Maine Educational Assessment

Grades 5, 8, and 11 Science

MeCAS Part I

2018–19 Technical Report

© 2019 Cognia, Inc.

TABLE OF CONTENTS

CHAPTER 1 OVERVIEW OF MAINE EDUCATIONAL ASSESSMENT SCIENCE 6
1.1 PURPOSE OF THE ASSESSMENT SYSTEM 6
1.2 CURRENT YEAR UPDATES 6
CHAPTER 2 TEST ADMINISTRATION 7
2.1 RESPONSIBILITY FOR ADMINISTRATION 7
2.2 ADMINISTRATION PROCEDURES 7
2.3 PARTICIPATION REQUIREMENTS AND DOCUMENTATION 7
2.4 ADMINISTRATOR TRAINING 8
2.5 DOCUMENTATION OF ACCOMMODATIONS 8
2.6 TEST SECURITY 9
2.7 TEST AND ADMINISTRATION IRREGULARITIES 9
2.8 TEST ADMINISTRATION WINDOW 10
2.9 CUSTOMER CARE CENTER 10
CHAPTER 3 TEST DESIGN AND DEVELOPMENT 11
3.1 TEST SPECIFICATIONS 11
3.1.1 Criterion-Referenced Test 11
3.1.2 Item Types 11
3.1.3 Description of Test Design 11
3.2 SCIENCE TEST SPECIFICATIONS 12
3.2.1 Standards 12
3.2.2 Item Types 13
3.2.3 Test Design 13
3.2.4 Blueprints 13
3.2.5 Depth of Knowledge 14
3.2.6 Use of Calculators and Reference Sheets 15
3.3 TEST DEVELOPMENT PROCESS 15
3.3.1 Item Development 15
3.3.2 Item Reviews at Cognia 15
3.3.3 Item Reviews at State Level 15
3.3.4 Bias and Sensitivity Review 16
3.3.5 External Expert Review 16
3.3.6 Reviewing and Refining 16
3.3.7 Item Editing 16
3.3.8 Item Selection and Operational Test Assembly 17
3.3.9 Operational Test Draft Review 18
3.3.10 Alternative Presentations 18
CHAPTER 4 SCORING 19
4.1 MACHINE-SCORED ITEMS 19

4.2 PERSON-SCORED ITEMS 19
4.2.1 Scoring Locations and Staff 20
4.2.2 Benchmarking Meetings 20
4.2.3 Scorer Recruitment and Qualifications 21
4.2.4 Methodology for Scoring Polytomous Items 21
4.2.5 Scorer Training 22
4.2.6 Leadership Training 24
4.2.7 Scoring Quality Control Methods 24
4.2.8 Scoring Quality Control Reports 27
CHAPTER 5 QUALITY ASSURANCE OF DATA 29
CHAPTER 6 CLASSICAL ITEM ANALYSIS 30
6.1 CLASSICAL DIFFICULTY AND DISCRIMINATION INDICES 30
6.2 DIFFERENTIAL ITEM FUNCTIONING 32
6.3 DIMENSIONALITY ANALYSES FOR MAINE SCIENCE 33
CHAPTER 7 ITEM RESPONSE THEORY SCALING AND EQUATING 37
7.1 ITEM RESPONSE THEORY 38
7.2 ITEM RESPONSE THEORY RESULTS 39
7.3 EQUATING 41
7.4 EQUATING RESULTS 41
7.5 ACHIEVEMENT STANDARDS 42
7.6 REPORTED SCALED SCORES 43
7.6.1 Description of Scale 43
7.6.2 Calculations 44
7.6.3 Distributions 45
CHAPTER 8 RELIABILITY 46
8.1 RELIABILITY AND STANDARD ERRORS OF MEASUREMENT 47
8.2 SUBGROUP RELIABILITY 47
8.3 SUBCATEGORY RELIABILITY 48
8.4 RELIABILITY OF ACHIEVEMENT-LEVEL CATEGORIZATION 48
8.4.1 Accuracy and Consistency 50
8.5 INTERRATER CONSISTENCY 50
CHAPTER 9 VALIDITY 52
9.1 QUESTIONNAIRE DATA 53
9.1.1 Time Spent 53
9.1.2 Self-Image 54
9.1.3 Attitude Toward Content Area 55
9.1.4 Match of Questions to What Is Learned in School 57
9.1.5 Difficulty of Assessment 59
REFERENCES 61
APPENDICES 63
APPENDIX A MAINE TECHNICAL ADVISORY COMMITTEE MEMBERS
APPENDIX B ITEM REVIEW AND BIAS AND SENSITIVITY REVIEW COMMITTEE MEMBERS

APPENDIX C DEMOGRAPHIC DATA OF PARTICIPANTS
APPENDIX D ACCOMMODATION FREQUENCIES
APPENDIX E ACCOMMODATIONS GUIDE
APPENDIX F PROCESSING AND REPORTING BUSINESS REQUIREMENTS
APPENDIX G INTERRATER AGREEMENT
APPENDIX H CLASSICAL ITEM ANALYSIS
APPENDIX I DIFFERENTIAL ITEM FUNCTIONING
APPENDIX J ITEM RESPONSE THEORY PARAMETERS
APPENDIX K TEST CHARACTERISTIC CURVES AND TEST INFORMATION FUNCTIONS
APPENDIX L DELTA AND RESCORE LOOKUP TABLES
APPENDIX M α-PLOTS AND b-PLOTS
APPENDIX N RAW TO SCALED SCORE LOOKUP TABLES
APPENDIX O SCORE DISTRIBUTIONS
APPENDIX P RELIABILITY
APPENDIX Q DECISION ACCURACY AND CONSISTENCY

CHAPTER 1 OVERVIEW OF MAINE EDUCATIONAL ASSESSMENT SCIENCE

The Maine Educational Assessment (MEA) currently assesses science and is administered to all students in grade 5, grade 8, and third year of high school via standard administration, administration with accommodations, and/or alternate assessment. The test was administered to approximately 38,000 students during April and May 2019. The alternate assessment administration began in December 2018 and concluded at the end of April 2019.

The MEA is designed to be the measure of both Maine's academic content standards in science and the *2007 Maine Learning Results* (MLRs), and to identify the knowledge and skills essential to prepare Maine students for work, higher education, citizenship, and personal fulfillment. These academic content standards express what students should know and should be able to do at various checkpoints during their education. They were developed using grade span standards in science (pre-K–2, 3–5, 6–8, and 9–12).

1.1 PURPOSE OF THE ASSESSMENT SYSTEM

The purpose of Maine's Comprehensive Assessment System is to provide point-in-time information about the academic achievement and progress of Maine students. The MEA is one portion of this system and provides information in science. Student results are reported according to academic achievement descriptors utilizing cut scores established in standard setting for each of four achievement levels: Well Below State Expectations, Below State Expectations, At State Expectations, Above State Expectations. The results from this assessment and others provide educators and the public with information to guide future educational practices to meet the needs of students, while monitoring the continuous improvement efforts of schools, school administrative units (SAUs), and the state of Maine in achieving a world-class education system for all students.

1.2 CURRENT YEAR UPDATES

There were not any significant changes to the 2018–19 MEA Science testing program.

CHAPTER 2 TEST ADMINISTRATION

2.1 RESPONSIBILITY FOR ADMINISTRATION

As indicated in the *Principal/Test Coordinator Manual*, principals and/or their designated MEA coordinators were responsible for the proper administration of the MEA Science assessment. Manuals were used to ensure the uniformity of administration procedures from school to school. These manuals—the *Principal/Test Coordinator Manual* and the *Test Administrator Manual*—stress the importance of test security and ethical administration while the tests are in the schools, and they contain explicit directions and scripts for test administrators to read aloud to test takers. These documents may be accessed on the Maine DOE's website at http://www.maine.gov/doe/Testing_Accountability/MECAS/materials/meascience.

