grades 3–8. The shift from the New England Common Assessment Program (NECAP) to the Smarter Balanced Assessment Consortium (SBAC) represented not only a change in assessment programs, but also adoption of new career and college readiness content standards and the establishment of new performance standards based on ensuring that students in grades 3–8 were on track to career and college readiness. Unlike the NECAP to SBAC transition, the SBAC to NH SAS shift reflected only a change in assessment programs. There was no change in the New Hampshire College and Career Ready Standards (NH CCRS) in ELA and mathematics, which were used to construct the NH SAS.

The NH SAS in science was also first administered in spring 2018, to students in grades 5, 8, and 11, replacing the New England Common Assessment Program (NECAP) for science in grades 4, 8, and 11. This shift effectively implemented the 2016 adoption of NH's College and Career Ready

Science Standards, or the Next Generation Science Standards (NGSS), which were used to construct the NH SAS in science.

The NH SAS assessments were built using AIRCore items, which are discussed further in Section 3 of this volume as well as Volume 2. The NH SAS ELA and mathematics tests are delivered as online, adaptive assessments for grades 3–8 students. The NH SAS science assessment is administered online to students in grades 5, 8, and 11 using a linear-on-the-fly (LOFT) test design.

## 1.2. Purpose and Intended Uses of the New Hampshire Statewide Assessment System

The primary purpose of NH SAS is to yield test scores at the student level and other levels of aggregation that reflect student performance, including English learners (ELs) and students with disabilities, relative to the NH CCRS. The NH SAS is a criterion-referenced test that applies principles of evidence-centered design (described further in Volume 2) to yield overall and reporting category-level test scores at the student level and other levels of aggregation that reflect student achievement of the NH CCRS.

The NH SAS in ELA, mathematics, and science draw all items from the AIRCore item banks (see Volume 2), which are rigorously developed banks of items aligned to recognized career and college readiness standards that have been widely adopted by many states. The American Institutes for Research (AIR) and the NHDOE worked together to ensure that the items in the tests constructed for all grades uniquely measure students' mastery of the NH CCRS in ELA, mathematics, and science.

The NH SAS supports instruction and student learning by providing timely feedback to educators and parents, which can be used to target resources and inform instructional strategies that remediate or enrich instruction. For spring 2018, scores were reported after setting performance standards (see Volume 3 for more information). In spring 2019 and all future summative administrations, scores are available and reported immediately. An array of reporting metrics allows performance to be monitored at both student and aggregate levels and growth to be measured at both student and group levels over time. Assessments can be used as an indicator to determine whether students in New Hampshire are ready with the knowledge and skills that are essential for college and career readiness.

## 1.3. Participants in the Development and Analysis of the NH SAS

The NHDOE manages the New Hampshire state assessment program with the assistance of several participants, including New Hampshire educators, a Technical Advisory Committee (TAC), and several vendors listed below. NHDOE fulfills the diverse requirements of implementing New Hampshire'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, 1999, 2014).

### 1.3.1. New Hampshire Department of Education

The Bureau of Instructional Support and Student Assessment coordinates, implements, and monitors statewide assessment system, including coordination with other NHDOE offices, New Hampshire public schools, and vendors.

### 1.3.2. New Hampshire Educators

New Hampshire educators participate in most aspects of the conceptualization and development of the NH SAS. Educators participate in developing the academic standards, including clarifying how these standards are assessed, test design, and review of test questions and passages. See Volume 2 for more details.

### 1.3.3. Technical Advisory Committee

NHDOE convenes a panel multiple times a year to discuss psychometric, test development, administrative, and policy issues of relevance to current and future New Hampshire testing. This committee is composed of several highly experienced practitioners from multiple New Hampshire school districts and recognized assessment experts who have provided guidance on other state and national testing programs.

### 1.3.4. American Institutes for Research

AIR is the vendor that was selected through the state-mandated competitive procurement process. In Fall 2017, AIR became the primary party responsible for developing NH SAS test content, building item pools and forms, conducting psychometric analyses, test administration, and scoring, and reporting results. Additionally, AIR is responsible for developing and maintaining the AIRCore bank (see Volume 2 for more information), which is used to construct the NH SAS for ELA, mathematics, and science.

### 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. TEST DESIGN

The NH SAS ELA and mathematics assessments are administered to students in grades 3–8 as online assessments using an adaptive item selection algorithm (Volume 2, Appendix J) with several technology-enhanced item types such as those shown in Table 12 and Table 13. Students in each grade responded to one writing prompt, administered online. Reading and Writing item responses were combined so that the data could be scored together to form an overall ELA score. In this document, the term *ELA* is used when referring to the combined Reading and Writing test, *Reading* is used when referring to only the Reading test or items, and *Writing* is used for only the writing prompt items. The 2019 NH SAS ELA and mathematics tests also contain new field-test items.

The NH SAS science assessment is administered online to students in grades 5, 8, and 11 using a LOFT test design. Science items are centered around a scientific phenomenon. They can consist of shorter items (stand-alone) or items with several parts (item clusters) requiring the student to interact with the item in various ways. All AIRCore science items adhere to the framework of the Next Generation Science Standards (NGSS). The science test was an operational field test in 2018, the first year of the new science assessment. In 2019 and onwards, additional items are field-tested to build out the item bank.

Students unable to participate in the online administration and requiring use of designated support or accommodation had the option to use print-on-request, a feature that provided the same items administered to students online but in a paper-pencil format. More information about designated supports and accommodations is available in Section 2.1 and Volume 5.

## 1.5. Student Participation

All New Hampshire public school students are required to participate in the statewide assessments. Table 1 shows the number of students who were tested and the number of students who were reported in the spring 2019 NH SAS by grade and subject area. It is expected that the number of students with reported scores is slightly less than the number of students tested, due to instances of incomplete tests where students did not respond to enough items for a score report to be generated.

*Table 1: Number of Students Participating in NH SAS Spring 2019*

| | Mathematics | | ELA | | Science | |
|---|---|---|---|---|---|---|
| ***Grade*** | ***Number Tested*** | ***Number Reported*** | ***Number Tested*** | ***Number Reported*** | ***Number Tested*** | ***Number Reported*** |
| 3 | 11,261 | 11,251 | 12,282 | 12,227 | - | - |
| 4 | 12,824 | 12,818 | 11,648 | 11,616 | - | - |
| 5 | 11,880 | 11,867 | 11,864 | 11,815 | 13,191 | 13,187 |
| 6 | 12,348 | 12,330 | 12,365 | 12,277 | - | - |
| 7 | 12,301 | 12,268 | 12,409 | 12,343 | - | - |
| 8 | 13,175 | 13,144 | 13,224 | 13,070 | 12,070 | 12,060 |
| 11 | - | - | - | - | 11,398 | 11,385 |

# 2. Summary of Operational Procedures

## 2.1. Test Administration

The test administration schedule for the 2019 New Hampshire Statewide Assessment System (NH SAS) is presented by content area in Table 2.

*Table 2: 2019 Testing Windows by Subject Area*

| Assessment | Subject | Grade(s) | Testing Window |
|---|---|---|---|
| Summative NH SAS | ELA (Reading & Writing) | 3–8 | March 19–June 7, 2019 |
| | Mathematics | 3–8 | March 19–June 7, 2019 |
| | Science | 5, 8, and 11 | March 19–June 7, 2019 |

The key personnel involved with the NH SAS administration include the district administrators (DAs), District Test Coordinators (DCs), School Test Coordinators (SCs), and Test Administrators (TAs) who proctor the test. A test administration manual (TAM; Volume 5 Appendix C) is provided so that personnel involved with statewide assessment administrations can maintain both standardized administration conditions and test security.

A secure browser developed by AIR is required to access the online NH SAS assessments. The online browser provides 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). 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 a majority of students for test administration planning purposes.

Students participating in the computer-based NH SAS are able to use the standard online testing features in the test delivery system (TDS), which includes a selection of font color and size and the ability to zoom in and out or highlight text. In addition to the resources available to all students, there are options such as braille, American Sign Language (ASL), and closed captioning available to accommodate students who are English learners (ELs) or students with accommodations prescribed by an Individualized Education Program (IEP) or Section 504 Plan. For ELs, Spanish language versions of the NH SAS mathematics and science are available. TAs and SCs in New Hampshire are responsible for ensuring that arrangements for accommodations are made before the test administration dates. During test development, it was ensured that scores obtained on the Spanish language version or other alternative modes of administrations are comparable to those received on the standard online test adhering to the same blueprints. For more information, see Volume 2.

## 2.2. SIMULATIONS

Prior to the operational testing window, AIR employs a simulation approach. Simulations are performed for all NH SAS, including English language arts (ELA), mathematics, and science tests.

For ELA and mathematics, simulations are used to configure the adaptive algorithm (described further in Volume 2, Appendix J), seeking to maximize test score precision while meeting blueprint specifications based on the available pool of test items. Psychometricians review ELA and mathematics simulation results for the following key diagnostic factors:

- Match-to-test blueprint: Determines that the tests have the correct number of test items overall and the appropriate proportion by content strands, as specified in the test blueprints for every grade and subject.

- Precision: Determines whether the size of the standard error of measurement is within the acceptable range and whether there is any possible bias in the estimates of student ability.
- Item exposure rate: Evaluates the utility of item pools and identifies overexposed and underexposed items.

These diagnostics are interrelated. For example, if the test pool for a particular content strand is limited (i.e., if there are only a few items available), achieving a 100% match to the blueprint for this content strand will lead to a high item exposure rate, which means that a large number of students will see the same items. A high item exposure rate results in decreased benefits from adaptive testing relative to using a fixed form, such as the usual increased security caused by a larger pool of items. The software system that performs the simulation allows the adjustment of test configuration to attain the best possible balance among these diagnostics. The simulation involves an iterative process that reviews initial results, adjusts these system parameters, runs new simulations, reviews the new results, and repeats the exercise until an optimal balance is achieved. The final setting is then applied for operational tests. The ELA and mathematics simulation reports in Appendix A describe in detail the simulation approach and results evaluated based on blueprint, precision, and item exposure rate.

For science, administered under a LOFT test design, the same algorithm is used to select items as for the adaptive tests, but 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 for science are discussed in Volume 2.

## 2.3. DESIGNATED SUPPORTS AND ACCOMMODATIONS

Designated support features are available for those for whom the need has been identified by an informed educator or team of educators. All educators making these decisions are trained on the process and understand the range of designated supports available. Scores achieved by students using designated supports are included for federal accountability purposes.

Accommodations are available for students for whom there is documented need on an IEP or Section 504 Plan. 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. State-approved accommodations do not compromise the learning expectations, constructs, or grade-level standards. Such accommodations help students with a documented need in an IEP or Section 504 Plan generate valid testing results so that they can fully demonstrate what students 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 available accommodation options for eligible students included the following: ASL, braille, embossing, presentation, print-on-demand, streamlined mode, text-to-speech (TTS), ASL human signer, abacus, 100s Number Table, read aloud, scribe, and speech-to-text. During test development, detailed in Volume 2, it was ensured that scores obtained from alternative modes of administrations, such as print-on-request, are comparable to those from the regular online testing adhering to the same test blueprints.

Embedded designated supports or accommodations are provided through instructional or assessment technology as part of the computer administration, whereas non-embedded features are non-digital and provided locally outside of the administration system. Further information about designated supports and accommodations is available in Volume 5.

Table 3 through Table 11 list the number of testing sessions in which a student was provided with each accommodation or designated support during the spring 2019 test administration.

*Table 3: Total Sessions with Allowed Embedded Designated Supports, ELA*

| Designated Supports | Grade | | | | | |
|---|---|---|---|---|---|---|
| | ***3*** | ***4*** | ***5*** | ***6*** | ***7*** | ***8*** |
| American Sign Language | 3 | 3 | 2 | - | - | 2 |
| Braille | - | - | - | - | - | 2 |
| Color Choices: Non-Default | 8 | 26 | 18 | 18 | 12 | 11 |
| Emboss: Stimuli & Items | - | - | - | - | - | 2 |
| Mouse Pointer: Non-Default | 6 | 13 | 17 | 13 | 3 | 5 |
| Permissive Mode | 141 | 139 | 146 | 44 | 49 | 62 |
| Print-on-Request | 49 | 48 | 22 | 10 | 12 | 18 |
| Streamlined Mode | 65 | 49 | 48 | 46 | 27 | 56 |
| Text-to-Speech | 1,549 | 1,516 | 1,358 | 1,222 | 986 | 883 |

*Table 4: Total Sessions with Allowed Non-Embedded Designated Supports, ELA*

| Designated Supports | Grade | | | | | |
|---|---|---|---|---|---|---|
| | ***3*** | ***4*** | ***5*** | ***6*** | ***7*** | ***8*** |
| ASL Human Signer | - | 1 | - | - | - | - |
| Amplification | 8 | 2 | 4 | 3 | 5 | 3 |
| Alternate Response | 1 | 1 | - | 1 | 5 | 8 |
| Bilingual Dictionary | 9 | 5 | 20 | 28 | 27 | 20 |
| Color Contrast | 13 | 23 | 20 | 21 | 17 | 9 |
| Color Overlays | 4 | 1 | 8 | 16 | 6 | 4 |
| Magnification | 9 | 13 | 12 | 19 | 19 | 13 |
| Noise Buffer | 52 | 88 | 76 | 28 | 32 | 32 |
| Read Aloud: Items | 363 | 361 | 414 | 302 | 219 | 192 |
| Read Aloud: Stimuli | 307 | 309 | 361 | 226 | 181 | 132 |
| Scribe | 360 | 334 | 271 | 205 | 151 | 107 |
| Separate Setting | 1,400 | 1,473 | 1,356 | 1,234 | 1,052 | 990 |

| Designated Supports | Grade | | | | | |
|---|---|---|---|---|---|---|
| | *3* | *4* | *5* | *6* | *7* | *8* |
| Speech-to-Text | 83 | 93 | 125 | 92 | 78 | 42 |

*Table 5: Total Sessions with Allowed Embedded and Non-Embedded Accommodations, ELA*

| Accommodations | Grade | | | | | |
|---|---|---|---|---|---|---|
| | *3* | *4* | *5* | *6* | *7* | *8* |
| **Embedded** | | | | | | |
| American Sign Language | 3 | 3 | 2 | - | - | 2 |
| Braille | - | - | - | - | - | 2 |
| Emboss: Stimuli & Items | - | - | - | - | - | 2 |
| Streamlined Mode | 65 | 49 | 48 | 46 | 27 | 56 |
| Text-to-Speech | 1,549 | 1,516 | 1,358 | 1,222 | 986 | 883 |
| **Non-Embedded** | | | | | | |
| Abacus | - | - | - | - | - | 3 |
| ASL Human Signer | - | 1 | - | - | - | - |
| Print-on-Request | 32 | 40 | 24 | 20 | 17 | 16 |
| Read Aloud | 367 | 344 | 322 | 242 | 185 | 134 |
| Scribe | 319 | 336 | 228 | 191 | 153 | 113 |
| Speech-to-Text | 145 | 144 | 160 | 101 | 102 | 66 |

*Table 6: Total Sessions with Allowed Embedded Designated Supports, Mathematics*

| Designated Supports | Grade | | | | | |
|---|---|---|---|---|---|---|
| | *3* | *4* | *5* | *6* | *7* | *8* |
| Braille | - | - | - | - | - | 1 |
| Color Choices: Non-Default | 6 | 12 | 18 | 15 | 11 | 10 |
| Emboss: Stimuli & Items | - | - | - | - | - | 1 |
| Mouse Pointer: Non-Default | 6 | 17 | 16 | 13 | 2 | 5 |
| Permissive Mode | 128 | 158 | 151 | 64 | 44 | 69 |
| Print-on-Request | 46 | 39 | 22 | 12 | 11 | 17 |
| Streamlined Mode | 51 | 59 | 50 | 46 | 23 | 54 |

| Designated Supports | Grade | | | | | |
|---|---|---|---|---|---|---|
| | *3* | *4* | *5* | *6* | *7* | *8* |
| Text-to-Speech | 1,654 | 1,831 | 1,557 | 1,558 | 1,268 | 1,191 |

*Table 7: Total Sessions with Allowed Non-Embedded Designated Supports, Mathematics*

| Designated Supports | Grade | | | | | |
|---|---|---|---|---|---|---|
| | *3* | *4* | *5* | *6* | *7* | *8* |
| ASL Human Signer | - | 1 | - | - | - | - |
| Amplification | 8 | 4 | 4 | 3 | 5 | 3 |
| Alternate Response | - | - | - | - | 4 | 7 |
| Color Contrast | 13 | 23 | 24 | 22 | 16 | 11 |
| Color Overlays | 5 | 3 | 9 | 16 | 6 | 4 |
| Magnification | 6 | 15 | 22 | 18 | 18 | 16 |
| Noise Buffer | 44 | 105 | 80 | 26 | 32 | 36 |
| Read Aloud: Items | 433 | 531 | 514 | 369 | 238 | 212 |
| Read Aloud: Stimuli | 360 | 436 | 458 | 281 | 192 | 159 |
| Scribe | 320 | 383 | 259 | 213 | 149 | 112 |
| Separate Setting | 1,306 | 1,583 | 1,334 | 1,238 | 1,030 | 1,001 |
| Speech-to-Text | 69 | 111 | 119 | 85 | 75 | 50 |