2.2 ADMINISTRATION PROCEDURES

MEA coordinators (principals or their designees) were instructed to read the *Principal/Test Coordinator Manual* prior to testing and to be familiar with the instructions given in the *Test Administrator Manual*. The *Principal/Test Coordinator Manual* included checklists to help each school prepare for testing. The checklists outlined tasks for the schools to perform before, during, and after test administration. Along with these checklists, the *Principal/Test Coordinator Manual* outlined the nature of the testing materials being sent to each school and discussed test security issues, such as how to inventory materials, how to track the materials at all times during administration, and how to return the materials once testing was completed. The *Test Administrator Manual* also included checklists to help administrators prepare themselves, their classrooms, and their students for the administration of the test. The manual contained sections that detailed the procedures to be followed for each test session. The *Principal/Test Coordinator Manual* and *Test Administrator Manual* gave instructions on preparing the material for its return to Cognia. Principals were required to complete an online Principal Certification of Proper Administration (PCPA) form at the conclusion of testing, certifying that all testing was administered according to MEA protocols.

2.3 PARTICIPATION REQUIREMENTS AND DOCUMENTATION

The intent is for *all* students in grade 5, grade 8, and third year of high school to participate in the MEA Science assessment through standard administration, administration with accommodations, and/or alternate assessment. Any student who is absent during any session of the MEA is expected to take a makeup test within the testing window.

On those occasions where it was deemed necessary to exclude a student from sections of the assessment or from the assessment as a whole because of special considerations (medical emergency only), schools were asked to seek the approval of the Maine DOE MEA review team. The names of the excluded students were forwarded to Cognia so these students would not be included in any reports or as part of the denominator representing the total number of students. Appendix C presents student participation in the MEA Science for all students by demographic group.

2.4 ADMINISTRATOR TRAINING

In addition to distributing the *Principal/Test Coordinator Manual* and the *Test Administrator Manual*, the Maine DOE, along with Cognia, provided a statewide test administration webinar to train and inform school personnel about the MEA testing procedures. This training was posted on the Maine DOE website at http://www.maine.gov/doe/Testing_Accountability/MECAS/materials/meascience.

2.5 DOCUMENTATION OF ACCOMMODATIONS

The approved accommodations for eligible students were listed on page 2 of the student answer booklet. This information was coded by the appropriate staff member after testing was completed. The *Principal/Test Coordinator Manual* and the *Test Administrator Manual* provided directions for coding the information related to accommodations.

All students who were considered for accommodations on the MEA had their individual situations reviewed by a team within the school prior to the time of testing. For every student with an identified exceptionality requiring an individualized education program (IEP), schools were required to hold an IEP team meeting that addressed the student's needs for accommodations. For other students needing test accommodations who did not have an identified exceptionality, a meeting was required that included one of the student's teachers, the building principal, related services personnel, and, whenever possible, the student's parents or guardians. If it was not possible for the parents or guardians to attend the meeting, they were notified of the committee's recommendations for accommodations prior to the time of testing.

Recommended accommodations were to be consistent with those accommodations already being used in the student's instructional program. Any such accommodations were reflected either in the minutes of the IEP team meeting (for students requiring an IEP) or in a statement prepared for the cumulative folders of students not requiring IEPs. Schools were given the following statement as a model: *The student will participate in the [__]th grade Maine Educational Assessment as scheduled during the month of May 2018 with the following accommodations*. Table 2-1 lists the numbers of students who tested with and without accommodations.

Table 2-1. 2018–19 MEA Science: Numbers of Students Tested With and Without Accommodations by Subject and Grade

Subject	*Grade*	*Number of Students Tested*	
		Without Accommodations	*With Accommodations*
Science	5	10,520	2,489
	8	11,259	1,807
	11	11,312	745

Appendix D also shows the accommodation frequencies observed for the May 2019 MEA administration. The *MEA Accommodations Guide*, which includes detailed descriptions of approved accommodations and their proper application, is presented in Appendix E.

2.6 TEST SECURITY

Maintaining test security is critical to the success of the MEA. The *Principal/Test Coordinator Manual* and the *Test Administrator Manual* explain in detail all test security measures and test administration procedures. School personnel were informed that any concerns about breaches in test security were to be reported to the school's test coordinator and/or principal immediately. The test coordinator and/or principal was responsible for immediately reporting the concern to the district superintendent and the state assessment director at the Maine DOE. Test security was also strongly emphasized at the test administration workshops. Principals were required to log on to a secure website to complete the PCPA form for each grade level tested at their school; they also had to provide the number of secure tests received from Cognia, the number of tests administered to students, and the number of secure test materials that they were returning to Cognia. Principals were instructed to submit the form by entering a unique password, which acted as their digital signature. By signing and submitting the form, the principal certified that the tests were administered according to the test administration procedures outlined in the *Principal/Test Coordinator Manual* and the *Test Administrator Manual*; that the security of the tests was maintained; that no secure material was duplicated or in any way retained in the school; and that all test materials had been accounted for and returned to Cognia. In addition, all Principal/Test Coordinators and Test Administrators are required to sign a *Test Security Agreement* including the above, recognition of penalties for violating test security protocol, and Student Data Privacy Agreement. These agreements, kept on file with the Principal/Test Coordinator, are found on the Maine DOE website at:
https://www.maine.gov/doe/Testing_Accountability/MECAS/materials/meascience.

2.7 TEST AND ADMINISTRATION IRREGULARITIES

Table 2-2 lists the numbers of irregularities that occurred during the 2018–19 administration.

Table 2-2. 18–19 MEA Science: Test and Administration Irregularities

Irregularity and/or Invalidation Reason	*Number of Students*	*Number of Invalidations*
Cheating	1	1 Session
TA/Accommodation Error	6	18 Sessions
Off Grade/Wrong Cohort	23	23 Full Tests
Totals	**30**	**n/a**

2.8 TEST ADMINISTRATION WINDOW

The high school test administration window was April 1–12, 2019. The grades 5 and 8 test administration window was April 22–May 3, 2019.

2.9 CUSTOMER CARE CENTER

To provide additional support to schools before, during, and after testing, the field was able to call the Cognia Customer Care Center. The support of this service center is essential to the successful administration of any statewide test program. The customer care center provided a centralized location that individuals in the field could call using a toll-free number to ask specific questions or report any problems they might have been experiencing. Representatives were responsible for receiving, responding to, and tracking calls, and then routing issues to the appropriate person(s) for resolution. All calls were logged into a database that included notes regarding the issue and resolution of each call.

The Maine Customer Care Center is open to receive calls from 7:30 a.m. to 4:30 p.m. EST, Monday through Friday year-round.

CHAPTER 3 TEST DESIGN AND DEVELOPMENT

3.1 TEST SPECIFICATIONS

3.1.1 Criterion-Referenced Test

The MEA is a criterion-referenced science test. Items on the MEA are developed specifically for Maine and are directly linked to Maine's science content standards. These content standards are the basis for the reporting categories and are used to help guide the development of test items.

3.1.2 Item Types

Maine educators and students are familiar with the types of items used in the assessment program. The types of items and their functions are described below:

- **Multiple-choice (MC)** items are used to provide breadth of coverage within a content area. Because they require no more than a minute for most students to answer, multiple-choice items make efficient use of limited testing time and allow for coverage of a wide range of knowledge and skills.
- **Constructed-response (CR)** items typically require students to use higher-order thinking skills—evaluation, analysis, summarization, and so on—to construct satisfactory responses. Constructed-response items take most students approximately 5 to 10 minutes to complete.
- ***Modules** consist of an informational passage and a collection of multiple-choice and constructed-response items that refer to and assess students on the content contained within the passage. This provides students with a richer context requiring them to delve deeper and investigate a topic from multiple content perspectives.