*Table 8: Total Sessions with Allowed Embedded and Non-Embedded Accommodations, Mathematics*

| Accommodations | Grade | | | | | |
|---|---|---|---|---|---|---|
| | *3* | *4* | *5* | *6* | *7* | *8* |
| **Embedded** | | | | | | |
| Braille | - | - | - | - | - | 1 |
| Emboss: Stimuli & Items | - | - | - | - | - | 1 |
| Streamlined Mode | 51 | 59 | 50 | 46 | 23 | 54 |
| Text-to-Speech | 1,654 | 1,831 | 1,557 | 1,558 | 1,268 | 1,191 |
| **Non-Embedded** | | | | | | |
| 100s Number Table | 37 | 218 | 220 | 215 | 195 | 149 |
| Abacus | 9 | 2 | 3 | 4 | 2 | 12 |

| Accommodations | Grade | | | | | |
|---|---|---|---|---|---|---|
| | ***3*** | ***4*** | ***5*** | ***6*** | ***7*** | ***8*** |
| ASL Human Signer | 3 | 4 | - | - | - | - |
| Print-on-Request | 29 | 43 | 22 | 21 | 18 | 14 |
| Read Aloud: Stimuli | 376 | 490 | 379 | 265 | 186 | 155 |
| Scribe | 280 | 380 | 242 | 196 | 145 | 120 |
| Speech-to-Text | 119 | 160 | 152 | 95 | 95 | 68 |

*Table 9: Total Sessions with Allowed Embedded Designated Supports, Science*

| Designated Supports | Grade | | |
|---|---|---|---|
| | ***5*** | ***8*** | ***11*** |
| Color Contrast | - | - | - |
| Mouse Pointer | 19 | 5 | - |
| Masking | 21 | 14 | - |
| Print Size | 17 | 17 | 6 |
| Color Choices | 20 | 8 | - |

*Table 10: Total Sessions with Allowed Non-Embedded Designated Supports, Science*

| Designated Supports | Grade | | |
|---|---|---|---|
| | ***5*** | ***8*** | ***11*** |
| Amplification | 4 | 5 | 2 |
| Color Contrast | 23 | 8 | - |
| Color Overlays | 9 | 5 | - |
| Magnification | 12 | 12 | 3 |
| Noise Buffer | 79 | 32 | 1 |
| Read Aloud: Items | 508 | 182 | 22 |
| Read Aloud: Stimuli | 451 | 135 | 11 |
| Scribe | 285 | 92 | 7 |
| Separate Setting | 1,360 | 853 | 209 |
| Alternate Response | - | 8 | - |
| Speech-to-Text | 135 | 33 | 4 |

*Table 11: Total Sessions with Allowed Embedded and Non-Embedded Accommodations, Science*

| Accommodations | Grade | | |
|---|---|---|---|
| | ***5*** | ***8*** | ***11*** |
| **Embedded** | | | |
| Audio Transcription | - | - | - |
| Color Choices | 20 | 8 | - |
| Streamlined Mode | 51 | 48 | 3 |
| **Non-Embedded** | | | |
| ASL Human Signer | - | - | - |
| Print-on-Request | 28 | 12 | - |
| Scribe | 277 | 102 | 11 |
| Speech-to-Text | 165 | 54 | 5 |
| Read Aloud Stimuli | 406 | 135 | 14 |

# 3. Item Banks and Test Design

New Hampshire content specialists and psychometricians reviewed all items in the AIRCore item banks with respect to item statistics, bias, and sensitivity for the state of New Hampshire. The items that were selected after these reviews were used for the New Hampshire operational item pool. In this section, we describe the characteristics of the spring 2019 operational item pool for the computer adaptive tests (English language arts (ELA) and mathematics) and the online tests administered linearly on the fly (science). The characteristics include both content (e.g., item types) and statistical summaries. Test design and methodology of field-testing new items are also discussed.

## 3.1. ELA and Mathematics Item Bank

For ELA and mathematics, all operational items used on the New Hampshire Statewide Assessment System (NH SAS) tests are drawn from the AIRCore item bank. Volume 2 is a separate, stand-alone report containing complete details on the AIRCore item bank; here, we note that AIRCore for ELA and mathematics is a pre-equated item bank with item parameters estimated under the multigroup item response theory (IRT) framework described in a later section of this volume.

The operational item pool includes an array of item types used to measure the NH CCRS in ELA and mathematics. Table 12 and Table 13 describe each of the item types in the item pool, and Table 14 and Table 15 show the number of items by item type for ELA and mathematics, respectively.

*Table 12: ELA Item Types*

| Response Type | Description |
|---|---|
| Evidence-Based Selected Response (EBSR) | Student selects the correct answers from Part A and Part B. Part A often asks the student to make an analysis or inference, and Part B requires the student to use text to support Part A. |
| Extended Response (ER) | Student is directed to provide a longer, written response. |
| Editing Task Choice (ETC) | Student identifies an incorrect word or phrase and chooses the replacement from a number of options. |
| Grid (GI) | Student selects words, phrases, or images and uses the drag-and-drop feature to place them into a graphic organizer. |
| Hot Text (HT) | Student is directed to either select or use the drag-and-drop feature to use text to support an analysis or make an inference. |
| Matching (MI) | Student checks a box to indicate if information from a column header matches information from a row. |
| Multiple Choice (MC) | Student selects one correct answer from a number of options. |
| Multiple Select (MS) | Student selects all correct answers from a number of options. |
| Natural Language (NL) | Student uses the keyboard to enter a response into a text field. |

*Table 13: Mathematics Item Types*

| Response Type | Description |
|---|---|
| Equation (EQ) | Student uses a toolbar with a variety of mathematical symbols to create a response. |
| Editing Task Choice (ETC) | Student identifies an incorrect word or phrase and chooses the replacement from a number of options. |
| Grid (GI) | Student selects words, phrases, or images and uses the drag-and-drop feature to place them into a graphic organizer. |
| Matching (MI) | Student checks a box to indicate if information from a column header matches information from a row. |
| Multiple Choice (MC) | Student selects one correct answer from a number of options. |
| Multiple Select (MS) | Student selects all correct answers from a number of options. |
| Table (TI) | Student types numeric values into a given table. |

*Table 14: ELA Operational Item Pool by Item Type and Grade*

| Item Type | Grade | | | | | |
|---|---|---|---|---|---|---|
| | ***3*** | ***4*** | ***5*** | ***6*** | ***7*** | ***8*** |
| EBSR | 28 | 31 | 24 | 56 | 54 | 42 |
| ER | 2 | 2 | 2 | 2 | 2 | 2 |
| ETC | 48 | 58 | 52 | 43 | 47 | 42 |
| GI | 0 | 0 | 1 | 0 | 0 | 0 |
| HT | 25 | 31 | 35 | 35 | 33 | 33 |
| MC | 180 | 201 | 164 | 239 | 234 | 228 |
| MI | 9 | 6 | 8 | 8 | 3 | 5 |
| MS | 10 | 25 | 21 | 35 | 55 | 35 |
| NL | 0 | 1 | 1 | 1 | 3 | 0 |

*Table 15: Mathematics Operational Item Pool by Item Type and Grade*

| Item Type | Grade | | | | | |
|---|---|---|---|---|---|---|
| | ***3*** | ***4*** | ***5*** | ***6*** | ***7*** | ***8*** |
| EQ | 248 | 267 | 265 | 241 | 206 | 175 |
| ETC | 0 | 0 | 1 | 0 | 0 | 0 |
| GI | 74 | 51 | 24 | 38 | 35 | 49 |
| MC | 90 | 60 | 58 | 117 | 76 | 140 |
| MI | 10 | 24 | 11 | 8 | 5 | 3 |
| MS | 42 | 78 | 39 | 34 | 13 | 36 |
| TI | 15 | 15 | 11 | 30 | 3 | 8 |

### 3.1.1. Embedded Field-Testing

ELA and mathematics tests began to include field test items and item clusters in the spring 2019 NH SAS using an embedded field test (EFT) design. The EFT slots are given positions within the middle of tests, such that item location and motivation effects, if they exist, would not propagate into the estimates of the item parameters. To obtain high-quality responses to the EFT items, students were unaware of which items were operational and which were EFT.

In the adaptive NH SAS ELA and mathematics tests, field test items or item clusters were randomly drawn from the field test item pool to fill out the allotted slots. Clusters consist of several item parts that require students to interact with the item in various ways. For ELA reading, 7–9 EFT items or 1 EFT item cluster and 0–2 EFT items per test were administered; for mathematics, it was 8 EFT items or 1 EFT item cluster per test, except in the segmented grade 6 test, where 8 EFT items or 1 EFT item cluster and 3 EFT items were administered.

The spring 2019 ELA and mathematics EFT items were put onto the New Hampshire reporting scale by using a fixed anchor item calibration method. The field-test items were administered in multiple AIRCore states, such as Arizona, Wyoming, New Hampshire, West Virginia, and North Dakota. All of the operational (treated as fixed anchor) and field-test items were put into a single incomplete data matrix for a multigroup IRT (MGIRT) calibration. Operational item parameters were fixed to their bank values, while field-test item parameters were estimated in a single run. If a calibration run did not converge, then the reason was investigated. Usually one or two items with negative item-total correlations were the cause. Those items were removed from the calibration and sent to the AIR content team for further action, such revision or rejection. The state group means, provided in Appendix J, were free estimations.

### 3.1.2. Operational Test Design

ELA and mathematics tests are assembled using AIR's adaptive testing algorithm. The adaptive item-selection algorithm selects items based on their content value and information value. The algorithm ensures that each student receives a unique test that adheres to the content requirements described in the NH SAS test specifications, ensuring a comparable and sufficient coverage of the content of the New Hampshire College- and Career-Readiness Standards (NH CCRS). In addition, each student's unique test assembled by the algorithm contains the items that best match students' performance level, as defined by the blueprint. The details of the adaptive item selection algorithm for NH SAS ELA and mathematics are presented in Volume 2.

### 3.1.3. Operational Item Pool Statistics

As reported in Section 2.2, a simulation approach to configure the adaptive algorithm was conducted before the operational testing window in order to maximize test score precision while meeting blueprint specifications based on the available pool of test items. The blueprint match was monitored both for simulation and operational administration. The summary of the simulation versus operational blueprint match for spring 2019 ELA and mathematics is provided in Appendix B. This summary shows that, across all grades and subjects, the vast majority of tests met the blueprint specifications with a 100% match at the reporting category level in both simulation and operational administrations. There were a few exceptions in grade 7 and grade 8 ELA operation administrations, as a small number of students took the test for the same grade in both 2018 and 2019. The test delivery system (TDS) prevents administration of any item more than once to the same student, resulting in a smaller item pool available for students retaking the same test.

The IRT statistical properties of the operational item pool used for the 2019 NH SAS are summarized in Table 16 through

Table 21 for reading and mathematics. 3PL and 2PL refer to the three-parameter logistic model and the two-parameter logistic model, respectively, while GPCM is the generalized partial credit model. Minimum, maximum, and five-point percentiles are summarized for discrimination (*a*), difficulty (*b*), and guessing (*c*) parameters for 3PL items and *a* and *b* parameters for 2 PL items. For GPCM, step parameters (*b1* and *b2*) are summarized.

*Table 16: 3PL Operational Item Parameters Five-Point Summary and Range, ELA*

| Grade | Parameter | *N* Item | Min | 5th Percentile | 25th Percentile | 50th Percentile | 75th Percentile | 95th Percentile | Max |
|---|---|---|---|---|---|---|---|---|---|
| 3 | *a* | 180 | 0.30 | 0.57 | 0.86 | 1.16 | 1.51 | 2.21 | 11.83 |
| | *b* | 180 | -2.36 | -1.89 | -1.30 | -0.83 | -0.34 | 0.42 | 1.89 |
| | *c* | 180 | 0.03 | 0.07 | 0.12 | 0.18 | 0.23 | 0.31 | 0.45 |
| 4 | *a* | 201 | 0.19 | 0.37 | 0.72 | 0.99 | 1.28 | 1.80 | 2.44 |
| | *b* | 201 | -2.84 | -1.85 | -1.32 | -0.84 | -0.20 | 0.79 | 1.98 |
| | *c* | 201 | 0.01 | 0.03 | 0.10 | 0.16 | 0.22 | 0.29 | 0.37 |
| 5 | *a* | 154 | 0.23 | 0.42 | 0.73 | 0.98 | 1.30 | 1.67 | 2.49 |
| | *b* | 154 | -2.03 | -1.38 | -0.79 | -0.34 | 0.28 | 1.01 | 2.54 |
| | *c* | 154 | 0.03 | 0.06 | 0.13 | 0.17 | 0.22 | 0.31 | 0.42 |
| 6 | *a* | 239 | 0.18 | 0.38 | 0.68 | 0.95 | 1.22 | 1.68 | 3.73 |
| | *b* | 239 | -2.33 | -1.07 | -0.42 | 0.18 | 0.75 | 1.52 | 5.67 |
| | *c* | 239 | 0.01 | 0.06 | 0.12 | 0.18 | 0.24 | 0.33 | 0.42 |
| 7 | *a* | 234 | 0.11 | 0.43 | 0.66 | 0.84 | 1.12 | 1.56 | 2.76 |
| | *b* | 234 | -1.98 | -1.04 | -0.19 | 0.38 | 0.85 | 1.93 | 7.40 |
| | *c* | 234 | 0.01 | 0.03 | 0.11 | 0.17 | 0.23 | 0.33 | 0.40 |
| 8 | *a* | 228 | 0.05 | 0.39 | 0.71 | 0.92 | 1.14 | 1.48 | 2.04 |
| | *b* | 228 | -1.39 | -0.92 | -0.18 | 0.29 | 1.10 | 2.16 | 3.85 |
| | *c* | 228 | 0.00 | 0.03 | 0.11 | 0.17 | 0.25 | 0.32 | 0.43 |

*Table 17: 2PL Operational Item Parameters Five-Point Summary and Range, ELA*

| Grade | Parameter | *N* Item | Min | 5th Percentile | 25th Percentile | 50th Percentile | 75th Percentile | 95th Percentile | Max |
|---|---|---|---|---|---|---|---|---|---|
| 3 | *a* | 105 | 0.03 | 0.40 | 0.68 | 0.84 | 1.06 | 1.39 | 1.92 |
| | *b* | 105 | -4.99 | -2.93 | -1.51 | -0.69 | -0.18 | 0.88 | 1.97 |
| 4 | *a* | 130 | 0.04 | 0.33 | 0.47 | 0.68 | 0.88 | 1.18 | 1.59 |
| | *b* | 130 | -2.96 | -2.12 | -1.19 | -0.47 | 0.38 | 2.34 | 5.31 |
| 5 | *a* | 116 | 0.19 | 0.35 | 0.56 | 0.73 | 0.95 | 1.19 | 1.34 |
| | *b* | 116 | -2.15 | -1.62 | -1.03 | -0.21 | 0.73 | 2.69 | 4.98 |
| 6 | *a* | 160 | 0.12 | 0.29 | 0.51 | 0.70 | 0.88 | 1.14 | 1.54 |
| | *b* | 160 | -2.13 | -1.59 | -0.24 | 0.34 | 1.15 | 3.46 | 6.75 |
| 7 | *a* | 181 | 0.19 | 0.28 | 0.47 | 0.69 | 0.89 | 1.26 | 1.43 |
| | *b* | 181 | -2.31 | -1.33 | -0.24 | 0.41 | 1.27 | 2.58 | 4.91 |
| 8 | *a* | 139 | 0.06 | 0.29 | 0.47 | 0.63 | 0.83 | 1.04 | 1.22 |
| | *b* | 139 | -4.60 | -1.41 | -0.02 | 0.75 | 1.46 | 3.15 | 5.82 |