  ** Module item types are being pilot-tested by the Maine DOE at this time and as such are treated as field-test items.*

Note that the use of released MEA items to prepare students to respond to multiple-choice and constructed-response items is appropriate and encouraged.

3.1.3 Description of Test Design

The MEA is structured using both common and field-test items. Common items are taken by all students in a given grade level. Student scores are based only on common items. Field-test items are divided among the forms of the test for each grade level. Each student takes only one form of the test and therefore answers a fraction of the field-test items. Field-test items are not identifiable to test takers and have a

negligible impact on testing time. Because all students participate in the field test, it provides the minimum sample size (750–1,500 students per item) needed to produce reliable data that can be used to inform item selection for future tests.

3.2 SCIENCE TEST SPECIFICATIONS

3.2.1 Standards

The 2018–19 MEA Science test items were aligned to the content standards of "D: The Physical Setting" and "E: The Living Environment," as described in the Science and Technology section of Maine's *Learning Results: Parameters for Essential Instruction*. No other science content standards were subject to statewide assessment. Content specialists used the content standards, performance indicators, and descriptors to help guide the development of test questions, which may address one or more of the performance indicators listed below.

D. The Physical Setting

- D1: Universe and Solar System—Students explain the physical formation and changing nature of our universe and solar system, and how our past and present knowledge of the universe and solar system developed.
- D2: Earth—Students describe and analyze the biological, physical, energy, and human influences that shape and alter Earth systems.
- D3: Matter and Energy—Students describe the structure, behavior, and interaction of matter at the atomic level and the relationship between matter and energy.
- D4: Force and Motion—Students understand that the laws of force and motion are the same across the universe.

E. The Living Environment

- E1: Biodiversity—Students describe and analyze the evidence for relatedness among and within diverse populations of organisms and the importance of biodiversity.
- E2: Ecosystem—Students describe and analyze the interactions, cycles, and factors that affect short-term and long-term ecosystem stability and change.
- E3: Cells—Students describe the structure and function of cells at the intracellular and molecular levels, including differentiation to form systems, interactions between cells and their environment, and the impact of cellular processes and changes on individuals.
- E4: Heredity and Reproduction—Students examine the role of DNA in transferring traits from generation to generation, in differentiating cells, and in evolving new species.
- E5: Evolution—Students describe the interactions between and among species, populations, and environments that lead to natural selection and evolution.

3.2.2 Item Types

The science test includes multiple-choice and constructed-response items. Each multiple-choice item requires students to select the correct response from four choices. Each type of item is worth a specific number of points in the student's total science score, as shown in Table 3-1.

Table 3-1. 2018–19 MEA Science: Item Types

Item Type	*Possible Score Points*
MC	0 or 1
CR	0, 1, 2, 3, or 4

MC = multiple-choice
CR = constructed-response

Consistent with the new annual release policy, 25% of the multiple-choice items and 50% of the constructed-response items were released from the 2018–19 MEA Science test. A practice test comprising released science items is available on the Maine DOE website at http://www.maine.gov/doe/Testing_Accountability/MECAS/supports. Schools are encouraged to incorporate the use of the released items in their instructional activities so that students will be familiar with the types of questions found on the test.

3.2.3 Test Design

Table 3-2 summarizes the numbers and types of items that were used to compute student scores on the 2018–19 MEA Science test. Additionally, each test form had eight multiple-choice field-test items and one constructed-response field-test item that did not affect student scores.

Table 3-2. 2018–19 MEA Science: Science Items

Grade	*Session 1*	*Session 2*	*Session 3*	*TOTAL*	
				MC	*CR*
5	8 MC, 2CR	6 MC, 1 CR	18 MC, 1 CR	32	4
8	11 MC, 2 CR	11 MC, 1 CR	18 MC, 1 CR	40	4
11	16 MC, 2 CR	24 MC, 2 CR	N/A	40	4

MC = multiple-choice; CR = constructed-response
Note: Grade 11 does not have three sessions.

3.2.4 Blueprints

Table 3-3 shows the distribution of points across the science standards. For grade 5, D1–D2 had one constructed-response item; D3–D4 had one constructed-response item; and E1–E5 had two constructed-response items. For grade 8, D1–D2 had one constructed-response item; D3–D4 had one constructed-response

item; and E1–E5 had two constructed-response items. For grade 11, D1–D2 had one constructed-response item; D3–D4 had one constructed-response item; and E1–E5 had two constructed-response items.

Table 3-3. 2018–19 MEA Science: Distribution of Score Points

Science Standards	*Grade 5*	*Grade 8*	*Grade 11*
D1–D2 (Earth / Space)	12	17	12
D3–D4 (Physical)	12	17	22
E1–E5 (Life)	24	22	22
Total Score Points	48	56	56

3.2.5 Depth of Knowledge

Each item on the MEA Science test is assigned a Depth of Knowledge (DOK) level. The DOK level reflects the complexity of mental processing students use to answer an item. DOK is not synonymous with difficulty. Each of the four DOK levels is described below.

- **Level 1 (Recall).** This level requires the recall of information such as a fact, definition, term, or simple procedure. These items require students to demonstrate a rote response, use a well-known formula, or follow a set procedure.
- **Level 2 (Skill/Concept).** This level requires mental processing beyond that of recalling or reproducing a response. These items require students to make some decisions about how to approach the item.
- **Level 3 (Strategic Thinking).** This level requires reasoning, planning, and using evidence. These items require students to handle more complexity and abstraction than items at the previous two levels.
- **Level 4 (Extended Thinking).** This level requires planning, investigating, and complex reasoning over an extended period of time. Students are required to make several connections within and across content areas. This level may require students to design and conduct experiments. Because of the nature of this level, there are no Level 4 items on the MEA.

It is important that the MEA measures a range of depths of knowledge. Table 3-4 shows the distribution of score points across the DOK levels used on the MEA.

Table 3-4. 2018–19 MEA Science: Distribution of Score Points Across Depth of Knowledge (DOK)

DOK Level	*Grade 5*	*Grade 8*	*Grade 11*
1	4	8	6
2	35	39	33
3	9	9	17
Total	48	56	56

3.2.6 Use of Calculators and Reference Sheets

Calculators are not used or needed when taking the science tests. There are no science reference sheets.

3.3 TEST DEVELOPMENT PROCESS

3.3.1 Item Development

Items used on the MEA are developed and customized specifically for use on the MEA and are consistent with Maine content standards and performance indicators. Cognia test developers work with the Maine state science specialist and with Maine educators to verify the alignment of items to the appropriate Maine content standards.

The development process combined the expertise of Cognia test developers, the Maine state science specialist, and committees of Maine educators to help ensure that items met the needs of the MEA program. All items used on the common portions of the MEA were reviewed by a committee of Maine content experts, a committee of Maine bias experts, and three external content experts.

3.3.2 Item Reviews at Cognia

The test developers at Cognia reviewed newly developed items for

- alignment to the intended content standard;
- item integrity, including science content and structure, format, clarity, possible ambiguity, and single correct answer;
- appropriateness and quality of graphics;
- appropriateness of scoring guide descriptions and distinctions;
- completeness of associated item documentation (e.g., scoring guide, content codes, key, grade level, DOK, and contract identified); and
- appropriateness for the designated grade level.

3.3.3 Item Reviews at State Level

A committee of Maine classroom teachers from across the state reviewed the items before field-testing. Teacher participants were selected based on their content-area expertise and grade-level familiarity. The purpose of the review was to evaluate new items for the embedded field test and determine their suitability for the assessment by answering the following four questions:

- Does the item align with the assigned content standard and performance indicator?
- Is the science content accurate?

- Is the science content grade-level appropriate?
- Does the item provide maximum accessibility for all students?