*Table 18: GPCM Operational Item Parameters Five-Point Summary and Range, ELA*

| Grade | Parameter | *N* Item | Min | 5th Percentile | 25th Percentile | 50th Percentile | 75th Percentile | 95th Percentile | Max |
|---|---|---|---|---|---|---|---|---|---|
| 3 | *a* | 21 | 0.33 | 0.34 | 0.59 | 0.84 | 1.32 | 1.59 | 1.60 |
| | *b1* | 21 | -3.70 | -3.43 | -2.53 | -2.20 | -2.00 | -1.60 | 0.57 |
| | *b2* | 21 | -1.84 | -1.76 | -1.47 | -0.90 | -0.48 | 0.25 | 0.60 |
| 4 | *a* | 25 | 0.17 | 0.33 | 0.43 | 0.60 | 0.88 | 1.46 | 1.49 |
| | *b1* | 25 | -3.45 | -3.30 | -2.35 | -2.21 | -1.31 | -0.43 | -0.24 |
| | *b2* | 25 | -1.85 | -1.51 | -0.98 | -0.33 | 1.27 | 2.19 | 3.95 |
| 5 | *a* | 25 | 0.24 | 0.32 | 0.48 | 0.60 | 0.92 | 1.48 | 1.59 |
| | *b1* | 25 | -3.20 | -2.92 | -2.10 | -1.75 | -1.25 | -0.80 | 0.92 |
| | *b2* | 25 | -1.42 | -0.91 | -0.54 | -0.35 | 0.09 | 0.56 | 0.68 |
| 6 | *a* | 24 | 0.29 | 0.30 | 0.47 | 0.55 | 0.83 | 1.63 | 1.70 |
| | *b1* | 24 | -4.71 | -3.64 | -2.14 | -1.86 | -0.85 | -0.20 | 3.86 |
| | *b2* | 24 | -1.91 | -1.25 | -0.32 | 0.05 | 0.94 | 1.31 | 1.52 |
| 7 | *a* | 20 | 0.26 | 0.31 | 0.40 | 0.59 | 1.30 | 1.71 | 1.74 |
| | *b1* | 20 | -3.14 | -2.25 | -1.82 | -1.38 | -1.05 | 0.08 | 0.27 |
| | *b2* | 20 | -1.25 | -0.94 | -0.49 | 0.38 | 1.04 | 2.53 | 3.37 |
| 8 | *a* | 24 | 0.31 | 0.33 | 0.45 | 0.64 | 0.95 | 1.34 | 1.41 |
| | *b1* | 24 | -3.07 | -2.81 | -1.74 | -1.31 | -1.07 | -0.13 | 1.22 |
| | *b2* | 24 | -1.05 | -0.82 | -0.34 | -0.08 | 0.60 | 1.07 | 2.45 |

*Table 19: 3PL Operational Item Parameters Five-Point Summary and Range, Mathematics*

| Grade | Parameter | *N* Item | Min | 5th Percentile | 25th Percentile | 50th Percentile | 75th Percentile | 95th Percentile | Max |
|---|---|---|---|---|---|---|---|---|---|
| 3 | *a* | 90 | 0.68 | 0.89 | 1.22 | 1.56 | 1.92 | 2.64 | 3.30 |
| | *b* | 90 | -4.41 | -3.67 | -2.73 | -2.37 | -1.83 | -1.36 | -0.75 |
| | *c* | 90 | 0.01 | 0.05 | 0.12 | 0.19 | 0.25 | 0.40 | 0.59 |
| 4 | *a* | 60 | 0.39 | 0.58 | 1.01 | 1.22 | 1.53 | 2.09 | 2.97 |
| | *b* | 60 | -3.87 | -3.19 | -2.47 | -2.03 | -1.31 | 0.00 | 0.62 |
| | *c* | 60 | 0.05 | 0.08 | 0.13 | 0.18 | 0.27 | 0.40 | 0.53 |
| 5 | *a* | 58 | 0.22 | 0.43 | 0.82 | 1.03 | 1.41 | 1.98 | 2.33 |
| | *b* | 58 | -5.70 | -2.50 | -1.71 | -1.01 | -0.37 | 0.23 | 1.15 |
| | *c* | 58 | 0.04 | 0.08 | 0.14 | 0.20 | 0.24 | 0.36 | 0.56 |
| 6 | *a* | 117 | 0.18 | 0.44 | 0.70 | 0.98 | 1.16 | 1.58 | 4.79 |
| | *b* | 117 | -4.41 | -2.37 | -1.22 | -0.26 | 0.29 | 1.32 | 4.73 |
| | *c* | 117 | 0.01 | 0.06 | 0.13 | 0.18 | 0.23 | 0.36 | 0.40 |
| 7 | *a* | 76 | 0.10 | 0.47 | 0.65 | 0.84 | 1.04 | 1.73 | 7.62 |
| | *b* | 76 | -4.09 | -1.70 | -0.26 | 0.76 | 1.68 | 2.27 | 2.91 |
| | *c* | 76 | 0.05 | 0.07 | 0.12 | 0.18 | 0.26 | 0.36 | 0.47 |
| 8 | *a* | 140 | 0.08 | 0.36 | 0.52 | 0.75 | 0.96 | 1.24 | 2.76 |
| | *b* | 140 | -2.15 | -1.34 | -0.18 | 1.04 | 2.11 | 3.14 | 5.90 |
| | *c* | 140 | 0.02 | 0.04 | 0.12 | 0.19 | 0.25 | 0.38 | 0.51 |

*Table 20: 2PL Operational Item Parameters Five-Point Summary and Range, Mathematics*

| Grade | Parameter | *N* Item | Min | 5th Percentile | 25th Percentile | 50th Percentile | 75th Percentile | 95th Percentile | Max |
|---|---|---|---|---|---|---|---|---|---|
| 3 | *a* | 377 | 0.27 | 0.77 | 1.20 | 1.48 | 1.76 | 2.13 | 2.60 |
| | *b* | 377 | -5.61 | -3.27 | -2.69 | -2.28 | -1.85 | -1.22 | 1.25 |
| 4 | *a* | 429 | 0.35 | 0.70 | 0.98 | 1.22 | 1.47 | 1.76 | 2.29 |
| | *b* | 429 | -3.42 | -2.78 | -2.10 | -1.56 | -1.03 | -0.27 | 0.84 |
| 5 | *a* | 336 | 0.20 | 0.57 | 0.81 | 1.03 | 1.25 | 1.55 | 2.06 |
| | *b* | 336 | -4.11 | -2.42 | -1.45 | -0.91 | -0.40 | 0.49 | 2.72 |
| 6 | *a* | 337 | 0.10 | 0.53 | 0.76 | 0.95 | 1.13 | 1.42 | 1.92 |
| | *b* | 337 | -3.58 | -2.15 | -0.89 | -0.14 | 0.51 | 1.39 | 6.97 |
| 7 | *a* | 250 | 0.25 | 0.46 | 0.67 | 0.89 | 1.10 | 1.40 | 2.47 |
| | *b* | 250 | -1.47 | -0.90 | -0.03 | 0.68 | 1.55 | 2.52 | 3.65 |
| 8 | *a* | 263 | 0.11 | 0.39 | 0.58 | 0.75 | 0.88 | 1.16 | 1.72 |
| | *b* | 263 | -5.51 | -0.20 | 1.22 | 1.93 | 2.52 | 3.43 | 6.69 |

*Table 21: GPCM Operational Item Parameters Five-Point Summary and Range, Mathematics*

| Grade | Parameter | *N* Item | Min | 5th Percentile | 25th Percentile | 50th Percentile | 75th Percentile | 95th Percentile | Max |
|---|---|---|---|---|---|---|---|---|---|
| 3 | *a* | 12 | 0.80 | 0.85 | 0.98 | 1.16 | 1.44 | 1.64 | 1.66 |
| | *b1* | 12 | -3.41 | -3.07 | -2.46 | -1.99 | -1.69 | -1.02 | -0.68 |
| | *b2* | 12 | -2.85 | -2.79 | -2.68 | -2.01 | -1.31 | -0.46 | -0.19 |
| 4 | *a* | 6 | 0.46 | 0.48 | 0.63 | 0.91 | 1.02 | 1.18 | 1.23 |
| | *b1* | 6 | -4.02 | -3.77 | -2.99 | -2.35 | -0.39 | 0.33 | 0.40 |
| | *b2* | 6 | -1.99 | -1.91 | -1.70 | -1.56 | -1.05 | -0.45 | -0.30 |
| 5 | *a* | 15 | 0.51 | 0.52 | 0.64 | 0.77 | 0.83 | 1.15 | 1.19 |
| | *b1* | 15 | -2.10 | -2.09 | -1.79 | -1.15 | -0.44 | 0.08 | 0.47 |
| | *b2* | 15 | -2.69 | -2.38 | -1.43 | -0.37 | -0.05 | 0.50 | 0.85 |
| 6 | *a* | 14 | 0.56 | 0.64 | 0.78 | 0.83 | 0.86 | 0.89 | 0.92 |
| | *b1* | 14 | -1.84 | -1.63 | -1.02 | -0.26 | 1.11 | 2.22 | 2.38 |
| | *b2* | 14 | -2.01 | -1.21 | -0.31 | 0.08 | 0.70 | 1.68 | 2.29 |
| 7 | *a* | 12 | 0.50 | 0.52 | 0.57 | 0.67 | 0.80 | 1.10 | 1.22 |
| | *b1* | 12 | -1.17 | -0.92 | -0.08 | 0.62 | 1.36 | 2.58 | 3.67 |
| | *b2* | 12 | -0.15 | 0.07 | 0.76 | 1.08 | 1.59 | 2.65 | 2.86 |
| 8 | *a* | 8 | 0.37 | 0.40 | 0.49 | 0.57 | 0.64 | 0.72 | 0.74 |
| | *b1* | 8 | -1.38 | -1.24 | -0.86 | -0.59 | 1.24 | 1.95 | 2.23 |
| | *b2* | 8 | -0.77 | -0.50 | 0.64 | 1.31 | 1.83 | 2.24 | 2.27 |

## 3.2. SCIENCE ITEM BANK AND TEST DESIGN

AIR works with a group of states to develop 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 has also built the AIRCore science item bank in partnership with these states. These AIR-owned items comprise a substantial part of the item bank and are shared with partner states. All of 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) survived and were available as operational items in 2019. The 2019 operational AIRCore items comprise 85 item clusters (performance tasks) and 82 stand-alone items. In 2019, 347 item clusters and stand-alone items were field-tested, of which 265 have survived rubric validation and item data review. For AIRCore, 75 item clusters and stand-alone items were field-tested, of which 61 will be added to the operational AIRCore item pool.

The New Hampshire science assessment uses only AIRCore items, but because the AIRCore science items are part of the larger across-state science item bank, the latter will be described in this section of the technical report. The larger science item bank is in use for operational accountability tests in seven states (2019), including New Hampshire. An additional three states 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. Here, note that best practices have been implemented at every turn:

- The goals, uses, and claims that the test is designed to support were identified in a collaborative meeting on 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 that helped co-author the NGSS. Two nationally recognized psychometricians also participated.
- AIR staff and participating states collaborate to develop items and test specifications. The item specifications are generally accompanied by sample items meeting those specifications. All specifications and sample items are reviewed by state content experts and committees of educators in at least one of the states.
- Items have been 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 is field tested, and items with questionable data are reviewed further by committees of educators.

### 3.2.1. Field Testing

All items that were part of the 2019 operational pool were field-tested in 2018 as described in Section 3.2.2. Additional items were field-tested in 2019, which are described in section 3.2.1.2.

### 3.2.2. 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 conducted an independent field test in which all students participated and were administered a full set of items, but no scores were reported (Connecticut and Rhode Island). In the remaining four states (New Hampshire, West Virginia, Utah, and Vermont), 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 22 presents the number of item clusters and stand-alone items administered in each grade for each state.

*Table 22: Number of Item Clusters and Stand-Alone Items Administered in Spring 2018, Science*

| Grade Band/Item Type | CT | HI | MSSA (RI, VT) | NH | OR | UT | WV | WY | Whole 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), field-test forms were constructed using a balanced incomplete design and spiraled across students. For the independent and operational field tests (except Utah), including New Hampshire, items were administered under a LOFT design. The difference between the test design for the independent field tests and operational field tests depended upon the test blueprint. For the independent field tests, the only blueprint constraint imposed was that students received four stand-alone items and two clusters 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 NH SAS science test is discussed in Section 3.2.4.

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 23 through Table 25 present the number of 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 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 23 presents the results for elementary schools, Table 24 the results for middle schools, and Table 25 the results for high schools. 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 items in some states.

*Table 23: Number of Common Items for Elementary School Administered in Spring 2018, Science*

| | State | Connecticut | Hawaii | MSSA (RI, VT) | 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 (RI, VT) | 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 (RI, VT) | 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 (RI, VT) | 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 24: Number of Common Items for Middle School Administered in Spring 2018, Science*

| | State | Connecticut | Hawaii | MSSA (RI, VT) | 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 (RI, VT) | 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 (RI, VT) | 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 |
| | UT | 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 (RI, VT) | 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 25: Number of Common Items for High School Administered in Spring 2018, Science*

| | State | Connecticut | Hawaii | MSSA (RI, VT) | 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 (RI, VT) | 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 (RI, VT)** | 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 (RI, VT)** | 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.

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 (a) low-scored responses from otherwise high-scoring students and (b) 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 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 26 presents the number of AIRCore items field tested in New Hampshire, the number of items that were rejected before or during rubric validation, the number of items that were sent out to data review, and the number of items that were rejected during data review.

*Table 26: Overview of Test Administration, Rubric Validation, and Item Data Review in Spring 2018, Science*

| Grade Band/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 |
|---|---|---|---|---|---|
| **Elementary School** | **58** | **0** | **23** | **2** | **56** |
| *Clusters* | 34 | 0 | 7 | 1 | 33 |
| *Stand-Alone items* | 24 | 0 | 16 | 1 | 23 |
| **Middle School** | **55** | **0** | **24** | **2** | **53** |
| *Clusters* | 30 | 0 | 11 | 1 | 29 |
| *Stand-Alone items* | 25 | 0 | 13 | 1 | 24 |
| **High School** | **60** | **2** | **31** | **2** | **56** |
| *Clusters* | 33 | 2 | 15 | 0 | 31 |
| *Stand-Alone items* | 27 | 0 | 16 | 2 | 25 |
| **Total** | **173** | **2** | **78** | **6** | **165** |

Table 27 summarizes the item pool that was used in New Hampshire for each of three science disciplines.

*Table 27: Overview of Items Field-Tested and Operationally Scored in Spring 2018, Science*

| Grade Band/Item Type | Items Field-Tested in Spring 2018 | | | | Scored Operational Items | | | |
|---|---|---|---|---|---|---|---|---|
| | Total | Earth and Space Sciences | Life Sciences | Physical Sciences | Total | Earth and Space Sciences | Life Sciences | Physical Sciences |
| **Elementary School** | **58** | **18** | **20** | **20** | **56** | **18** | **19** | **19** |
| *Cluster* | 34 | 11 | 12 | 11 | 33 | 11 | 11 | 11 |
| *Stand-Alone* | 24 | 7 | 8 | 9 | 23 | 7 | 8 | 8 |
| **Middle School** | **55** | **17** | **18** | **20** | **52*** | **14** | **18** | **20** |
| *Cluster* | 30 | 10 | 7 | 13 | 28 | 8 | 7 | 13 |
| *Stand-Alone* | 25 | 7 | 11 | 7 | 24 | 6 | 11 | 7 |
| **High School** | **60** | **19** | **23** | **18** | **56** | **17** | **22** | **17** |
| *Cluster* | 33 | 7 | 16 | 10 | 31 | 7 | 15 | 9 |
| *Stand-Alone* | 27 | 12 | 7 | 8 | 25 | 10 | 7 | 8 |
| **Total** | **173** | **54** | **61** | **58** | **164** | **49** | **59** | **56** |

**Note:* Item 2261 had a student-facing issue that was fixed during the testing window; no student took it after fixing the issue.

### 3.2.3. 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 within 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 28 presents the numbers of field-tested item clusters and stand-alone items administered in each grade for each state.

*Table 28: Number of Field-Test Items Administered in Spring 2019, Science*

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

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

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 LOFT design, where the only blueprint constraint imposed was that students received four stand-alone items and two clusters 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. In New Hampshire, a student received either an item cluster or a set of five stand-alone items as a field-test set. 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. The test design for the NH SAS science test is discussed in Section 3.2.4.

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 29 through Table 31 present the number of clusters 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 29 presents the results for elementary schools, Table 30 presents the results for middle schools, and Table 31 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 29: Number of Common Field-Test Items for Elementary School Administered in Spring 2019, Science*

| | State | Connecticut | Hawaii | Idaho | MSSA (RI, VT) | 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 (RI, VT)** | 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)** |

| | State | Connecticut | Hawaii | Idaho | MSSA (RI, VT) | New Hampshire | Oregon | West Virginia | Wyoming |
|---|---|---|---|---|---|---|---|---|---|
| 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 (RI, VT) | 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 (RI, VT) | 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 30: Number of Common Field-Test Items for Middle School Administered in Spring 2019, Science*

| | State | Connecticut | Hawaii | Idaho | MSSA (RI, VT) | 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 (RI, VT) | 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 (RI, VT) | 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)** |

| | State | Connecticut | Hawaii | Idaho | MSSA (RI, VT) | New Hampshire | Oregon | West Virginia | Wyoming |
|---|---|---|---|---|---|---|---|---|---|
| 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 (RI, VT) | 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 31: Number of Common Field-Test Items for High School Administered in Spring 2019, Science*

| | State | Connecticut | Hawaii | Idaho | MSSA (RI, VT) | 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 (RI, VT) | 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 (RI, VT) | 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 (RI, VT) | 20 | 9 | - | **6 (5)** | 6 | 12 | - | 14 |
| | NH | 8 | 0 | - | 6 | **11 (10)** | 1 | - | 7 |
| | OR | 22 | 17 | - | 13 | 1 | **1 (1)** | - | 8 |
| | WV | - | - | - | - | - | - | **-** | - |

| State | Connecticut | Hawaii | Idaho | MSSA (RI, VT) | New Hampshire | Oregon | West Virginia | Wyoming |
|---|---|---|---|---|---|---|---|---|
| **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 of this volume.

Following essentially the same process as explained in Section 3.2.1.1, items went through a substantial validation process. The modifications to the process followed in 2018 were minor:

- 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 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), Rhode Island, Vermont, New Hampshire, Oregon, 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 4.5 on DIF.

Table 32 presents the numbers of AIRCore items field tested in New Hampshire or another state, items rejected before or during rubric validation, items sent out for data review, and items rejected during data review.

*Table 32: Overview of Science Administration, Rubric Validation, and Item Data Review*

| 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 (18)** | **2 (0)** | **72 (16)** | **24 (0)** | **91 (18)** |
| *Clusters* | 50 (0) | 1 (0) | 16 (0) | 10 (0) | 39 (0) |
| *Stand-Alone items* | 67 (18) | 1 (0) | 56 (16) | 14 (0) | 52 (18) |
| **Middle School** | **127 (28)** | **6 (5)** | **66 (15)** | **21 (1)** | **97 (19)** |
| *Clusters* | 38 (4) | 1 (1) | 12 (0) | 5 (0) | 29 (0) |
| *Stand-Alone items* | 89 (24) | 5 (4) | 54 (15) | 16 (1) | 68 (19) |
| **High School** | **103 (29)** | **6 (3)** | **52 (12)** | **15 (1)** | **80 (24)** |
| *Clusters* | 35 (2) | 2 (1) | 15 (0) | 5 (0) | 26 (0) |
| *Stand-Alone items* | 68 (27) | 4 (2) | 37 (12) | 10 (1) | 54 (24) |
| **Total** | **347 (75)** | **14 (8)** | **190 (43)** | **60 (2)** | **268 (61)** |

*Note:* AIRCore 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 33 summarizes the AIRCore item pool after adding the items that were field-tested in 2019 and survived rubric validation and item data review.

*Table 33: Overview of Combined AIRCore Item Pool in Spring 2019, Science*

| Grade Band/Item Type | AIRCore Item Pool | | | |
|---|---|---|---|---|
| | Total | Earth and Space Sciences | Life Sciences | Physical Sciences |
| **Elementary School** | **79** | **23** | **28** | **28** |
| *Cluster* | 32 | 11 | 11 | 10 |
| *Stand-Alone* | 47 | 12 | 17 | 18 |
| **Middle School** | **70** | **23** | **26** | **21** |
| *Cluster* | 25 | 8 | 6 | 11 |
| *Stand-Alone* | 45 | 15 | 20 | 10 |
| **High School** | **79** | **17** | **46** | **16** |
| *Cluster* | 28 | 6 | 14 | 8 |
| *Stand-Alone* | 51 | 11 | 32 | 8 |
| **Total** | **228** | **63** | **100** | **65** |

### 3.2.4. Operational Test Design

For science, tests were assembled under a LOFT 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. By assigning weights of zero to the information value of an item with respect to the underlying latent variable, the items are solely selected based on their contribution to meeting the blueprint. The blueprint for science is given in Table 34 through Table 36.

*Table 34: Science Test Blueprint, Grade 5 Science*

| Grade 5 | Min Clusters | Max Clusters | Min Stand-Alone Items | Max Stand-Alone Items | Min Clusters + Min Stand-Alone Items | Max Clusters + Max Stand-Alone Items |
|---|---|---|---|---|---|---|
| **Discipline – Physical Science, 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** |

| Grade 5 | Min Clusters | Max Clusters | Min Stand-Alone Items | Max Stand-Alone Items | Min Clusters + Min Stand-Alone Items | Max Clusters + Max Stand-Alone Items |
|---|---|---|---|---|---|---|
| 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 & 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 & Properties of Matter | 0 | 1 | 0 | 1 | 0 | 1 |
| 5-PS1-2: Structure & Properties of Matter | 0 | 1 | 0 | 1 | 0 | 1 |
| 5-PS1-3: Structure & Properties of Matter | 0 | 1 | 0 | 1 | 0 | 1 |
| 5-PS1-4: Structure & Properties of Matter | 0 | 1 | 0 | 1 | 0 | 1 |
| **Discipline – Life Science, 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 & 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 & 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 |

| Grade 5 | Min Clusters | Max Clusters | Min Stand-Alone Items | Max Stand-Alone Items | Min Clusters + Min Stand-Alone Items | Max Clusters + Max Stand-Alone Items |
|---|---|---|---|---|---|---|
| 3-LS4-4: Ecosystems* | 0 | 1 | 0 | 1 | 0 | 1 |
| **Discipline – Earth and Space Science, PE Total = 13** | **2** | **2** | **4** | **4** | **6** | **6** |
| **DCI – Earth's Systems** | **0** | **1** | **0** | **2** | **0** | **3** |
| 3-ESS2-1: Weather & Climate | 0 | 1 | 0 | 1 | 0 | 1 |
| 3-ESS2-2: Weather & Climate | 0 | 1 | 0 | 1 | 0 | 1 |
| 4-ESS2-1: Earth's Systems & Processes | 0 | 1 | 0 | 1 | 0 | 1 |
| 4-ESS2-2: Earth's Systems & 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 & Climate* | 0 | 1 | 0 | 1 | 0 | 1 |
| 4-ESS3-2: Earth's Systems & 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 & 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 35: Science Test Blueprint, Grade 8 Science*

| Grade 8 | Min Clusters | Max Clusters | Min Stand-Alone Items | Max Stand-Alone Items | Min Clusters + Stand-Alone Items | Max Clusters + Stand-Alone Items |
|---|---|---|---|---|---|---|
| **Discipline – Physical Science, PE Total = 19** | **2** | **2** | **4** | **4** | **6** | **6** |
| **DCI – Matter and Its Interactions** | **0** | **1** | **0** | **2** | **0** | **3** |
| MS-PS1-1: Structure & Properties of Matter | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-PS1-2: Chemical Reactions | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-PS1-3: Structure & Properties of Matter | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-PS1-4: Structure & Properties of Matter | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-PS1-5: Chemical Reactions | 0 | 1 | 0 | 1 | 0 | 1 |

| Grade 8 | Min Clusters | Max Clusters | Min Stand-Alone Items | Max Stand-Alone Items | Min Clusters + Stand-Alone Items | Max Clusters + Stand-Alone Items |
|---|---|---|---|---|---|---|
| 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 & Interactions* | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-PS2-2: Forces & Interactions | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-PS2-3: Forces & Interactions | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-PS2-4: Forces & Interactions | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-PS2-5: Forces & 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 & Electromagnetic Radiation | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-PS4-2: Waves & Electromagnetic Radiation | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-PS4-3: Waves & Electromagnetic Radiation | 0 | 1 | 0 | 1 | 0 | 1 |
| **Discipline – Life Science, 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 & Energy | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-LS1-7: Matter & 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 & Energy | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-LS2-2: Interdependent Relationships in Ecosystems | 0 | 1 | 0 | 1 | 0 | 1 |

| Grade 8 | Min Clusters | Max Clusters | Min Stand-Alone Items | Max Stand-Alone Items | Min Clusters + Stand-Alone Items | Max Clusters + Stand-Alone Items |
|---|---|---|---|---|---|---|
| MS-LS2-3: Matter & Energy | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-LS2-4: Matter & Energy | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-LS2-5: Interdependent Relationships in Ecosystems* | 0 | 1 | 0 | 1 | 0 | 1 |
| **DCI – Heredity: 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 & Adaptation | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-LS4-2: Natural Selection & Adaptation | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-LS4-3: Natural Selection & Adaptation | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-LS4-4: Natural Selection & Adaptation | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-LS4-5: Growth, Development, Reproduction | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-LS4-6: Natural Selection & Adaptation | 0 | 1 | 0 | 1 | 0 | 1 |
| **Discipline – Earth and Space Science, 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 & Climate | 0 | 1 | 0 | 1 | 0 | 1 |
| MS-ESS2-6: Weather & 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 |
| MS-ESS3-5: Weather & Climate | 0 | 1 | 0 | 1 | 0 | 1 |
| **Total PE = 55** | **6** | **6** | **12** | **12** | **18** | **18** |

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

*Table 36: Science Test Blueprint, Grade 11 Science*

| Grade 11 | Min Clusters | Max Clusters | Min Stand-Alone Items | Max Stand-Alone Items | Min Clusters + Stand-Alone Items | Max Clusters + Stand-Alone Items |
|---|---|---|---|---|---|---|
| **Discipline – Physical Science, 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 |
| 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 Science, 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 |

| Grade 11 | Min Clusters | Max Clusters | Min Stand-Alone Items | Max Stand-Alone Items | Min Clusters + Stand-Alone Items | Max Clusters + Stand-Alone Items |
|---|---|---|---|---|---|---|
| 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 |
| 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 Science, 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 |
| 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 |

| Grade 11 | Min Clusters | Max Clusters | Min Stand-Alone Items | Max Stand-Alone Items | Min Clusters + Stand-Alone Items | Max Clusters + Stand-Alone Items |
|---|---|---|---|---|---|---|
| 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 NH SAS science test design were as follows. There were four segments on the test, each with its own item pool. The segments and respective item pools were:

- Life Sciences;
- Earth and Space Sciences;
- Physical Sciences; and
- Embedded field-test segment (all three disciplines).

For the three segments corresponding to science disciplines, which constituted the operational segments of the test, a student received two clusters and four stand-alone items of the respective discipline (see also the Min and Max cluster values of the blueprint in Table 34 through Table 36 at the discipline level). The fourth segment was an unscored EFT segment consisting of either one cluster or a set of five stand-alone items from the AIRCore field test pool. The order of the four segments was randomized across students. Further 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 34 through Table 36); 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 items in total could be sampled from the same Disciplinary Core Idea, as indicated by the Min and Max values in the rows representing Disciplinary Core Ideas.

# 4. Field Test Classical Analyses Overview

Following test administration, all field-test items are evaluated for discrimination, difficulty, and differential item functioning (DIF). In addition, distractor analysis is conducted on multiple-choice (MC) items in English language arts (ELA) and mathematics, and response time analysis is performed for science items. Any items flagged for out-of-range statistics are reviewed by the AIR content and psychometric staff; poorly performing items are then rejected from the item bank. The criteria for flagging and reviewing ELA and mathematics items is provided in Table 37.

*Table 37: Thresholds for Flagging Items in Classical Item Analysis, ELA, and Mathematics*

| Analysis Type | Flagging Criteria |
|---|---|
| Item Discrimination | Point biserial correlation for the correct response is < 0.20. |
| Distractor Analysis | Point biserial correlation for any distractor response is > 0. |
| Item Difficulty (MC items) | The proportion of students (*p*-value) is < 0.15 or > 0.90. |
| Item Difficulty (non-MC items) | Relative mean is < 0.10 or > 0.95. |

As explained in Section 0 of this volume, science items administered as field-test items in 2018 and 2019 in New Hampshire or any of the states that signed the Memorandum of Understanding 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, Rhode Island, Vermont, Oregon, and West Virginia). Furthermore, for the computation of differential item functioning statistics, the data of all states with an operational or independent field test were combined in order to obtain a sufficient number of students for each demographic group. The criteria for flagging and reviewing items is provided in **Error! Not a valid bookmark self-reference.**, and a description of the statistics is provided below. Items that were flagged for data review were reviewed by a committee, as explained in Section 0.

*Table 38: Thresholds for Flagging Items in Classical Item Analysis, Science*

| 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 a cluster). |
| Item Difficulty (Stand-Alones) | Average *p*-value < .15 or > 0.95 (across the assertions within a stand-alone). |
| Timing (Clusters) | Percentile 80* > 15 minutes. |
| Timing (Stand-Alones) | Percentile 80* > 3 minutes. |
| Timing | Assertions per (percentile 80*) minute < 0.5 |

| Analysis Type | Flagging Criteria |
|---|---|
| DIF (Clusters) | Two or more assertions show 'C' DIF in the same direction |
| DIF (Stand-Alones) | One or more assertions show 'C' DIF in the same direction |

*A percentile 80 of *x* minutes: 80% of the students spent *x* minutes or fewer 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. In general, the higher the value, the better the item was able to differentiate between high- and low-achieving students. For science, the discrimination index was calculated for each assertion 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 were either very difficult or very easy were flagged for review but were not necessarily removed if they were grade-level appropriate and aligned with the test specifications. For multiple-choice items, the proportion of students in the sample selecting the correct answer (the *p*-value) was computed in addition to the proportion of students selecting incorrect responses. For constructed-response items, item difficulty was calculated using the item's relative mean score and the average proportion correct (analogous to *p*-value and indicating the ratio of the item's mean score divided by the maximum possible score points). Conventional item *p*-values are summarized in Section 4.6. For science, for all assertions for all items, the proportion of students for which the assertion was true (the *p*-value) was computed. AIR also aggregated the *p*-value at the item level by computing the average *p*-value across all assertions of an item. The average *p*-values are summarized in Section 4.6.

## 4.3. ELA AND MATHEMATICS DISTRACTOR ANALYSIS

Distractor analysis for multiple-choice items is used to identify items that may have marginal distractors, ambiguous correct responses, the wrong key, or more than one correct answer that attracts high-scoring students. For MC items, the correct response should be the most frequently selected option by high-scoring students. The discrimination value of the correct response should be substantial and positive, and the discrimination values for distractors should be lower and, generally, negative.

## 4.4. SCIENCE RESPONSE TIME

Given that the science clusters consist of multiple student interactions, they typically require more time for students to complete. To ensure a good balance between the amount of information provided by an item 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. A field-test item was flagged for additional review when:

- percentile 80 > 15 minutes, if the item was a cluster;

- percentile 80 > 3 minutes, if the item was a stand-alone; or
- assertions per (percentile 80) minute < 0.5.

## 4.5. DIFFERENTIAL ITEM FUNCTIONING ANALYSIS

Differential item functioning (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. The Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, 2014) state that when sample sizes permit, subgroup differences in performance should be examined and appropriate actions should be taken to ensure that differences in performance are not attributable to construct-irrelevant factors.

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 and 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 generalized Mantel–Haenszel (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 standardized mean difference (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 was evaluated for the embedded field test items for spring 2019 ELA and mathematics. Appendix H presents the DIF analysis results using the generalized Mantel-Haenszel (MH) procedure. Due to privacy regulations, this analysis did not include New Hampshire students. It was performed on three states (West Virginia, North Dakota, and Wyoming), which share items with New Hampshire. The generalized MH classified items into three categories (A, B, or C) for DIF, ranging from no evidence of DIF to severe DIF (Table 39). Furthermore, items were categorized positively (i.e., +A, +B, or +C), signifying that an item favored the focal group (e.g., African American/Black, Hispanic, or female), or negatively (i.e., –A, –B, or–C), signifying that an item favored the reference group (e.g., white or male). Items were flagged if their DIF statistics indicated the "C" category for any group. A DIF classification of "C" indicates that the item shows significant DIF and should be reviewed for potential content bias, differential validity, or other issues that may reduce item fairness.

Due to the limited number of students in some groups, DIF analyses were performed for the following groups in ELA and mathematics:

- Male/Female;
- White/African-American;
- White/Hispanic;
- White/Asian or Pacific Islander;
- White/American Indian or Alaskan Native; and
- White/Multi-racial.