3.3.4 Bias and Sensitivity Review

Bias review is an essential component of the development process. During the bias review process, items were reviewed by a committee of Maine educators who represented various student subgroups, including the visually or hearing impaired and students for whom English is a second language. Items were examined for content that might cause the test to be inaccessible for these students or that might generally offend or dismay students, teachers, or parents/guardians. Being aware of these considerations in the development of assessment items and materials can avoid many unduly controversial issues, and unfounded concerns can be allayed before the test forms are produced.

3.3.5 External Expert Review

The field-test items from the previous year were classified into three groups based on science content. Three science experts (one each for Earth/space science, life science, and physical science) reviewed the group of items corresponding to their area of expertise. The expert reviewers primarily evaluated each item for correct science content. For the multiple-choice items, the experts also indicated whether the keyed answer was correct and whether it was the only correct answer among the options given. The Maine DOE state science specialist and Cognia test developers reviewed the experts' evaluations and adjusted the items as necessary.

3.3.6 Reviewing and Refining

Recommended changes from the Item Review and Bias and Sensitivity meetings, as well as the comments from the three external science experts, were reviewed and considered by the Maine DOE state science specialist. Cognia test developers made the edits that were approved by the Maine DOE state science specialist.

3.3.7 Item Editing

Cognia editors reviewed and edited the items to ensure adherence to sound testing principles and to style guidelines in the *Chicago Manual of Style*, 16th edition. These principles include the stipulations that items

- demonstrate correct grammar, punctuation, usage, and spelling;
- are written in a clear, concise style;
- contain unambiguous explanations that tell students what is required to attain a maximum score;

- are written at a reading level that allows students to demonstrate their knowledge of the subject matter being tested regardless of reading ability;
- exhibit high technical quality regarding psychometric characteristics;
- have appropriate answer options or score point descriptors; and
- are free of potentially insensitive content.

3.3.8 Item Selection and Operational Test Assembly

Cognia test developers met with the Maine state science specialist to select the common items. In preparation for the meeting, the test developers and psychometricians at Cognia considered the following in selecting sets of items to propose for the common test:

- **Content coverage/match to test design and blueprints.** The test designs and blueprints stipulate a specific number of multiple-choice and constructed-response items. Item selection for the embedded field test was based on the number of items in the existing pool of items that are eligible for the common test.
- **Item difficulty and complexity.** Item statistics drawn from the data analysis of previously field-tested items were used to ensure quality psychometric characteristics as well as similar levels of difficulty and complexity from year to year.
- **"Cueing" items.** Items were reviewed for any information that might "cue" or provide information that would help to answer another item.

The Maine state science specialist reviewed the proposed sets of items and made the final selection of items for the common test.

The test developers then sorted and laid out the items into test forms. During assembly of the test forms, the following criteria were considered:

- **Key patterns.** The sequence of keys (correct answers) was reviewed to ensure that key order appeared random.
- **Option balance.** Items were balanced across forms so that each form contained a roughly equivalent number of key options (As, Bs, Cs, and Ds).
- **Page fit.** Item placement was modified to ensure the best fit and arrangement of items on any given page.
- **Relationships among forms.** Although field-test items differ from form to form, these items must take up the same number of pages in all forms so that sessions begin on the same page in every form. Therefore, the number of pages needed for the longest form often determines the layout of each form.
- **Visual appeal.** The visual accessibility of each page of the form was always taken into consideration, including such aspects as the amount of "white space," the density of the test, and the number of graphics.

3.3.9 Operational Test Draft Review

After the forms were laid out as they would appear in the final test booklets, the forms were again thoroughly reviewed by Cognia editors to ensure that the items appeared exactly as intended. Any changes made during test construction were reviewed and approved by the test developer. The Maine DOE state science specialist then read the forms for final approval.

3.3.10 Alternative Presentations

The common test for each grade was translated into the Braille format as determined by student's Individual Educational Program (IEP) by the American Printing House for the Blind, a subcontractor that specializes in test materials for blind and visually impaired students. In addition, Form 1 for each grade was adapted into a large-print version.

CHAPTER 4 SCORING

4.1 MACHINE-SCORED ITEMS

Multiple-choice item responses were compared to scoring keys using item analysis software. Correct answers were assigned a score of one point and incorrect answers were assigned zero points. Student responses with multiple marks and blank responses were also assigned zero points.

The hardware elements of the scanners monitor themselves continuously for correct data reads, as does the software that drives these scanners. Standard checks include recognition of a sheet that does not belong or is upside down or backward and identification of critical data that are missing (e.g., a student ID number), test forms that are out of range or missing, and page or document sequence errors. When a problem is detected, the scanner stops and displays an error message directing the operator to investigate and correct the situation.

4.2 PERSON-SCORED ITEMS

The images of student responses to constructed-response items were hand-scored using Cognia's electronic scoring system, iScore. Use of iScore minimized the need for scorers to physically handle answer booklets and related scoring materials. Thus, student confidentiality was easily maintained, since all MEA scoring was "blind" (i.e., district, school, and student names were not visible to scorers). The iScore system maintained the linkage between the student response images and their associated test booklet numbers.

Through iScore, qualified scorers accessed electronically scanned images of student responses. Scorers evaluated each response and recorded its score via keypad or mouse entry through the iScore system. When a scorer finished evaluating one response, the next response appeared immediately on the computer screen.

Imaged responses from all answer booklets were sorted into item-specific groups for scoring purposes. Scorers reviewed responses from only one test item at a time. However, imaged responses from a student's entire booklet were always available for viewing by leadership when necessary, such as to score wrong locations or review books identified with a crisis response. The physical booklet was also available to the Scoring Content Specialist on-site. (Scoring Content Specialist and other scoring roles are described in the following section.)

The use of iScore also helped to ensure that access to all student response images was limited only to scorers and other authorized staff within Cognia scoring facilities.

4.2.1 Scoring Locations and Staff

The iScore database, its operation, and its administrative controls are all based in Dover, NH. Table 4-1 presents the locations by grade where 2018–19 MEA test item responses were scored.

Table 4-1. 2018–19 MEA Science: Operational Scoring Locations by Grade

Grade	*Dover, NH*	*Menands, NY*	*Longmont, CO*
5	x		
8			x
11		x	

Staff Positions

The following staff members were involved with scoring the 2018–19 MEA responses:

- The MEA Scoring Project Manager oversaw communication and coordination of scoring.
- The iScore operational manager coordinated technical communication and managed scorer access to the iScore system.
- A Scoring Content Specialist in the science content area ensured consistency of scoring for all grades tested in that content area. The Scoring Content Specialist also provided read-behind activities (defined in a later section) for Scoring Supervisors. Scoring Supervisors, selected from a pool of experienced Scoring Team Leaders (STLs) for their ability to score accurately and to instruct and train scorers, participated in benchmarking activities for each specific grade of the science content area. Scoring Supervisors provided read-behind activities (defined in a later section) for STLs at the scoring site. For the MEA, the role of Scoring Content Specialist was filled by the Scoring Science Content Manager.
- Scoring Supervisors, who are typically selected from experienced STLs, worked under the direct supervision of the Scoring Content Specialist to prepare training materials, train and supervise scorers, prepare materials for field-test meetings, attend benchmarking meetings, and train and supervise the scoring of field-test items.
- STLs, selected from a pool of skilled and experienced scorers, provided read-behind activities (defined in a later section) for the scorers at their scoring tables (2 to 12 scorers at each table). The ratio of Scoring Supervisors and STLs to scorers was approximately 1 : 11.
- Scorers at the scoring sites scored operational and field-test MEA 2018–19 student responses. Recruitment of scorers is described in Section 4.2.3.