*Table 39: DIF Classification Rules, ELA and Mathematics*

| **Dichotomous Items** | |
|---|---|
| *Category* | *Rule* |
| C | $MH_{X^2}$ is significant and $\lvert\hat{\Delta}_{MH}\rvert \geq 1.5$ |
| B | $MH_{X^2}$ is significant and $1 \leq \lvert\hat{\Delta}_{MH}\rvert < 1.5$ |
| A | $MH_{X^2}$ is not significant or $\lvert\hat{\Delta}_{MH}\rvert < 1$ |
| **Polytomous Items** | |
| *Category* | *Rule* |
| C | $MH_{X^2}$ is significant and $\lvert SMD\rvert / \lvert SD\rvert > .25$ |
| B | $MH_{X^2}$ is significant and $.17 < \lvert SMD\rvert / \lvert SD\rvert \leq .25$ |
| A | $MH_{X^2}$ is not significant or $\lvert SMD\rvert / \lvert SD\rvert \leq .17$ |

In science, a similar DIF categorization rule was applied at the assertion level. Items were flagged for review according to additional item-level criteria set based on the results of the assertion-level categorizations. The item level criteria also considered the item type (i.e., cluster or standalone). All science DIF statistics were computed after the testing windows closes. All DIF statistics were computed after the testing windows closed. For states with the field test segment embedded in their legacy science test, business rules for data review are defined on the DIF statistics computed on student responses in the state that owns the item. For items owned by states with an operational or independent field test, as well as for the AIRCore science items, the data of all operational and independent states were combined in order to minimize the number of items with insufficient sample sizes for one or more demographic groups. Note that the student background variables used to define groups in the DIF analyses were not available for New Hampshire students; therefore, the responses of New Hampshire students are not included in the computation of DIF statistics for AIRCore science items. A minimum sample of 200 responses (Zwick, 2012) per item in each subgroup was applied. DIF analyses were performed for the following groups (*not all items had sufficient sample sizes of 200 or more for DIF analyses in these groups):

- Male/Female

- White/African-American*
- White/Hispanic*
- White/Asian and Pacific Islander*
- White/American Indian and Alaskan Natives*
- English Learner (EL)/Non-EL*
- Economically Disadvantaged/Non-Economically Disadvantaged*
- Special Education/Non-Special Education*

DIF statistics were calculated at the assertion level. 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 serve DIF. Classification rules shown in Table 40 were applied to the assertions in each item (cluster or stand-alone). Furthermore, assertions were categorized positively (i.e., +A, +B, or +C), signifying that an item favored the focal group (e.g., African America/Black, Hispanic, or female), or negatively (i.e., -A, -B, or –C), signifying that an item favored the reference group (e.g., white or male).

An item is flagged for data review according to the following criterion:

- 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 40: DIF Classification Rules, Science*

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

Compared to 2018, the business rule for flagging items for DIF was more liberal. Specifically, in 2019 a cluster was flagged for DIF whenever two or more assertions showed ‘C’ DIF in the same direction for a demographic group, regardless of the number of assertions in the cluster. In 2018, the same rule was followed for clusters with fewer than 10 assertions, but a stricter criterion of 3 ‘C’ DIF flags in the same direction was used for clusters with 10 or more assertions. 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).

Content experts reviewed all items flagged on the basis of DIF statistics. They were encouraged to discuss these items and were asked to decide whether each item should be excluded from the pool of potential items given its performance. Not all items that exhibit DIF are biased; characteristics of the education system may also lead to DIF. For example, if schools in certain areas are less likely to offer rigorous mathematics classes, students at those schools might perform

more poorly on mathematics items than would be expected, given their proficiency on other types of items. In this example, it is not the item that exhibits bias but, rather, the instruction. However, DIF can indicate bias, so all items that are flagged for DIF were reviewed by content experts for potential bias.

## 4.6. Classical Analyses Results

This section presents a summary of results from the classical item analysis of the 2019 field-tested AIRCore items in ELA, mathematics, and science.

Table 41 through Table 45 provide summaries of the *p*-values and biserial correlations by percentile as well as the range by grade and subject for items field tested in ELA, mathematics, and science. The statistics were computed across the values of the items. The distribution of the operational item *p*-values presents a desired variability across the scale in all grades and all subjects. Note that the column *Total FT Items* in ELA and mathematics shows the number of items in the field test pool that were used in the computation of the percentiles. The three-dimension scores for writing items are counted as three items in ELA. For science, the *p*-values are computed on the combined data of states that used AIRCore items. The average values across the assertions within an item was used in the computation of the percentiles and ranges.

*Table 41: Distribution of P-Values for Field-Test Items, ELA*

| Grade | Total FT Items | Min | 5th Percentile | 25th Percentile | 50th Percentile | 75th Percentile | 95th Percentile | Max |
|---|---|---|---|---|---|---|---|---|
| 3 | 101 | 0.05 | 0.16 | 0.33 | 0.44 | 0.57 | 0.73 | 0.83 |
| 4 | 76 | 0.11 | 0.27 | 0.38 | 0.49 | 0.62 | 0.77 | 0.86 |
| 5 | 77 | 0.11 | 0.20 | 0.40 | 0.51 | 0.61 | 0.79 | 0.89 |
| 6 | 92 | 0.09 | 0.20 | 0.36 | 0.51 | 0.63 | 0.77 | 0.82 |
| 7 | 76 | 0.12 | 0.15 | 0.31 | 0.44 | 0.59 | 0.74 | 0.85 |
| 8 | 75 | 0.20 | 0.23 | 0.37 | 0.50 | 0.58 | 0.77 | 0.87 |

*Table 42: Distribution of Item Biserial Correlations for Field-Test Items, ELA*

| Grade | Total FT Items | Min | 5th Percentile | 25th Percentile | 50th Percentile | 75th Percentile | 95th Percentile | Max |
|---|---|---|---|---|---|---|---|---|
| 3 | 101 | 0.05 | 0.11 | 0.26 | 0.33 | 0.42 | 0.48 | 0.55 |
| 4 | 76 | 0.02 | 0.20 | 0.28 | 0.37 | 0.43 | 0.53 | 0.58 |
| 5 | 76 | 0.02 | 0.20 | 0.28 | 0.37 | 0.43 | 0.53 | 0.58 |
| 6 | 92 | 0.05 | 0.12 | 0.30 | 0.36 | 0.42 | 0.50 | 0.53 |
| 7 | 76 | 0.09 | 0.13 | 0.25 | 0.33 | 0.40 | 0.46 | 0.49 |
| 8 | 75 | 0.02 | 0.10 | 0.26 | 0.35 | 0.42 | 0.50 | 0.55 |

*Table 43: Distribution of P-Values for Field-Test Items, Mathematics*

| Grade | Total FT Items | Min | 5th Percentile | 25th Percentile | 50th Percentile | 75th Percentile | 95th Percentile | Max |
|---|---|---|---|---|---|---|---|---|
| 3 | 93 | 0.06 | 0.09 | 0.37 | 0.53 | 0.69 | 0.85 | 0.94 |
| 4 | 106 | 0.06 | 0.15 | 0.31 | 0.44 | 0.59 | 0.81 | 0.90 |
| 5 | 80 | 0.07 | 0.16 | 0.34 | 0.49 | 0.59 | 0.76 | 0.92 |
| 6 | 141 | 0.02 | 0.06 | 0.21 | 0.38 | 0.60 | 0.82 | 0.90 |
| 7 | 57 | 0.02 | 0.04 | 0.11 | 0.31 | 0.46 | 0.67 | 0.83 |
| 8 | 85 | 0.01 | 0.05 | 0.15 | 0.35 | 0.50 | 0.68 | 0.85 |

*Table 44: Distribution of Item Biserial Correlations for Field-Test Items, Mathematics*

| Grade | Total FT Items | Min | 5th Percentile | 25th Percentile | 50th Percentile | 75th Percentile | 95th Percentile | Max |
|---|---|---|---|---|---|---|---|---|
| 3 | 93 | 0.16 | 0.25 | 0.37 | 0.45 | 0.51 | 0.56 | 0.63 |
| 4 | 106 | 0.12 | 0.26 | 0.33 | 0.44 | 0.52 | 0.57 | 0.61 |
| 5 | 80 | 0.05 | 0.19 | 0.35 | 0.42 | 0.49 | 0.56 | 0.60 |
| 6 | 141 | -0.03 | 0.17 | 0.33 | 0.41 | 0.48 | 0.56 | 0.62 |
| 7 | 57 | 0.11 | 0.18 | 0.33 | 0.38 | 0.49 | 0.55 | 0.58 |
| 8 | 85 | 0.09 | 0.18 | 0.32 | 0.40 | 0.47 | 0.56 | 0.61 |

*Table 45: Distribution of P-Values for Field-Test Items, Science*

| Grade | Total FT Items | Min | 5th Percentile | 25th Percentile | 50th Percentile | 75th Percentile | 95th Percentile | Max |
|---|---|---|---|---|---|---|---|---|
| 5 | 18 | 0.24 | 0.25 | 0.44 | 0.68 | 0.78 | 0.86 | 0.87 |
| 8 | 20 | 0.13 | 0.24 | 0.36 | 0.42 | 0.50 | 0.64 | 0.65 |
| 11 | 26 | 0.01 | 0.10 | 0.28 | 0.42 | 0.51 | 0.60 | 0.62 |

*Table 46: Distribution of Item Biserial Correlations for Field-Test Items, Science*

| Grade | Total FT Items | Min | 5th Percentile | 25th Percentile | 50th Percentile | 75th Percentile | 95th Percentile | Max |
|---|---|---|---|---|---|---|---|---|
| 5 | 18 | 0.23 | 0.24 | 0.39 | 0.49 | 0.55 | 0.63 | 0.66 |
| 8 | 20 | -0.03 | 0.14 | 0.26 | 0.38 | 0.58 | 0.68 | 0.69 |
| 11 | 26 | -0.04 | 0.23 | 0.36 | 0.44 | 0.55 | 0.61 | 0.66 |

Among the AIRCore science items that were field-tested in 2019, 10 items were flagged for item discrimination, 3 items were flagged for *p*-value, 34 items were flagged for response time, and 7 items were flagged for DIF according to the criteria outlined in the sections above. Some items were flagged for multiple reasons and were therefore reviewed by educators during the process of data review. The classical statistics of all operational and field-tested AIRCore science items presented in Appendix K. The statistics were computed separately for each state where the item was administered. The average value of the statistics across states is also presented. Codebooks are provided to help navigating the view of a specific subset of items (e.g., checking the statistics of the AIRCore field-test item calculated using NH data only by applying relevant filters).

# 5. Item Calibration and Equating

## 5.1. ELA and Mathematics Item Calibration and Equating

### 5.1.1. Item Calibration

For English language arts (ELA) and mathematics, AIRCore is a pre-equated item bank with item parameters estimated under the multigroup item response theory (MGIRT) framework

Item response theory (IRT; van der Linden & Hambleton, 1997) was used to calibrate all items and derive scores for all AIRCore items used for the New Hampshire Statewide Assessment System. IRT is a general framework that models test responses resulting from an interaction between students and test items.

IRT encompasses many related measurement models that allow for varied assumptions about the nature of the data. Simple unidimensional models are the most common models used in K–12 operational testing programs, and items are often calibrated using a sample of students from within a state population. AIRCore items were administered across samples of students in different states. This grouping structure leads to a natural extension of the basic IRT models to data collected from multiple populations; hence, the MGIRT model (Bock & Zimowski, 1997) is used to calibrate all AIRCore items.

All individuals in the calibration sample are considered to have the observed responses $z_{jik_j}$, corresponding to test taker $j$, in group $k$ to the $i$th item. The MGIRT assumes local (conditional) independence of item responses and further assumes that the $j$th individual is a member of the $k_j$th population with density function $f(\theta; \mu_{k_j}, \sigma_{k_j}^2)$.

The generalized approach to item calibration begins with familiar probability models, including the three-parameter logistic model (3PL; Lord & Novick, 1968) for binary items and the Generalized Partial Credit Model (GPCM; Muraki, 1992) for items scored in multiple categories.

The probability model for binary items is denoted as

$$\mathrm{P_{ij}}\left(z_{ijk_j} = 1 | \theta_{jk_j}\right) = c_i + \frac{1 - c_i}{1 + \exp\left[-Da_i\left(\theta_{jk_j} - b_i\right)\right]},$$

where $\mathrm{P_{ij}}\left(z_{ijk_j} = 1|\theta_{jk_j}\right)$ is the probability of test taker $j$ answering item $i$ correct, $c_i$ is the lower asymptote of the item response curve (the pseudo-guessing parameter), $b_i$ is the location parameter, $a_i$ is the slope parameter (the discrimination parameter), and $D$ is a constant fixed at 1.7, bringing the logistic into coincidence with the probit model. Student ability is represented by $\theta_{jk_j}$.

The GPCM is typically expressed as the probability for individual $j$ of scoring in the $(z_{ijk_j} + 1)$th category to the $i$th item as

$$\mathrm{P_{ij}}\left(z_{ijk_j}\middle|\theta_{\mathrm{jk_j}}\right) = \frac{\exp\sum_{k=1}^{z_{ijk_j}} Da_i\left(\theta_{jk_j} - b_{ki}\right)}{1 + \sum_{h=1}^{m_i}\exp\sum_{k=0}^{h} Da_i\left(\theta_{jk_j} - b_{ki}\right)},$$

where $b_{ki}$ is the $k$th step value, $z_{ijk_j} = \{0,1,..,m_i\}$, $m_i$ is the maximum possible score of the item.

The conditional independence assumption then provides for the likelihood of the individual response pattern to be expressed as

$$\Pr\left(\mathbf{z}_{jk_j}\middle|\theta_j, \boldsymbol{\gamma}\right) = \prod_{i=1}^{N} Pr\left(z_{ijk_j}|\theta_{jk_j}, \boldsymbol{\gamma}\right)$$

where $\boldsymbol{\gamma}$ is a vector of item parameters, leading to the marginal likelihood of the responses within group $k$ as

$$L_j(\boldsymbol{\gamma}) = \int \prod_{i=1}^{I} Pr\left(z_{ij\boldsymbol{\gamma}}|\theta_{j\boldsymbol{\gamma}}, \boldsymbol{\gamma}\right) f\left(\theta|\mu_{k_j}, \sigma_{k_j}^2\right) d\theta$$

Then, assuming independence between different groups, the overall likelihood to be maximized with respect to the item parameters is

$$\arg\max L(\boldsymbol{\gamma}) = \prod_{j=1}^{N} L_{jk}(\boldsymbol{\gamma})$$

All item parameter estimates were obtained with IRTPRO version 4.1 (Cai, Thissen, & du Toit, 2011). IRTPRO uses the marginal maximum likelihood estimation. Identification of the model requires fixing the population parameters for one group to $N(0,1)$ and then the means of all other groups are freely estimated relative to the reference group. Each group's means and standard deviations are reported in Appendix B.

### 5.1.2. Equating to the Scale for ELA and Mathematics

Equating to the established reporting scale is done using the Stocking-Lord procedure (Stocking & Lord, 1983). The methods are implemented by calibrating the item response data using the same MGIRT model as described above and then using the methods described in this section to equate them to the AIRCore bank. Without loss of generality, the subscript notation is simplified here as the grouping structure for the MGIRT is unused for establishes linkages between tests.

First, define the probability of response for the class of binary IRT models on the *bank* scale, that is, the scale we are linking items to and let the subscripts $I$ and $J$ denote the item parameters for the bank and items to be rescaled, respectively:

$$p(z_{i,I} = 1|\theta) = c_{i,I} + \frac{1 - c_{i,I}}{1 + \exp[-Da_{i,I}(\theta - b_{i,I})]}$$

and for the polytomous IRT models:

$$p(z_{i,I}|\theta) = \frac{\exp(\sum_{k=0}^{z_i} Da_i(\theta - b_{ki,I}))}{\sum_{j=0}^{m_i} \exp \sum_{k=0}^{j} Da_{i,I}(\theta - b_{ki,I})}$$

where $z_i$ denotes score point $\mathrm{z}_i = \{1, \dots, \mathrm{m}_i\}$ to item $i$. The expected score for the polytomous models is:

$$E(z_{i,I}|\theta) = \sum_{z_{i,I}=1}^{m_i} z_{i,I} p(z_{i,I}|\theta)$$

The form of the IRT models for the new items that are to be linked onto the bank scale, or the *rescaled* items, have a similar form, but the transformation coefficients $A$ and $B$ are introduced as:

$$p(z_{i,I}^* = 1|\theta) = c_{i,J} + \frac{1 - c_{i,J}}{1 + \exp[-D\frac{a_{i,J}}{A}(\theta - (b_{i,J} * A + B))]}$$

and

$$p(z_{i,I}^*|\theta) = \frac{\exp(\sum_{k=0}^{z_i} D\frac{a_{i,J}}{A}(\theta - (b_{ki,J} * A + B)))}{\sum_{j=0}^{m_i} \exp \sum_{k=0}^{j} D\frac{a_{i,J}}{A}(\theta - (b_{ki,J} * A + B))}$$

The "*" is used when transformation coefficients appear in the IRT model. The notation $p(z_{i,J}|\theta)$ denotes the same IRT model, but without the transformation coefficients $A$ and $B$.

The symmetric approach uses the reverse transform for the bank items:

$$p(z_{i,I}^* = 1|\theta) = c_{i,I} + \frac{1 - c_{i,I}}{1 + \exp[-DAa_{i,I}\left(\theta - \frac{(b_{i,I} - B)}{A}\right)]}$$

and for the polytomous IRT models:

$$p(z_{i,I}^*|\theta) = \frac{\exp\left(\sum_{k=0}^{z_i} DAa_i\left(\theta - \frac{(b_{ki,I} - B)}{A}\right)\right)}{\sum_{j=0}^{m_i} \exp \sum_{k=0}^{j} DAa_{i,I}\left(\theta - \frac{(b_{ki,I} - B)}{A}\right)}$$

And then the objective function to be minimized with respect to the transformation coefficients, $A$ and $B$, is

$$\arg\min SL = \int \left[\sum_{i=1}^{K} E\left(z_{i,I}\middle|\theta_1\right) - \sum_{i=1}^{K} E\left(z_{i,J}^{*}\middle|\theta_1\right)\right]^2 f(\theta_1|\mu_1,\sigma_1^2)\, d\theta_1 + \int \left[\sum_{i=1}^{K} E\left(z_{i,I}^{*}\middle|\theta_2\right) - \sum_{i=1}^{K} E\left(z_{i,J}\middle|\theta_2\right)\right]^2 f(\theta_2|\mu_2,\sigma_2^2)\, d\theta_2$$

where $f(\theta_1|\mu_1,\sigma_1^2)$ is the normal population density associated with putting operational items onto the bank scale, and $f(\theta_2|\mu_2,\sigma_2^2)$ is the density associated with putting bank items onto the operational scale. Implementation is performed using Gauss-Hermite quadrature, and the integral is replaced with summation over $q$ quadrature points

$$\arg\min SL = \sum_{q_1=1}^{Q_1} \left[\sum_{i=1}^{K} E\left(z_{i,I}\middle|\theta_{1,q_1}\right) - \sum_{i=1}^{K} E\left(z_{i,J}^{*}\middle|\theta_{1,q_1}\right)\right]^2 w_{q_1} + \sum_{q_2=1}^{Q_2} \left[\sum_{i=1}^{K} E\left(z_{i,I}^{*}\middle|\theta_{2,q_2}\right) - \sum_{i=1}^{K} E\left(z_{i,I}\middle|\theta_{2,q_2}\right)\right]^2 w_{q_2}$$

where $\theta_{1,q_1}$ is node $q_1$ associated with $\theta_1$, $w_{q_1}$ is the weight at node $q_1$, $\theta_{2,q_2}$ is node $q_2$ associated with $\theta_2$, and $w_{q_2}$ is the weight at node $q_2$.

### 5.1.3. Establishing the Initial AIRCore Bank

This section describes the process of establishing the initial set of item parameters and equating the items over the years they were used. The AIRCore item bank currently spans three different years (2015–2017) of field testing. Initially, every grade was calibrated separately within a given year using MGIRT. For example, grade 5 mathematics items in 2015 and those in 2016 were calibrated separately. These year-over-year separate item calibrations were then equated using the Stocking-Lord procedure (Stocking & Lord, 1983) to place all AIRCore items from the separate calibrations onto a single scale.

This equating chain was established using a common-item non-equivalent groups design where a set of common items were administered in the pools each year. All common items in the pool were used unless the item's *A* parameter is less than 0.1 or greater than 3, and the absolute *B* parameter is greater than 6.

Table 47 below displays year-to-year equating constants.

*Table 47: Linking Across Years Results, ELA and Mathematics*

| Subject | Grade | 2016 to 2015 | | | 2017 to 2016 | | |
|---|---|---|---|---|---|---|---|
| | | Number of Anchor Items | Slope | Intercept | Number of Anchor | Slope | Intercept |
| ELA | 3 | 113 | 0.9413 | 0.0085 | 138 | 0.9749 | 0.1082 |
| | 4 | 128 | 0.8711 | 0.0091 | 185 | 0.9531 | 0.1451 |
| | 5 | 125 | 1.0497 | -0.0374 | 172 | 1.0340 | 0.0708 |
| | 6 | 173 | 1.0635 | 0.0953 | 184 | 0.9756 | 0.0750 |
| | 7 | 163 | 1.1462 | -0.0069 | 178 | 1.0259 | 0.1838 |
| | 8 | 135 | 0.9785 | -0.1097 | 155 | 1.0279 | -0.1285 |
| Math | 3 | 101 | 0.9765 | 0.0563 | 255 | 0.9444 | 0.0570 |
| | 4 | 96 | 1.0017 | 0.0011 | 229 | 1.0287 | 0.0394 |
| | 5 | 218 | 1.0586 | 0.0284 | 271 | 1.0392 | 0.0682 |
| | 6 | 194 | 1.0266 | 0.0949 | 228 | 1.0530 | 0.0961 |
| | 7 | 178 | 1.0682 | -0.0574 | 259 | 1.0901 | -0.0606 |
| | 8 | 194 | 1.1290 | -0.1380 | 269 | 1.0763 | -0.0296 |

### 5.1.4. Linking the Initial AIRCore Bank to SAGE Bank

The methods above are used to calibrate and equate the AIRCore bank. Once that bank was established, these items were then linked to the Utah Student Assessment of Growth and Excellence (SAGE) item bank, which provides a vertical reporting scale. Linking the AIRCore bank and SAGE bank also used the Stocking-Lord procedure (Stocking & Lord, 1983) using the same common-item non-equivalent groups design. Table 48 shows linking constants for each grade and subject between the initial AIRCore bank and SAGE. These linking constants were used to put the initial AIRCore bank into the SAGE on-grade level scale.

Appendix C documents the design and results of the vertical linking study that was implemented to develop the SAGE ELA and mathematics item bank.

*Table 48: Linking to SAGE Results, ELA and Mathematics*

| Subject | Grade | Number of Anchor Items | Slope | Intercept |
|---|---|---|---|---|
| ELA | 3 | 177 | 1.0026 | 0.0729 |
| | 4 | 227 | 1.0267 | -0.0131 |
| | 5 | 182 | 0.9873 | 0.0860 |
| | 6 | 244 | 1.0085 | 0.0228 |
| | 7 | 159 | 1.0189 | -0.0243 |
| | 8 | 160 | 0.9983 | 0.1773 |

| Subject | Grade | Number of Anchor Items | Slope | Intercept |
|---|---|---|---|---|
| Mathematics | 3 | 295 | 1.1081 | 0.1386 |
| | 4 | 276 | 1.0609 | 0.0979 |
| | 5 | 247 | 1.0406 | 0.1034 |
| | 6 | 211 | 1.0056 | 0.0525 |
| | 7 | 217 | 1.0125 | 0.1035 |
| | 8 | 252 | 0.9671 | 0.2525 |

Table 49 and Table 50 display the number of students in each participating state contributing to the AIRCore multigroup IRT model.

*Table 49: Number of Students Used in AIRCore MGIRT Calibration, ELA*

| Grade | Year | Utah | Florida | Arizona | Oregon (2015) / Ohio (2016) |
|---|---|---|---|---|---|
| 3 | 2015 | 39,279 | - | 33,687 | 9,323 |
| | 2016 | 46,901 | - | 62,242 | 85,972 |
| | 2017 | 47,317 | - | 72,754 | - |
| 4 | 2015 | 39,753 | - | 33,091 | 11,858 |
| | 2016 | 43,190 | 207,867 | 61,065 | 95,211 |
| | 2017 | 45,537 | 206,341 | 73,195 | - |
| 5 | 2015 | 38,976 | 35,780 | 32,398 | 8,398 |
| | 2016 | 36,196 | 199,326 | 60,210 | 97,451 |
| | 2017 | 43,825 | 209,984 | 72,289 | - |
| 6 | 2015 | 38,340 | 42,565 | 33,114 | 8,234 |
| | 2016 | 38,106 | 196,409 | 57,635 | 101,799 |
| | 2017 | 39,662 | 200,039 | 69,837 | - |
| 7 | 2015 | 36,082 | 56,752 | 30,911 | 10,688 |
| | 2016 | 45,469 | 193,186 | 58,050 | 105,249 |
| | 2017 | 45,484 | 197,752 | 69,754 | - |
| 8 | 2015 | 36,445 | 82,159 | 32,277 | 13,590 |
| | 2016 | 42,530 | 195,125 | 57,349 | 104,360 |
| | 2017 | 42,018 | 197,269 | 69,481 | - |

*Table 50: Number of Students Used in AIRCore MGIRT Calibration, Mathematics*

| Grade | Year | Utah | Florida | Arizona | Oregon (2015) / Ohio (2016) |
|---|---|---|---|---|---|
| 3 | 2015 | 48,473 | - | 43,543 | 27,642 |
| | 2016 | 49,762 | - | 62,586 | 94,869 |
| | 2017 | 49,688 | 185,609 | 72,857 | - |
| 4 | 2015 | 47,088 | - | 43,464 | 27,102 |
| | 2016 | 48,367 | - | 61,384 | 95,765 |
| | 2017 | 49,727 | 173,825 | 73,438 | - |
| 5 | 2015 | 47,098 | 87,436 | 42,419 | 26,957 |
| | 2016 | 46,702 | 201,278 | 60,448 | 97,308 |
| | 2017 | 48,021 | 212,008 | 72,428 | - |
| 6 | 2015 | 46,160 | 87,831 | 40,512 | 27,550 |
| | 2016 | 46,380 | 193,158 | 57,868 | 101,015 |
| | 2017 | 46,263 | 195,425 | 70,034 | - |
| 7 | 2015 | 43,517 | 79,949 | 39,887 | 26,753 |
| | 2016 | 43,718 | 170,453 | 57,467 | 102,933 |
| | 2017 | 43,623 | 171,940 | 68,366 | - |
| 8 | 2015 | 43,745 | 60,958 | 39,997 | 26,969 |
| | 2016 | 43,377 | 125,120 | 49,781 | 78,629 |
| | 2017 | 44,035 | 120,321 | 59,171 | - |
| | 2016 | 28,212 | 137,337 | 39,249 | - |
| | 2017 | 9,763 | 120,631 | 50,063 | - |

## 5.2. ITEM CALIBRATION AND LINKING FOR SCIENCE

### 5.2.1. Model Description

In discussing IRT models for the New Hampshire science assessment, 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 taken into account.

**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:

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

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's $j$ 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 AIRCore science bank 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 science items developed under the Next Generation Science Standards (NGSS) framework, multiple assertions are typically developed around a single item so that assertions are clustered within items.

One approach would be 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 there are local dependencies 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. 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 to an underestimation of the standard error of measurement (e.g., Sireci, Wainer, & Thissen, 1991; Yen, 1993).

Many current ELA assessments also contain groups of items that pertain to the same stimulus. For example, several items may share the same reading passage. Currently, item clustering effects and the resulting conditional dependencies are typically ignored, an approach that seems to work reasonably well in practice. This may be because, in ELA assessments, the individual items within a group of items pertaining to the same passage are often written so that the effects of sharing the same stimulus material are kept to a minimum, such as by relating items to different parts of the reading passage. However, for the NGSS science items, the conditional dependencies between the assertions of an item (and item cluster) are too substantial to be ignored, because those assertions are more intrinsically related to one another. For example, the assertions within an item are organized around a single performance expectation.

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

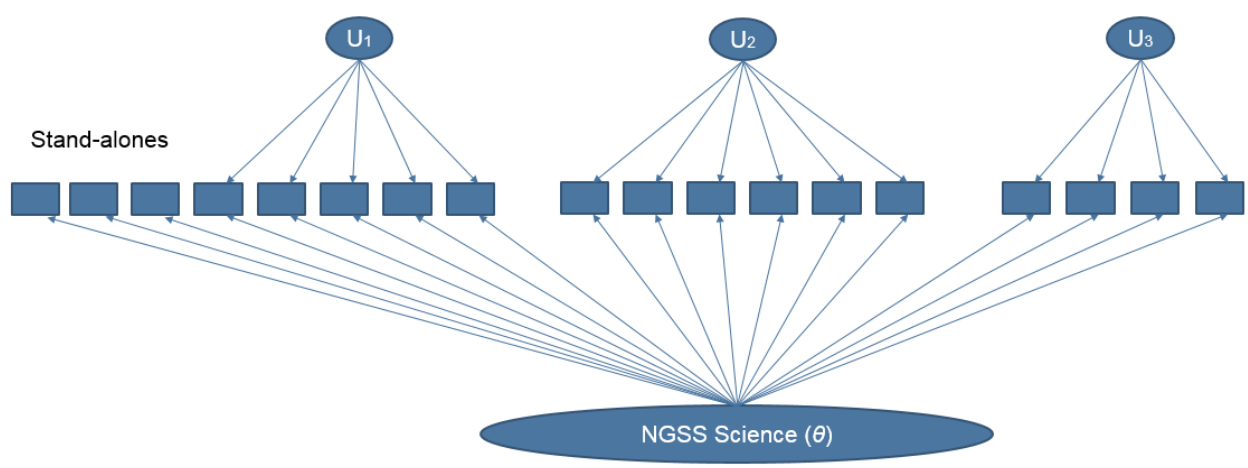
models assume that all assertions (the basic units of analysis) are independent given a single underlying trait $\theta$, we now assume conditional independence of assertions, given the underlying latent trait $\theta$ and all nuisance dimensions:

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

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 $\mathbf{u}$ 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 note 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 not necessary 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 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 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*




## 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:

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

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 the item response function. 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.

**Multigroup Model**

The item bank for science was calibrated concurrently using all the items administered in any of the states that collaborate with the American Institutes for Research (AIR) on their new science assessments. In the calibration, each state was treated as a population of students, or as a 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 $k$, $k = 1, \ldots, K$, 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 0 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:

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

### 5.2.2. Item Calibration

**Estimation**

A separate IRT model was fit for each grade band. The parameters of each 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 $u_j$ for each student $j$ are treated as random effects and integrated out to obtain the marginal log likelihood corresponding to the observed response pattern $\mathbf{z}_j$ for student $j$,

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

where $\boldsymbol{\Sigma}$ is a diagonal matrix with diagonal elements $\sigma_{u_g}^2$. Across all students and groups, the overall log likelihood to be maximized with respect to the vector $\gamma$ 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 spaces $(\theta, u_{\mathrm{g}})$, (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 2018 parameters since NH reports scores before the testing window closes (immediate score reporting). The 2019 parameters will 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 below for both years.

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®, in order to speed up 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 next 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 parameters estimates were very similar across software packages.

*Table 51: Groups Per Grade for the 2018 Science Calibration*

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

*Note:* Shaded rows represent states that were part of the core calibration.

### 2018 Calibration Sequence

Table 51 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 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, in order to divide the large set of items (and assertions) into more manageable pieces, a separate calibration was carried out for two states with a 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
    a. All the item responses of New Hampshire and West Virginia. These states administered items from (see bank sharing matrix Table 52). A more detailed overlap of the common items at the time of the 2018 calibration was given in Section 0.1 (see Table 29 through Table 31):
        i. AIRCore
        ii. Connecticut
        iii. Hawaii
        iv. Rhode Island
        v. Vermont
        vi. Utah
        vii. West Virginia
    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
        i. AIRCore
        ii. Connecticut
        iii. Hawaii
        iv. Rhode Island
        v. Vermont
        vi. Utah
        vii. West Virginia
        viii. Wyoming (items were treated as not administered, and responses are replaced by missing code)
        ix. Oregon (items were treated as not administered, and responses are replaced by missing code)
    c. Item responses from Hawaii to items also administered in another state (Hawaii items were used in Connecticut, Hawaii, Rhode Island, Vermont, and West Virginia).

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

*Table 52: Science State Sharing Matrix*

| Source Bank and State Owned | CT | HI | MSSA | NH (from ITS Sandbox) | OR | UT | WV | WY |
|---|---|---|---|---|---|---|---|---|
| **AIRCore** | X | X | X | X | X | | X | X |
| **Connecticut** | X | | X | | | | X | |
| **Hawaii** | X | X | X | | | | X | |
| **Oregon** | X | | X | | X | | | |
| **MSSA** | 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 Utah and Hawaii had a substantial proportion of items that were administered only in Utah and Hawaii, respectively. Hawaii has both Hawaii and AIRCore items in common with the states of the core calibration (Hawaii administered only Hawaii and AIRCore items); Utah has only Utah items in common (Utah administered only Utah items). The parameters for the unique Hawaii items depend only on responses from students from Hawaii, and the parameters for the unique Utah items depend only on responses from students from Utah. For both states, the state-specific items were calibrated separately 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.

**2019 Calibration Sequence**

The calibration was done 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 3 items, there were 1,500 or more student responses). In this step, only the data of states with an operational test were included. Table 53 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 RI allowed students to skip items. There were 9 items administered as operational items in 2019 for which the sample size was smaller than 1,000, out of a total of 438 items. Table 54 through Table 56 present the number of operational clusters and stand-alone items that were shared between the item pools of any two states. The numbers below the diagonal represent the number of common items at the time of the 2019 calibration. The shaded diagonal elements represent the number of unique items at the time of calibration. Table 54 presents the results for elementary schools, Table 55 the results for middle schools, and Table 56 the results for high schools. The numbers at operational administration are slightly different from the numbers at calibration because items with sample size smaller than 1000 were excluded from the calibration.