4.2.2 Benchmarking Meetings

In preparation for implementing MEA scoring guidelines, Cognia scoring staff prepared and facilitated benchmarking meetings held with the Maine DOE state science specialist representing the

department of education. The purpose of these meetings was to establish guidelines for scoring MEA items during the current field-test scoring session and for future operational scoring sessions.

Scoring staff selected several dozen student responses for each item that were identified as illustrative midrange examples of the respective score points. The Scoring Content Specialist presented these responses to the MEA Science content specialist during benchmarking meetings and worked collaboratively to finalize an authoritative set of score point exemplars for each field-test item. As a matter of practice, these sets are included in the scoring training materials each time an item is administered.

This repeated use of MEA-approved sets of midrange score point exemplars helps ensure that scorers follow established guidelines each time a particular MEA item is scored.

4.2.3 Scorer Recruitment and Qualifications

For scoring the 2018–19 MEA, Cognia actively sought a diverse pool of scorers with a broad range of backgrounds: teachers, scientists, business professionals, teachers, graduate school students, and retired educators, and the like. To verify this diversity, demographic information such as gender, race, educational background-among other information - was captured electronically.

Via temporary employment agencies, Scoring Services recruited individuals who have successfully attained a four-year college degree or higher. In all cases, potential scorers were required to submit documentation (e.g., resume and/or transcripts) as evidence of meeting the education requirements. In addition, each scorer and scoring leadership staff signed a non-disclosure/confidentiality agreement, which was retained by the staffing agency at the respective scoring facility and is available for review.

Table 4-2 summarizes the qualifications of the 2018–19 MEA scoring leadership and scorers.

Table 4-2. 2018–19 MEA Science: Qualifications of Scoring Leadership and Scorers—Spring Administration

Scoring Responsibility	*Educational Credentials*				*Total*
	Doctorate	*Master's*	*Bachelor's*	*Other*	
Scoring Leadership*	2	8	6		16
Scorers	5	21	35		61

*Scoring Leadership = Scoring Supervisors and Scoring Team Leaders

4.2.4 Methodology for Scoring Polytomous Items

Possible Score Points

The ranges of possible score points for the different polytomous items are shown in Table 4-3.

Table 4-3. 2018–19 MEA Science: Possible Score Points for Polytomous Types

Polytomous Item Type	*Possible Score Point Range*
Constructed-response	0–4
Nonscorable items	Blank, Unreadable

Nonscorable Items

Scorers could designate a response as nonscorable for any of the following reasons:

- Response was blank (no attempt to respond to the question).
- Response was unreadable (illegible, too faint to see, or only partially legible/visible).
- Response was written in the wrong location (seemed to be a legitimate answer to a different question).

Note: "Unreadable" and "wrong location" responses were ultimately resolved by researching the actual answer document (electronic copy or hard copy, as needed) to identify the correct location (in the answer document) or to more closely examine the response and then assign a score.

Scoring Procedures

Scoring procedures for polytomous items included both single scoring and double-blind scoring. Single-scored items were scored by one scorer. Double-blind-scored items were scored independently by two scorers whose scores were tracked for interrater agreement (for further discussion of double-blind scoring and interrater agreement, see Section 4.2.7 and Appendix G).

4.2.5 Scorer Training

Scorer training began with an introduction of the on-site scoring staff and an overview of the purpose and goals of MEA program (including discussion about the security, confidentiality, and proprietary nature of testing materials, scoring materials, and procedures).

Next, scorers were trained by Scoring Supervisors with a thorough review and discussion of the scoring guide for each item to be scored. Each item-specific scoring guide included the actual item and score point descriptions.

Following review of an item's scoring guide, scorers reviewed the particular response set organized for that training: anchor and practice sets. Qualifying set(s) were administered thereafter. (These are defined on the following pages.)

Anchor Set

Scorers first reviewed an anchor set of exemplar responses for an item which was approved by the science content specialist representing the Maine DOE. Responses in anchor sets are typical, mid-range examples of score points. They are also true exemplars, i.e., their scores cannot be changed by anyone other than by a joint decision between the MEA client and Cognia Scoring Services staff. Each anchor set contains one client-approved sample response per score point and includes, where necessary, the rationale provided by the MEA Science content specialist for choosing that response as a representative anchor response for that particular score point. The set includes a second sample response for each score point if there is more than one plausible way to illustrate the merits and intent of a score point.

Responses were read aloud to scorers in descending score order. Announcing the true score of each anchor response, trainers facilitated discussion of responses in relation to score point descriptions to help scorers internalize the typical characteristics of score points.

This anchor set continued to serve as a reference for scorers as they went on to qualification, scoring, and recalibration activities for that item.

Practice Set

To mimic live scoring, scorers practiced applying the scoring guide and anchors to responses in the practice set. The practice set included a selection of student responses designed to help establish both the full score point range and the range of possible responses within each score point. The practice set often included unusual responses that were less clear or solid (e.g., shorter than normal, employing atypical approaches, simultaneously containing very low and very high attributes, and written in ways difficult to decipher). Responses in the practice set were presented in randomized score point order.

After scorers independently read and scored a practice set response, trainers would poll scorers or use an online training system reports to record the initial range of scores. Trainers then led a group discussion of one or two responses, directing scorer attention to difficult scoring issues (e.g., the borderline between two adjacent score points). Throughout the process, trainers modeled how to evaluate student responses by referring to the scoring standards as defined by the scoring guide and exemplified in the anchor set.

Qualifying Set

After the practice set was completed, scorers were required to score responses accurately and reliably in qualifying sets assembled for constructed-response items. The 10 responses in each qualifying set were selected from an array of responses that clearly represented and illustrated the range of score points for that item as reviewed and approved by the state specialist. Hard copies of the responses were available to scorers after the qualification round, if needed, for additional training.

To be eligible to score, scorers were required to demonstrate accuracy rates of at least 80% exact agreement (i.e., to exactly match the predetermined score on at least 8 of the 10 responses) and at least 90% exact or adjacent agreement (i.e., to exactly match or be within one score point of the predetermined score on 9 or 10 of the 10 responses). Scorers were allowed one discrepant score (i.e., 1 score out of 10 that was more than one score point from the predetermined score) provided they had at least eight exact scores.

Retraining

Scorers who did not pass the first qualifying set were retrained by reviewing their performance with scoring leadership and then scoring a second qualifying set of responses. If they achieved the required accuracy rate on the second qualifying set, they could score operational responses.

Scorers who did not achieve the required scoring accuracy rates on the second qualifying set were not allowed to score responses for that item. Instead, they either began training on a different item or were dismissed from scoring for that day.

4.2.6 Leadership Training

Scoring Supervisors and select STLs were trained in a separate session prior to scorer training. In addition to discussing the items and their responses, Scoring Supervisors and STLs went into greater detail about the scoring rationale as provided by the client so that they are well-equipped to handle questions from scorers.

4.2.7 Scoring Quality Control Methods

Scorers were required to demonstrate and maintain their ability to score student responses accurately and consistently throughout the scoring process. The iScore scoring tools that measured MEA scoring quality were as follows:

- Embedded Papers
- Read-Behind Scoring
- Double-Blind Scoring
- Recalibration Sets

Any scorer whose accuracy fell below the expected rate related to the monitoring method was retrained on that item. Upon approval by the Scoring Supervisor or Scoring Content Specialist, the scorer could resume scoring. Scorers who met or exceeded the expected accuracy rates continued evaluating student responses.

The accuracy rate required of a scorer to qualify to score responses live was higher than that required to continue to score responses live. The reason for the difference is that an “exact score” in double-blind scoring requires that two scorers agree on the same score for potentially borderline responses, i.e., an exact

score is dependent on two peers agreeing on responses that often do not sit neatly in the middle of their score point spectrum. An "exact score" in qualification requires that a scorer match a score that is pre-established as sitting in the middle of a score point. Scoring Services employs a variety of monitoring techniques to ensure scorer accuracy throughout the process of live scoring.

Embedded Papers

Embedded Papers are previously scored responses that are loaded ("embedded") into iScore and distributed "blindly" to scorers during scoring. Embedded Papers may be chosen either before or during scoring and are inserted into the scoring queue so that they do not distinguish themselves from all other live student responses.

Ten Embedded Papers were distributed at random points throughout the first full day of scoring to ensure that scorers were sufficiently calibrated at the beginning of the scoring period.

Any scorer who fell below the required scoring accuracy rate was retrained before being allowed by the Scoring Supervisor to continue scoring. Once allowed to resume scoring, scoring leadership carefully monitored these scorers by increasing the number of read-behinds.

Embedded Papers were used for all constructed-response items.

Read-Behind Procedures

The practice of read-behind scoring was applied to all constructed-response items in all grades. Read-behind scoring allowed scoring leadership to monitor each scorer's scoring performance by way of an immediate real-time snapshot of the scorer's accuracy. The data that was generated by read-behind scoring presented leadership with opportunities to answer questions and to provide counsel to scorers who may have had trouble maintaining the scoring standards. iScore is designed such that the selection of any scored student responses for read-behind scoring was done without a scorer knowing which response was selected for a read-behind. The STL would, at various points throughout the scoring session, instruct the system to assign the next one, two, or three responses per scorer to be placed into the read-behind queue at a time. Responses could be pulled for all scorers who were assigned to the STL, or only for certain scorers. Each read-behind response was scored blindly by the STL; that is, each scorer's response score was only revealed to the STL after the STL had submitted his/her score to the system. The STL would then have an opportunity to compare his/her score against the score assigned by the scorer and the score-of-record (i.e., the reported score) was determined as follows:

- If there was exact agreement between the scores, no action was necessary; the scorer's score remained.
- If the scores were adjacent (i.e., differed by one point), the STL's score became the score-of-record. (A significant number of adjacent scores for a scorer triggered an individual scoring

consultation with the STL, after which the Scoring Supervisor determined whether and when the scorer could resume scoring.)

- If the scores were discrepant (i.e., differed by more than one point), the STL's score became the score of record. (This triggered an individual consultation with the STL, after which the Scoring Supervisor determined whether and when the scorer could resume scoring that item.)

Table 4-4 illustrates how scores were resolved by the read-behind method.

Table 4-4. 2018–19 MEA Science: Examples of Read-Behind Scoring Resolutions

Scorer Score	*Leadership Score*	*Score of Record*
3	4	4
2	1	1
1	3	3

STLs were tasked with conducting, on average, five read-behinds per scorer throughout each scoring half-day; however, STLs conducted a proportionally greater number of read-behinds for scorers who seemed to be struggling to maintain, or who fell below, accuracy standards.

In addition to regular read-behinds, scoring leadership could choose to conduct read-behinds on any scorer at any point during the scoring process to gain an overview of a scorer's current accuracy.

Double-Blind Scoring

Double-blind scoring refers to two scorers independently scoring a response without knowing whether the response was to be double-blind scored. The practice was applied to all constructed-response item types. The scoring system was set to automatically select 10% of student responses to be double scored. In addition, responses that were marked as blank or were assigned a condition code could also be double scored, even if not originally selected to be part of the 10% double-scoring batch.

If there was a discrepancy (a difference greater than one score point) between double-blind scores, the response was placed into an arbitration queue. Arbitration responses were reviewed by scoring leadership (STL or Scoring Supervisor) without knowledge of the two scorers' scores. Scoring leadership assigned the final score. Appendix G provides the MEA 2018–19 percentages of agreement between scorers for each common item for each grade.

Scoring leadership consulted individually with any scorer whose scoring rate fell below the required accuracy rate, and the Scoring Supervisor determined whether and when the scorer could resume scoring on that item. Once the scorer resumed scoring, scoring leadership carefully monitored the scorer's accuracy by increasing the number of read-behinds.

Recalibration Sets

Beginning with the second day of scoring, scorers were required to take an online recalibration set to determine whether they were still calibrated to the scoring standard. Each recalibration set consisted of five responses representing the entire range of possible scores, including some responses with a score point of 0. Any recalibration results that showed discrepant scores or 2 or more adjacent scores required a review with scoring leadership before the scorer could begin scoring. Recalibration sets were employed for all constructed-response items. As noted above, they were administered at the start of the second day of scoring on an item since the accuracy on the first day of scoring an item is monitored using the item's initial qualification set and the set of Embedded Papers. In the event an item was scored during a third day, newly assembled recalibration sets were administered similar to the second day's administration.

The Scoring Content Specialist voided all scores recorded on that item for that day by scorers who did not meet the accuracy requirements. Responses associated with voided scores were reset and redistributed to scorers with demonstrated accuracy for that item.

4.2.8 Scoring Quality Control Reports

iScore generated multiple reports that were used by scoring leadership to measure and monitor scorers for scoring accuracy, consistency, and productivity. Due to the complexity of scoring a large-scale assessment project such as the MEA, computer-generated reports were necessary to ensure that

- overall group-level accuracy, consistency, and reliability of scoring were maintained at acceptable levels;
- immediate, real-time individual scorer data was available to allow early intervention when necessary;
- scoring schedules were maintained.

The following reports were produced by iScore for internal use throughout each scoring day by scoring leadership (including STLs, Scoring Supervisors, Scoring Content Specialists, and the Scoring Project Manager, where applicable):

- **The Read-Behind Summary** showed the total number of read-behind responses for each scorer and noted the number and percentages of exact, adjacent, and discrepant scores with scoring leadership. Scoring leadership could generate this report by choosing options (such as "Today," "Past Week," and "Cumulative") from a drop-down menu. The report could also be filtered to select data for a particular item or across all items. This report was used in conjunction with other reports to determine whether a scorer's scores would be voided (and rescored by other qualified scorers). The benefit of this report is that it can reveal scorer to STL accuracy agreement.

- **The Double-Blind Summary** showed the total number of double-scored responses by each scorer and noted the number and percentages of exact, adjacent, and discrepant scores with second scorers. This report was used in conjunction with other reports to determine whether a scorer's scores should be voided. The benefit of this report is that it can reveal scorer to scorer accuracy agreement.
- **The Accuracy Summary** combined read-behind and double-blind data, showing the total number for the scorers, their accuracy rates, and their score point distributions.
- **The Embedded CRR Summary** showed, for each scorer (by item or across all items), the total number of responses scored, the number of Embedded Papers scored, and the numbers and percentages of exact, adjacent, and discrepant scores compared against the Scoring Content Specialist. This report was used in conjunction with other reports to determine whether a scorer's scores should be voided. The benefit of this report is that it can reveal the degree to which an individual scorer agrees with the scores assigned by the Scoring Content Specialist. Also, since Embedded Papers are administered during the first hours of scoring, this report can provide an early illustration of agreement between scorers and Scoring Content Specialists.
- **The Qualification Statistics Summary** listed each scorer by name and ID number, identified which qualifying set(s) he or she did and did not take, and, for the ones taken, his or her pass rate. In addition to the pass rates of individuals, the report also showed numbers of scorers passing or failing a qualifying set. The Scoring Supervisor used this report to determine how scorers within their assigned scoring group performed on specific qualifying sets.
- **The Summary Statistics Report** showed the total number of student responses for an item and identified, for the time at which the report was generated, the following:
    - the number of single and double-blind scorings that had been performed; and
    - the number of single and double-blind scorings yet to be performed.