*Table 53: 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 54: Number of Common Operational Elementary School Items Administered in Spring 2019, Science*

| | State | Connecticut | MSSA (RI, VT) | New Hampshire | Oregon | West Virginia |
|---|---|---|---|---|---|---|
| Cluster | **CT** | **1 (1)** | 44 | 24 | 42 | 55 |
| | **MSSA (RI, VT)** | 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)** |

| | State | Connecticut | MSSA (RI, VT) | New Hampshire | Oregon | West Virginia |
|---|---|---|---|---|---|---|
| Stand-Alone | CT | 3 (3) | 34 | 26 | 30 | 47 |
| | MSSA (RI, VT) | 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 (RI, VT) | 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 55: Number of Common Operational Middle School Items Administered in Spring 2019, Science*

| | State | Connecticut | MSSA (RI, VT) | New Hampshire | Oregon | West Virginia |
|---|---|---|---|---|---|---|
| Cluster | CT | 3 (3) | 26 | 24 | 54 | 92 |
| | MSSA (RI, VT) | 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 (RI, VT) | 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 (RI, VT) | 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 56: Number of Common Operational High School Items Administered in Spring 2019, Science*

| | State | Connecticut | MSSA (RI, VT) | New Hampshire | Oregon | West Virginia |
|---|---|---|---|---|---|---|
| Cluster | CT | 5 (5) | 33 | 22 | 30 | 0 |
| | MSSA (RI, VT) | 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 (RI, VT) | 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 (RI, VT) | 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 a 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 57 provides an overview of the groups per grade for calibration of the field test items.

*Table 57: 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 |

### 5.2.3. 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 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 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

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

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

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

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

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

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

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 for which from the 2019 assessments was lower than 1,000 (9 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 58. 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 (all states except for RI). So, out of an abundance of caution, these three items were not anchored to their 2018 parameter values.

*Table 58: 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 | 38549 | 0.0000 | 0.0234 | 39347 | 0.0000 | 0.1443 | 37616 |
| **NEW HAMPSHIRE** | 0.0631 | 0.1083 | 13187 | 0.0940 | 0.1108 | 12060 | 0.0798 | 0.2278 | 11385 |
| **OREGON** | -0.0101 | 0.0096 | 44989 | 0.0028 | 0.0156 | 42043 | -0.0383 | 0.1030 | 41630 |
| **RHODE ISLAND** | -0.0312 | 0.0142 | 10751 | -0.1044 | -0.0692 | 10306 | -0.2261 | -0.0879 | 9612 |
| **VERMONT** | 0.1069 | 0.1504 | 6017 | 0.0781 | 0.1133 | 5894 | 0.0179 | 0.1545 | 5332 |
| **WEST VIRGINIA** | -0.1970 | -0.1529 | 19540 | -0.3012 | -0.2783 | 19043 | - | - | - |
| | $\frac{1}{K}\sum_{k=1}^{K}\hat{\mu}_{k\,2019}$ | $\frac{1}{K}\sum_{k=1}^{K}\hat{\mu}_{k\,2018}$ | $\hat{B}$ | $\frac{1}{K}\sum_{k=1}^{K}\hat{\mu}_{k\,2019}$ | $\frac{1}{K}\sum_{k=1}^{K}\hat{\mu}_{k\,2018}$ | $\hat{B}$ | $\frac{1}{K}\sum_{k=1}^{K}\hat{\mu}_{k\,2019}$ | $\frac{1}{K}\sum_{k=1}^{K}\hat{\mu}_{k\,2018}$ | $\hat{B}$ |
| **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), as well as the estimated group means and variances, are presented in Appendix L. 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 L display the histogram of the difficulty parameters for elementary and middle school for all items that are part of the New Hampshire operational pool. The figures also display the proficiency distributions. The distribution of the difficulty parameter overlaps well with the proficiency distribution in grade 8. The grade 5 items are slightly easier than the student proficiency, while the grade 11 items are slightly more difficult than the student proficiency in general.

# 6. SCORING

## 6.1. MAXIMUM LIKELIHOOD ESTIMATION FOR ELA AND MATHEMATICS

Ability estimates were generated using *pattern scoring*, a method that scores students depending on how they answer individual items. Scoring details are provided below.

### 6.1.1. Likelihood Function

The likelihood function for generating the maximum likelihood estimates (MLEs) is based on a mixture of item types and can therefore be expressed as

$$L(\theta) = L(\theta)^{MC} L(\theta)^{CR},$$

where

$$L(\theta)^{MC} = \prod_{i=1}^{N_{MC}} P_i^{z_i} Q_i^{1-z_i}$$

$$L(\theta)^{CR} = \prod_{i=1}^{N_{CR}} \frac{exp \sum_{k=0}^{z_i} Da_i(\theta - \delta_{ki})}{\sum_{h=0}^{m_i} exp \sum_{k=0}^{h} Da_i(\theta - \delta_{ki})}$$

$$P_i = c_i + \frac{1 - c_i}{1 + exp[-Da_i(\theta - b_i)]}$$

$$Q_i = 1 - P_i,$$

where $c_i$ is the lower asymptote of the item response curve (i.e., the pseudo-guessing parameter), $a_i$ is the slope of the item response curve (i.e., the discrimination parameter), $b_i$ is the location parameter, $z_i$ is the observed response to the item, $i$ indexes item, $h$ indexes step of the item, $m_i$ is the maximum possible score point (starting from 0), $\delta_{ki}$ is the kth step for item $i$ with m total categories, and $D = 1.7$.

A student's theta (i.e., MLE) is defined as $\arg\max_{\theta} log\big(L(\theta)\big)$ given the set of items administered to the student.

### 6.1.2. Derivatives

Finding the maximum of the likelihood requires an iterative method, such as Newton-Raphson iterations. The estimated MLE is found via the following maximization routine:

$$\theta_{t+1} = \theta_t - \frac{\partial \ln L(\theta_t)}{\partial \theta_t} \Big/ \frac{\partial^2 \ln L(\theta_t)}{\partial^2 \theta_t},$$

where

$$\frac{\partial \ln L(\theta)}{\partial \theta} = \frac{\partial \ln L(\theta)^{3PL}}{\partial \theta} + \frac{\partial \ln L(\theta)^{CR}}{\partial \theta}$$

$$\frac{\partial^2 \ln L(\theta)}{\partial^2 \theta} = \frac{\partial^2 \ln L(\theta)^{3PL}}{\partial^2 \theta} + \frac{\partial^2 \ln L(\theta)^{CR}}{\partial^2 \theta}$$

$$\frac{\partial \ln L(\theta)^{3PL}}{\partial \theta} = \sum_{i=1}^{N_{3PL}} Da_i \frac{(P_i - c_i)Q_i}{1 - c_i} \left( \frac{z_i}{P_i} - \frac{1 - z_i}{Q_i} \right)$$

$$\frac{\partial^2 \ln L(\theta)^{3PL}}{\partial^2 \theta} = -\sum_{i=1}^{N_{3PL}} D^2 a_i^2 \frac{(P_i - c_i)Q_i}{(1-c_i)^2}\left(1 - \frac{z_i c_i}{P_i^2}\right)$$

$$\frac{\partial \ln L(\theta)^{CR}}{\partial \theta} = \sum_{i=1}^{N_{CR}} Da_i \left( exp\left( \sum_{k=1}^{z_i} Da_i(\theta - \delta_{ki}) \right) \right) \left( \frac{z_i}{1 + \sum_{j=1}^{m_i} exp\left(\sum_{k=1}^{j} Da_i(\theta - \delta_{ki})\right)} - \frac{\sum_{j=1}^{m_i} j exp\left(\sum_{k=1}^{j} Da_i(\theta - \delta_{ki})\right)}{\left(1 + \sum_{j=1}^{m_i} exp\left(\sum_{k=1}^{j} Da_i(\theta - \delta_{ki})\right)\right)^2} \right)$$

$$\frac{\partial^2 \ln L(\theta)^{CR}}{\partial^2 \theta} = \sum_{i=1}^{N_{CR}} D^2 a_i^2 \left( \left( \frac{\sum_{j=1}^{m_i} j exp\left(\sum_{k=1}^{j} Da_i(\theta - \delta_{ki})\right)}{1 + \sum_{j=1}^{m_i} exp\left(\sum_{k=1}^{j} Da_i(\theta - \delta_{ki})\right)} \right)^2 - \frac{\sum_{j=1}^{m_i} j^2 exp\left(\sum_{k=1}^{j} Da_i(\theta - \delta_{ki})\right)}{1 + \sum_{j=1}^{m_i} exp\left(\sum_{k=1}^{j} Da_i(\theta - \delta_{ki})\right)} \right)$$

and where $\theta_t$ denotes the estimated $\theta$ at iteration *t*. $N_{CR}$ is the number of items that are scored using the generalized partial credit model (GPCM) model, and $N_{3PL}$ is the number of items scored using three-parameter logistic (3PL) or two-parameter logistic (2PL) model.

### 6.1.3. Standard Errors of Estimate

When the MLE is available, the standard error (SE) of the MLE is estimated by

$$se(\hat{\theta}) = \frac{1}{\sqrt{-\left(\frac{\partial^2 \ln L(\hat{\theta})}{\partial^2 \theta}\right)}},$$

where

$$\frac{\partial^2 \ln L(\hat{\theta})}{\partial^2 \theta} = \sum_{i=1}^{N_{CR}} D^2 a_i^2 \left( \left( \frac{\sum_{j=1}^{m_i} j Exp\left(\sum_{k=1}^{j} Da_i(\hat{\theta} - b_{ik})\right)}{1 + \sum_{j=1}^{m_i} Exp\left(\sum_{k=1}^{j} Da_i(\hat{\theta} - b_{ik})\right)} \right)^2 - \frac{\sum_{j=1}^{m_i} j^2 Exp\left(\sum_{k=1}^{j} Da_i(\hat{\theta} - b_{ik})\right)}{1 + \sum_{j=1}^{m_i} Exp\left(\sum_{k=1}^{j} Da_i(\hat{\theta} - b_{ik})\right)} \right) - \sum_{i=1}^{N_{3PL}} D^2 a_i^2 \frac{(P_i - c_i)Q_i}{(1-c_i)^2}\left(1 - \frac{z_i c_i}{P_i^2}\right)$$

where $N_{CR}$ is the number of items that are scored using the GPCM model, and $N_{3PL}$ is the number of items scored using 3PL or 2PL model.

### 6.1.4. Extreme Case Handling

When students answer all items correctly or all items incorrectly, the likelihood function is unbounded, and an MLE cannot be generated. In addition, when a student's raw score is lower

than the expected raw score due to guessing, the likelihood is not identified. For New Hampshire Statewide Assessment System (NH SAS) scoring, the extreme cases were handled as follows:

i. Assign the Lowest Obtainable Theta (LOT) value of –4 to a raw score of 0.
ii. Assign the Highest Obtainable Theta (HOT) value of 4 to a perfect score.
iii. Generate MLE for every other case and apply the following rule:
    a. If MLE is lower than –4, assign theta to –4.
    b. If MLE is higher than 4, assign theta to 4.

As NH SAS used a vertical score for scoring, the truncated LOT and HOT were converted to the vertical scale before being applied. These truncated LOT and HOT in vertical scale and the associated scale scores for each grade and subject are provided in Table 59 and Table 60.

*Table 59: ELA Theta and Corresponding Scaled-Score Limits for Extreme Ability Estimates*

| Grade | Lowest Obtainable Theta (LOT) | Highest Obtainable Theta (HOT) | Lowest Obtainable Scale Score (LOSS) | Highest Obtainable Scale Score (HOSS) |
|---|---|---|---|---|
| 3 | -4.61 | 2.03 | 420 | 750 |
| 4 | -4.39 | 2.73 | 430 | 790 |
| 5 | -4.01 | 3.11 | 450 | 810 |
| 6 | -3.72 | 3.48 | 460 | 830 |
| 7 | -3.75 | 3.77 | 470 | 850 |
| 8 | -3.84 | 4.24 | 480 | 870 |

*Table 60: Mathematics Theta and Corresponding Scaled Score Limits for Extreme Ability Estimates*

| Grade | Lowest Obtainable Theta (LOT) | Highest Obtainable Theta (HOT) | Lowest Obtainable Scale Score (LOSS) | Highest Obtainable Scale Score (HOSS) |
|---|---|---|---|---|
| 3 | -4.85 | -0.05 | 300 | 550 |
| 4 | -4.77 | 1.15 | 310 | 610 |
| 5 | -4.63 | 2.17 | 320 | 660 |
| 6 | -4.52 | 3.40 | 330 | 720 |
| 7 | -4.05 | 4.03 | 340 | 750 |
| 8 | -4.28 | 5.64 | 350 | 830 |

### 6.1.5. Standard Error of LOT/HOT Scores

When the MLE is available and within the LOT and HOT, the standard error (SE) is estimated based on Fisher information.

When the MLE is not available (such as for extreme score cases) or the MLE is censored to the LOT or HOT, the SE for student *s* is estimated by

$$se(\theta_s) = \frac{1}{\sqrt{I(\theta_s)}},$$

where $I(\theta_s)$ is the test information for student *s*. The NH SAS included items that were scored using the 3PL 2PL model, and the GPCM from the item response theory (IRT). The 2PL can be visualized as either a 3PL item with no pseudo-guessing parameter or a dichotomously scored GPCM item. The test information was calculated as

$$I(\theta_s) = \sum_{i=1}^{N_{CR}} D^2 a_i^2 \left( \frac{\sum_{j=1}^{m_i} j^2 Exp\left(\sum_{k=1}^{j} Da_i(\theta_s - b_{ik})\right)}{1 + \sum_{j=1}^{m_i} Exp\left(\sum_{k=1}^{j} Da_i(\theta_s - b_{ik})\right)} - \left( \frac{\sum_{j=1}^{m_i} jExp\left(\sum_{k=1}^{j} Da_i(\theta_s - b_{ik})\right)}{1 + \sum_{j=1}^{m_i} Exp\left(\sum_{k=1}^{j} Da_i(\theta_s - b_{ik})\right)} \right)^2 \right) + \sum_{i=1}^{N_{3PL}} D^2 a_i^2 \left( \frac{Q_i}{P_i} \left[ \frac{P_i - c_i}{1 - c_i} \right]^2 \right),$$

where $N_{CR}$ is the number of items that are scored using the GPCM model, and $N_{3PL}$ is the number of items scored using the 3PL or 2PL model.

For SE of LOT/HOT scores, theta in the formula above is replaced with the LOT/HOT values. The upper bound of the SE was set to 1.5 and converted to the vertical scale. Any value larger than 1.5 was truncated at 1.5. The truncated standard error of measurement (SEM) values on the vertical scale are provided in Table 61.

*Table 61: SEM Truncation Values for Each Grade, ELA and Mathematics*

| Subject | Grade | SEM Truncation Values on Theta Metric | SEM Truncation Values on Vertical Scale |
|---|---|---|---|
| ELA | 3 | 1.5 | 1.25 |
| | 4 | 1.5 | 1.34 |
| | 5 | 1.5 | 1.34 |
| | 6 | 1.5 | 1.35 |
| | 7 | 1.5 | 1.41 |
| | 8 | 1.5 | 1.52 |
| Mathematics | 3 | 1.5 | 0.90 |
| | 4 | 1.5 | 1.11 |
| | 5 | 1.5 | 1.28 |
| | 6 | 1.5 | 1.49 |

| Subject | Grade | SEM Truncation Values on Theta Metric | SEM Truncation Values on Vertical Scale |
|---|---|---|---|
| | 7 | 1.5 | 1.52 |
| | 8 | 1.5 | 1.86 |

### 6.1.6. Transforming Vertical Scale Scores to Reporting Scale Scores

NH SAS scale scores are reported for each student who takes the ELA, mathematics, or science assessments. Scale scores are based on the operational items presented to the student and do not include any field-test items or linking items. AIRCore item parameters are converted to a vertical scale in the item bank, and a single scale across all grades is used within ELA and mathematics. The reporting scale scores are calculated as

$$SS = slope * \theta_{vertical} + intercept,$$

where $slope$ and $intercept$ are the reporting scaling constants, and $\theta_{vertical}$ is the post-vertically scaled IRT ability estimate. For ELA, the slope and intercept are fixed at 50 and 650, and for mathematics, at 50 and 550, respectively. In this transformation, the following rules are applied:

1. The same linear transformation is used for all students within a grade.
2. Scale scores are rounded to the nearest integer (e.g., 302.4 to 302; 302.5 to 303).
3. A standard error is provided for each score, using the same set of items used to derive the score. The standard error of the scaled score is calculated as

$$se(SS) = se(\theta) * slope.$$

4. Truncated scale scores use actual SEs from the vertical scale theta estimates.

A summary of spring 2019 NH SAS scale scores using means, standard deviations, and percentages of students within each of the performance levels for each test is provided in Appendix D. The summary of scale scores for each reporting category is provided in Appendix E. All scores are based on the operational items presented to the student.