CHAPTER 5 QUALITY ASSURANCE OF DATA

Quality-assurance measures are embedded throughout the entire process of data analysis. The data processor, data analyst, and psychometrician assigned to work on MEA implement quality-control checks of their respective computer programs and intermediate products. Moreover, when data are handed off to different functions within the Data and Reporting Services, and Psychometrics Departments, the sending function verifies that the data are accurate before handoff. Additionally, when a function receives a data set, the first step is to verify the data for accuracy.

Another type of quality-assurance measure is parallel processing. Students' scaled scores are assigned by a psychometrician through a process of equating and scaling. The scaled scores are also computed by a data analyst to verify that scaled scores and corresponding achievement levels are assigned accurately. Respective scaled scores and assigned achievement levels are compared across all students for 100% agreement. Different exclusions that determine whether each student receives scaled scores and/or is included in different levels of aggregation are also parallel processed. Using the *Processing and Reporting Business Requirements* document, two data analysts independently write a computer program that assigns students' exclusions. For each grade and content-area combination, the exclusions assigned by each data analyst are compared across all students. Only when 100% agreement is achieved can the rest of the data analysis be completed.

The first set includes those that satisfy the following criteria:

- One-school SAU
- Two-school SAU
- Multi-school SAU

The second set of samples includes SAUs or schools that have unique reporting situations as indicated by the *Processing and Reporting Business Requirements* document. This second set is necessary to ensure that each rule is applied correctly. The second set includes the following criteria:

- Private school that receives a stipulated percentage of federal and/or state funds
- Small school that receives no school report
- Small SAU that receives no SAU report
- SAU that receives a report but with schools that are too small to receive a school report
- School with excluded (not tested) students
- School with homeschooled students

CHAPTER 6 CLASSICAL ITEM ANALYSIS

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

Both qualitative and quantitative analyses are conducted to ensure that MEA Science items meet these standards. Qualitative analyses are described in earlier chapters of this report; this chapter focuses on quantitative evaluations. Statistical evaluations are presented in four parts: (1) difficulty indices, (2) item-test correlations, (3) differential item functioning (DIF) statistics, and (4) dimensionality analyses. The item analyses presented here are based on the statewide administration of the MEA Science in spring 2018.

6.1 CLASSICAL DIFFICULTY AND DISCRIMINATION INDICES

All multiple-choice and constructed-response items are evaluated in terms of item difficulty according to standard classical test theory practices. Difficulty is defined as the average proportion of points achieved on an item and is measured by obtaining the average score on an item and dividing it by the maximum possible score for the item. Multiple-choice items are scored dichotomously (correct versus incorrect); thus, for these items, the difficulty index is simply the proportion of students who correctly answered the item. Constructed-response items are scored polytomously, meaning that a student can achieve a score of 0, 1, 2, 3, or 4. By computing the difficulty index as the average proportion of points achieved, the indices for the different item types are placed on a similar scale, ranging from 0.0 to 1.0 regardless of the item type. Although this index is traditionally described as a measure of difficulty, it is properly interpreted as an easiness index, because larger values indicate easier items. An index of 0.0 indicates that all students received no credit for the item, and an index of 1.0 indicates that all students received full credit for the item.

Items that are answered correctly by almost all students provide little information about differences in student abilities, but they do indicate knowledge or skills that have been mastered by most students. Similarly, items that are correctly answered by very few students provide little information about differences in student abilities, but they may indicate knowledge or skills that have not yet been mastered by most students. In general, to provide the best measurement, difficulty indices should range from near-chance performance (0.25 for four-option multiple-choice items or essentially zero for constructed-response items) to 0.90, with the majority of items generally falling between around 0.4 and 0.7. However, on a standards-referenced

assessment such as the MEA Science, it may be appropriate to include some items with very low or very high item difficulty values to ensure sufficient content coverage.

A desirable characteristic of an item is for higher-ability students to perform better on the item than lower-ability students do. The correlation between student performance on a single item and total test score is a commonly used measure of this characteristic of the item. Within classical test theory, the item-test correlation is referred to as the item's discrimination because it indicates the extent to which successful performance on an item discriminates between high and low scores on the test. For constructed-response items, the item discrimination index used was the Pearson product-moment correlation; for multiple-choice items, the corresponding statistic is commonly referred to as a point-biserial correlation. The theoretical range of these statistics is -1.0 to 1.0, with a typical observed range from 0.2 to 0.6.

Discrimination indices can be thought of as measures of how closely an item assesses the same knowledge and skills assessed by other items contributing to the criterion total score. That is, the discrimination index can be thought of as a measure of construct consistency.

A summary of the item difficulty and item discrimination statistics for each grade is presented in Table 6-1. Note that the statistics are presented for all items as well as by item type (multiple-choice and constructed-response). The mean difficulty and discrimination values shown in the table are within generally acceptable and expected ranges.

Table 6-1. 2018–19 MEA Science Part I: Summary of Item Difficulty and Discrimination Statistics by Grade

Grade	*Item Type*	*Number of Items*	p-*Value*		*Discrimination*	
			Mean	*Standard Deviation*	*Mean*	*Standard Deviation*
5	ALL	36	0.63	0.13	0.33	0.10
	MC	32	0.65	0.12	0.31	0.09
	CR	4	0.46	0.09	0.48	0.07
8	ALL	44	0.63	0.12	0.36	0.08
	MC	40	0.65	0.12	0.35	0.07
	CR	4	0.46	0.04	0.51	0.02
11	ALL	44	0.54	0.14	0.38	0.10
	MC	40	0.57	0.12	0.36	0.07
	CR	4	0.29	0.05	0.59	0.01

A comparison of indices across grade levels is complicated because these indices are population dependent. Direct comparisons would require that either the items or the students were common across groups. Since that is not the case, it cannot be determined whether differences in performance across grade levels are because of differences in student abilities, differences in item difficulties, or both. With this caveat in mind, it appears that generally students in grade 11 found their items somewhat more difficult than students in grades 5 and 8.

Comparing the difficulty indices of multiple-choice items and constructed-response items is inappropriate because multiple-choice items can be answered correctly by guessing. Thus, it is not surprising that the difficulty indices for multiple-choice items tend to be higher (indicating that students performed better on these items) than the difficulty indices for constructed-response items. Similarly, discrimination indices for the four-point constructed-response items were larger than those for the dichotomous items because of the greater variability of the former (i.e., the partial credit these items allow) and the tendency for correlation coefficients to be higher given greater variances of the correlates.

In addition to the item difficulty and discrimination summaries presented above, item-level classical statistics and item-level score point distributions were also calculated. Item-level classical statistics are provided in Appendix H; item difficulty and discrimination values are presented for each item. The item difficulty and discrimination indices are within generally acceptable and expected ranges. Very few items were answered correctly at near-chance or near-perfect rates. Similarly, the positive discrimination indices indicate that students who performed well on individual items tended to perform well overall. There was a small number of items with low discrimination indices, but none was negative. While it is not inappropriate to include items with low discrimination values or with very high or very low item difficulty values to ensure that content is appropriately covered, there were very few such cases on the MEA Science. Item-level score point distributions are provided for constructed-response items in Appendix H; for each item, the percentage of students who received each score point is presented.

6.2 Differential Item Functioning

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

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

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

subgroup differences in performance could be traced to differential experience (such as geographical living conditions or access to technology), the inclusion of such items should be reconsidered.

Computed DIF indices have a theoretical range from -1.0 to 1.0 for multiple-choice items, and the index is adjusted to the same scale for constructed-response items. Dorans and Holland (1993) suggested that index values between -0.05 and 0.05 should be considered negligible. The preponderance of MEA Science items fell within this range. Dorans and Holland further stated that items with values between -0.10 and -0.05 and between 0.05 and 0.10 (i.e., "low" DIF) should be inspected to ensure that no possible effect is overlooked and that items with values outside the [-0.10, 0.10] range (i.e., "high" DIF) are more unusual and should be examined very carefully.[1]

For the 2018–19 MEA Science, the following subgroup comparisons were evaluated for DIF:

- Male versus Female
- No Disability versus Disability
- Not Economically Disadvantaged versus Economically Disadvantaged
- Non-LEP versus LEP
- White (non-Hispanic) versus Black or African American or Hispanic

The tables in Appendix I present the number of items classified as either "low" or "high" DIF, overall and by group favored.