### 6.1.7. Overall Performance Classification

Each student is assigned an overall performance category according to his or her overall scale score. Table 62 and Table 63 provide the scale score range for performance standards for ELA and mathematics, respectively. The lower bound of the level 3, Proficient, marks the minimum cut score for proficiency.

*Table 62: Performance Levels for ELA by Grade*

| Grade | Level 1 Below Proficient | Level 2 Approaching Proficient | Level 3 Proficient | Level 4 Above Proficient |
|---|---|---|---|---|
| 3 | 420–556 | 557–586 | 587–615 | 616–750 |
| 4 | 430–579 | 580–604 | 605–634 | 635–790 |
| 5 | 450–593 | 594–620 | 621–663 | 664–810 |
| 6 | 460–604 | 605–641 | 642–687 | 688–830 |
| 7 | 470–607 | 608–643 | 644–696 | 697–850 |
| 8 | 480–624 | 625–660 | 661–710 | 711–870 |

*Table 63: Performance Levels for Mathematics by Grade*

| Grade | Level 1 Below Proficient | Level 2 Approaching Proficient | Level 3 Proficient | Level 4 Above Proficient |
|---|---|---|---|---|
| 3 | 300–409 | 410–430 | 431–454 | 455–550 |
| 4 | 310–430 | 431–459 | 460–491 | 492–610 |
| 5 | 320–459 | 460–494 | 495–521 | 522–660 |
| 6 | 330–478 | 479–517 | 518–555 | 556–720 |
| 7 | 340–506 | 507–551 | 552–586 | 587–750 |
| 8 | 350–538 | 539–590 | 591–624 | 625–830 |

### 6.1.8. Reporting Category Performance Classification

In addition to overall performance classification, subscale-level classification is computed to classify student performance levels for each of the content standard subscales. For each subscale, classification into one of three performance levels is determined by following the rules

- if $(\theta_{tt} < \theta_{Proficient} - 1.5 \times SE_{RC})$, then performance is classified as *Low*;
- if $(\theta_{Proficient} - 1.5 \times SE_{RC} \leq \theta_{tt} < \theta_{Proficient} + 1.5 \times SE_{RC})$, then performance is classified as *At or Approaching*; and
- if $(\theta_{tt} \geq \theta_{Proficient} + 1.5 * SE_{RC})$, then performance is classified as *On or Above*;

where $\theta_{Proficient}$ is the minimum proficiency cut score based on the overall test, $\theta_{tt}$ is the student's score on a given subscale, and $SE_{RC}$ is the standard error of the given subscale. Zero and perfect scores are assigned *Low* and *On or Above*, respectively.

### 6.1.9. Strengths and Weaknesses Scores

For individual students, strengths and weaknesses scores at reporting categories are computed relative to their individual overall estimated abilities.

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

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

where $E(z_{ij})$ is the expected score on item *i* for student *j* with estimated ability $\hat{\theta}_j$.

Residuals are summed for items within a reporting category. The sum of residuals is divided by the total number of points possible for items within the reporting category, *T*,

$$\delta_{jT} = \frac{\sum_{i \in T} \delta_{ji}}{\sum_{i \in T} m_i}.$$

For an aggregate unit, a target score for the reporting category is computed by averaging the target scores of individual students with different abilities who receive different items that measure the same reporting category at different levels of difficulty,

$$\bar{\delta}_{Tg} = \frac{1}{n_g} \sum_{j \in g} \delta_{jT}, \text{ an } se(\bar{\delta}_{Tg}) = \sqrt{\frac{1}{n_g(n_g-1)} \sum_{j \in g} \left(\delta_{jT} - \bar{\delta}_{Tg}\right)^2},$$

where $n_g$ is the number of students who respond to any of the items that belong to the reporting category *T* for an aggregate unit $g$. If a student did not happen to see any items on a particular reporting category, the student is not included in the $n_g$ count for the aggregate.

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

For reporting category level strengths/weakness, the following is reported:

- If $\bar{\delta}_{Tg} \leq -1 * se(\bar{\delta}_{Tg})$, then performance is worse than on the overall test.
- If $\bar{\delta}_{Tg} \geq +1 * se(\bar{\delta}_{Tg})$, then performance is better than on the overall test.
- Otherwise, performance is similar to performance on the overall test.
- If $se(\bar{\delta}_{Tg}) > 0.2$, data are insufficient.

## 6.2. Marginal Maximum Likelihood Estimation for Science

### 6.2.1. Marginal Likelihood Function

Student scores are obtained by marginalizing out the nuisance dimensions $\mathbf{u}_j$ from the likelihood of the observed response pattern $\mathbf{z}_j$ for student $j$,

$$\ell_i\left(\theta_j\right) = \log \int_{\mathbf{u}_j} P\left(\mathbf{z}_j \middle| \theta_j, \mathbf{u}_j\right) N\left(\mathbf{u}_j \middle| \mathbf{0}, \mathbf{\Sigma}\right) d\mathbf{u}_j ,$$

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 calibration 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 $\mathbf{v} = \Sigma^{-\frac{1}{2}}\mathbf{u}$. Then we have

$$\int_{\mathbf{u}_j} P\left(\mathbf{z}_j \middle| \theta_j, \mathbf{u}_j\right) N\left(\mathbf{u}_j \middle| \mathbf{0}, \mathbf{\Sigma}\right) d\mathbf{u}_j = \int_{\mathbf{v}_j} P\left(\mathbf{z}_j \middle| \theta_j, \mathbf{\Sigma}^{\frac{1}{2}}\mathbf{v}_j\right) N\left(\mathbf{v}_j \middle| \mathbf{0}, \mathbf{I}\right) d\mathbf{v}_j .$$

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

$$\int_{\mathbf{u}_j} P\left(\mathbf{z}_j \middle| \theta_j, \mathbf{u}_j\right) N\left(\mathbf{u}_j \middle| \mathbf{0}, \mathbf{\Sigma}\right) d\mathbf{u}_j = P\left(\mathbf{z}_j \middle| \theta_j\right),$$

which is the likelihood under the unidimensional Rasch model.

### 6.2.2. Derivatives

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

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

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

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

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

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

Based on the previous equations, we only need 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

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

and

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

Therefore, we have

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

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

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

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

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

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

### 6.2.3. Extreme Case Handling

Just like 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 64 contains the LOT and HOT values for each grade.

### 6.2.4. Standard Errors of Estimate

The SEM of the MMLE score estimate is

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

where $I\left(\hat{\theta}_{MMLE}\right)$ 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 each student answers and their estimate of θ. Different students will 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 errors are set at the truncation value for the standard errors.

### 6.2.5. Student-Level Scale Scores

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 above, but only with items within the given discipline. Scores are truncated on the theta scale at the LOT and HOT values specified in Table 64, which correspond to values of the estimated mean minus/plus four times the estimated standard deviation of $\theta$.

The reporting scales are linear transformations 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 64). The standard error of estimate for the estimated scale score is obtained as:

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

In 2018, the slope $a$ and intercept $b$ were chosen so that the center of the reporting scale of each grade (550, 850, and 1150, respectively) is at the grade mean of the 2018 base-year and has a standard deviation of 12.5. Furthermore, for each grade the reporting scale ranges from the base-year mean minus four times the standard deviation to the base-year mean plus four times the standard deviation. Specifically, for grade 5, the slope and intercept were obtained as:

$$\begin{aligned} SS &= 12.5\theta^* + 550 \\ &= 12.5\frac{\theta - \hat{\mu}_\theta}{\hat{\sigma}_\theta} + 550 \\ &= \frac{12.5}{\hat{\sigma}_\theta}\theta + \left(550 - \frac{12.5\hat{\mu}_\theta}{\hat{\sigma}_\theta}\right), \end{aligned}$$

where the second line stems from standardizing theta, $\theta^* = \frac{\theta - \hat{\mu}_\theta}{\hat{\sigma}_\theta}$. For grades 8 and 11, the slope and intercept can also be derived in a similar fashion.

Table 64 presents the intercept and slope for the three grades that are assessed, as well as the LOT, HOT, LOSS, and HOSS values. Table 64 and Table 65 represents the values that were used for the 2018 and 2019 reporting scale.

As explained in section 5.2.3, the item bank was recalibrated in 2019 and the 2019 item parameter and $\theta$ scale will be the underlying scale going forward. Because $\theta_{j\ 2019} = \theta_{j\ 2018} + B$ , the reporting scale is linear transformation of the 2019 scale, with the slope and intercept updated as follows:

$$\begin{aligned} SS &= a * \hat{\theta}_{MMLE,2018} + b_{2018} \\ &= a * \left(\hat{\theta}_{MMLE,2019} - B\right) + b_{2018} \\ &= a * \hat{\theta}_{MMLE,2019} + b_{2019}, \end{aligned}$$

with $b_{2019}=b_{2018}-a*B$. Table 65 represents the updated slope and intercept for the linear transformation of the 2019 $\theta$ scale. Because the LOT and HOT are specified to correspond to values of the estimated mean minus/plus four times the estimated standard deviation of $\theta$, they are updated as well. The updated linear transformation ensures that the scales remain comparable across years.

*Table 64: Science Theta and Corresponding Scaled-Score Limits for Extreme Ability Estimates (for 2018 θ scale)*

| Grade | Slope | Intercept | Lowest Obtainable Theta (LOT) | Highest Obtainable Theta (HOT) | Lowest Obtainable Scale Score (LOSS) | Highest Obtainable Scale Score (HOSS) |
|---|---|---|---|---|---|---|
| 5 | 16.009 | 547.156 | -2.94 | 3.30 | 500 | 600 |
| 8 | 18.768 | 847.165 | -2.51 | 2.81 | 800 | 900 |
| 11 | 15.969 | 1146.898 | -2.93 | 3.32 | 1100 | 1200 |

*Table 65: Science Theta and Corresponding Scaled-Score Limits for Extreme Ability Estimates (for 2019 θ scale)*

| Grade | Slope | Intercept | Lowest Obtainable Theta (LOT) | Highest Obtainable Theta (HOT) | Lowest Obtainable Scale Score (LOSS) | Highest Obtainable Scale Score (HOSS) |
|---|---|---|---|---|---|---|
| 5 | 16.009 | 547.822 | -2.98 | 3.25 | 500 | 600 |
| 8 | 18.768 | 847.622 | -2.53 | 2.79 | 800 | 900 |
| 11 | 15.969 | 1149.161 | -3.07 | 3.18 | 1100 | 1200 |

### 6.2.6. Rules for Calculating Performance Levels

Performance levels and corresponding cut scores were set during standard setting in the summer of 2018. Students are classified into one of four performance levels, based on their total score. Table 66 contains the score ranges on the reporting scale metrics for each of the grades.

*Table 66: Performance Levels for Science by Grade*

| Grade | Level 1 Below Proficient | Level 2 Approaching Proficient | Level 3 Proficient | Level 4 Above Proficient |
|---|---|---|---|---|
| 5 | 500–543 | 544–553 | 554–565 | 566–600 |
| 8 | 800–844 | 845–853 | 854–869 | 870–900 |
| 11 | 1100–1145 | 1146–1152 | 1153–1175 | 1176–1200 |

**Strengths and Weaknesses for Disciplines Relative to Proficiency Cut Score**

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

- if ($\hat{\theta}_{discipline} < \theta_{proficient} - 1.5 * SEM\left(\hat{\theta}_{discipline}\right)$), then performance is classified as *Low*;
- if $(\theta_{proficient} - 1.5 * SEM\left(\hat{\theta}_{discipline}\right) \leq \hat{\theta}_{discipline} < \theta_{proficient} + 1.5 * SEM\left(\hat{\theta}_{discipline}\right))$ , then performance is classified as *At or Approaching*; and
- if ($\hat{\theta}_{discipline} \geq \theta_{proficient} + 1.5 * SEM\left(\hat{\theta}_{discipline}\right)$), then performance is classified as *On or Above*,

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 *Low*, and the HOT is always classified as *On or Above*.

### 6.2.7. Disciplinary Core Ideas Level Reporting

**Relative to Overall Performance**

For aggregated units (classrooms, schools, districts), there are reporting at levels below the science discipline level. In 2017–2018, reports were provided at the level of Disciplinary Core Ideas (DCI). Same reports were provided in 2018-2019.

The method for DCI reports is based on the use of residuals. The residuals for an individual student are aggregated within a DCI.

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,

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

where $\boldsymbol{\tau}_i$ is the vector of parameters for assertion *i* (e.g., for the Rasch testlet model, $\boldsymbol{\tau}_i = b_i$). 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

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

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 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 is indicated as well.

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

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

**Relative to Proficiency Cut Score**

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

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

The following is reported for DCIs for aggregate units:

- If $\bar{\delta}_{DCIg} \leq -\ 1.5 * SEM\left(\bar{\delta}_{DCIg}\right)$, then performance is *below* the proficiency cut score.
- If $\bar{\delta}_{DCIg} \geq 1.5 *\ SEM\left(\bar{\delta}_{DCIg}\right)$, then performance is *above* the proficiency cut score.
- Otherwise, performance 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 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, here two sources of AIR's quality control system are described. First, quality assurance (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, which yields item statistics, blueprint match rates, and item exposure rate reports. Table 67 provides a summary of indicators generated from each QA report.

*Table 67: Overview of Quality Assurance Reports*

| QA Report | Purpose | Rationale |
|---|---|---|
| *Item Statistics* | To confirm whether items work as expected | Early detection of errors (key errors for selected-response items and scoring errors for constructed-response, performance, or technology-enhanced items) |
| *Blueprint Match Rates* | To monitor unexpected low blueprint match rates | Early detection of unexpected blueprint match issues |
| *Item Exposure Rates* | To monitor unlikely high exposure rates of items or passages or unusually low item pool usage (high unused items/passages) | Early detection of any oversight in the blueprint specification |

### 7.1.1. Item Statistics Report

The item statistics 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 test-item performance, 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 item response theory. 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. The criteria for flagging and reviewing ELA and Mathematics items is provided in Table 68, and a description of the statistics is provided below in this section. 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 4.6, the classical statistics of all AIRCore operational and field-test science items are presented in Appendix K.

*Table 68: Thresholds for Flagging Items in Classical Item Analysis, ELA and Mathematics*

| Analysis Type | Flagging Criteria |
|---|---|
| Item Discrimination | Point biserial correlation for the correct response is < 0.10. |
| Distractor Analysis | Point biserial correlation for any distractor response is > 0. |
| Item Difficulty | The proportion of students (*p*-value) is 0 or 1. |

### 7.1.2. Blueprint Match Reports

The QA system generates blueprint match reports at the content standards level and for other content requirements, such as strand or depth of knowledge (DOK) level for ELA and Mathematics, or strand and affinity group 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 are met, the number of administrations in which the blueprint requirements are not met, and, for administrations in which specifications are not met, the number of items by which the requirement is not met.

While simulation results described in Appendix A (ELA and mathematics) and Volume 2 (science) indicate that the configuration resulted in test administrations meeting all blueprint match requirements, it is also important to evaluate the blueprint match rate for actual test administrations. Appendix F shows the detailed comparison for simulation and operational blueprint match for ELA and mathematics. Across all grades and subjects, every test met the blueprint specifications with a 100% match at the reporting category level.

For Science, blueprint match is discussed in detail in Volume 2 for both simulated and operational test administrations.

### 7.1.3. Item Exposure Report

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 which should be targeted for future item development. Such item overexposure is also usually anticipated in the simulation studies used to configure the adaptive algorithm.

Appendix G shows the item exposure rates for the operational test administrations for ELA and mathematics. As is consistent with the simulation results described in Appendix A, in spring 2019 most test items were administered to 20% or fewer test takers.

For Science, 22% to 29% of items were administered to 20% or more test takers across all grades. More details are discussed in Volume 2.

## 7.2. SCORING QUALITY CONTROL

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 maximum likelihood estimates for ELA and Mathematics and marginal maximum likelihood estimates for Science, using the procedures described within this report. Scores are approved and published by the New Hampshire Department of Education only when the two independent systems match.

# 8. References

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

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.

Bock, R. D., & Zimowski, M. F. (1997). Multiple Group IRT. In van der Linden W. J., Hambleton R. K. (Eds.), *Handbook of modern item response theory*. New York : Springer-Verlag.

Cai, L. (2017). Flexmirt version 3.51: Flexible multilevel multidimensional item response theory analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group, LLC.

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

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 bi-factor analysis. *Psychometrika*, *57*, 423–436. doi:10.1007/BF02295430

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

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer-Verlag.

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.

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

Muraki, E. (1992). A generalized partial credit model: Applications of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159–176.

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

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

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

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

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 assessment: 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.