6.3 Dimensionality Analyses for Maine Science

The MEA science tests were each designed to measure and report a single score on science achievement using a unidimensional scale. Thus, each of these tests is said to be measuring a single dimension, and the term "unidimensionality" is used to describe it.

Because each test is constructed with multiple content area subcategories and item types, and their associated knowledge and skills, the subtests associated with each of these could potentially result in a large number of secondary dimensions being invoked beyond the primary dimension that all the items on a test have in common. Generally, the scores on such subtests are highly correlated with each other; therefore, the primary dimension they share typically explains an overwhelming majority of variance in test scores. In fact, the presence of just such a dominant primary dimension is the psychometric assumption that provides the foundation for the unidimensional IRT models that were used for calibrating, linking, scaling, and equating the 2018-19 MEA science test forms.

The purpose of dimensionality analysis is to investigate whether violation of the assumption of test unidimensionality is statistically detectable and, if so, (a) the degree to which unidimensionality is violated

[1] It should be noted that DIF for items is evaluated initially at the time of field-testing. If an item displays high DIF, it is flagged for review by a Cognia content specialist.

and (b) the nature of the multidimensionality. Findings from dimensionality analyses performed on the 2018-19 MEA science common items for grades 5, 8, and 11 are reported below. (Note: only common items were analyzed since they are used for score reporting.)

Dimensionality analyses were conducted using the nonparametric IRT-based methods DIMTEST (Stout, 1987; Stout, Froelich, & Gao, 2001) and DETECT (Zhang & Stout, 1999). Nonparametric techniques were preferred for this analysis because such techniques avoid strong parametric modeling assumptions while still adhering to the fundamental principles of item response theory. Parametric techniques, such as nonlinear factor analysis, make strong assumptions that are often inappropriate for real data, such as assuming a normal distribution for ability and lower asymptotes of zero for the item characteristic curves.

Both DIMTEST and DETECT use as their basic statistical building block the estimated average conditional covariances for item pairs. A conditional covariance is the covariance between two items conditioned on expected total score for the rest of the test, and the average conditional covariance is obtained by averaging over all possible conditioning scores. When a test is strictly unidimensional, all conditional covariances are expected to take on values within random noise of zero, indicating statistically independent item responses for examinees with equal expected scores. Non-zero conditional covariances are essentially violations of the principle of local independence, and local *dependence* implies multidimensionality. Thus, non-random patterns of positive and negative conditional covariances are indicative of multidimensionality. In particular, when multiple dimensions are present, items measuring the same dimension will have positive conditional covariances with each other, whereas items measuring different dimensions will have negative conditional covariances with each other. For example, if multiple-choice (MC) items measure a different dimension from open-response (OR) items, we would expect MC items to have positive conditional covariances with each other, OR items to have positive conditional covariances with each other, and MC items to have negative conditional covariances with OR items.

DIMTEST is a hypothesis-testing procedure for detecting violations of local independence. For the exploratory analyses conducted for the MEA science tests, the data were first divided into a training sample and a cross-validation sample. Then an analysis of the conditional covariances was conducted on the training sample data to find the cluster of items that displays the greatest evidence of local dependence. The cross-validation sample was then used to test whether the conditional covariances of the selected cluster of items displays local dependence, conditioning on total score on the non-clustered items. The DIMTEST statistic follows a standard normal distribution under the null hypothesis of unidimensionality.

DETECT is an effect-size measure of multidimensionality. For the exploratory analyses conducted for the MEA science tests, as with DIMTEST, the data were first randomly divided into a training sample and a cross-validation sample. (Note: The training and cross-validation samples used for the DETECT analyses were randomly drawn independently of the samples used for the DIMTEST analyses.) The training sample was then used to find a set of mutually exclusive and collectively exhaustive clusters of items that best fit a systematic pattern of positive conditional covariances for pairs of items from the same cluster and negative

conditional covariances from different clusters. Next, the clusters from the training sample were used with the cross-validation sample data to average the conditional covariances: within-cluster conditional covariances were summed, from this sum the between-cluster conditional covariances were subtracted, this difference was divided by the total number of item pairs, and this average was multiplied by 100 to yield an index of the average violation of local independence for an item pair. DETECT values less than 0.2 indicate very weak multidimensionality (or near unidimensionality), values of 0.2 to 0.4 weak to moderate multidimensionality; values of 0.4 to 1.0 moderate to strong multidimensionality, and values greater than 1.0 very strong multidimensionality (Roussos & Ozbek, 2006).

As mentioned above, in applying DIMTEST and DETECT to the 2018-19 MEA science tests, the data for each grade were split into a training sample and a cross-validation sample. Each grade had at least 12,000 student examinees, so the training sample and cross-validation sample for each grade had at least 6,000 students. DIMTEST was then applied to each grade. DETECT was applied to each dataset for which the DIMTEST null hypothesis was rejected in order to estimate the effect size of the multidimensionality.

Because of the large sample sizes of both MEA science tests, DIMTEST would be sensitive to even quite small violations of unidimensionality, and the null hypothesis was rejected at level 0.01 for the datasets. These results were not surprising because strict unidimensionality is an idealization that almost never holds exactly for a given dataset. Thus, it was important to use DETECT to estimate the effect size of the violations of local independence found by DIMTEST. Table 1 displays the multidimensional effect size estimates from DETECT.

As shown in Table 1, all of the DETECT values indicate very weak (DETECT values less than or equal to 0.20) multidimensionality for every test analyzed. Also shown in Table 1 are the DETECT values from the previous two years' dimensionality analysis. This year's results are seen to be very similar to previous years' in that they indicated very weak multidimensionality for the tests.

Table 6-2. 2018–19 MEA Science Part I: Multidimensionality Effect Sizes

Grade	*Multidimensionality Effect Size*		
	2016-17	*2017-18*	*2018-19*
5	0.20	0.17	0.19
8	0.17	0.12	0.16
11	0.14	0.14	0.11

We also investigated how DETECT divided the tests into clusters to see if there were any discernible patterns with respect to the multiple-choice (MC) and open-response (OR) item types. Focusing only on MC-OR separation that occurs in both the DETECT clusters and in the pattern of the signs of the conditional

covariances (i.e., strong and clear separation), this year's analysis indicates no strong separation in either Grade 5, 8, or 11. This result agreed with the findings from most past years for the science tests.

The combined results of DETECT and DIMTEST indicate that multidimensionality is present in MEA science tests but the magnitude is small in all cases. Investigating the patterns of signs in the conditional covariance matrices, there were no indications of strong non-random patterns. We also investigated the DETECT clusters with respect to MC-OR separation, and found no strong evidence of MC-OR separation in either Grade 5, 8, or 11. A more thorough investigation employing experts in the substantive content of the test forms may result in identification of clusters related to the skills and knowledge areas measured by the items. In any case the violations of local independence from all such effects, as evidenced by the DETECT effect sizes, were very small and do not warrant any changes in test design or scoring.

CHAPTER 7 ITEM RESPONSE THEORY SCALING AND EQUATING

This chapter describes the procedures used to calibrate, equate, and scale the MEA Science test. During the course of these psychometric analyses, a number of quality-control procedures and checks on the processes were implemented. These procedures included evaluation of the calibration processes (e.g., checking the number of Newton cycles required for convergence for reasonableness, checking item parameters and their standard errors for reasonableness, and examining test characteristic curves [TCCs] and test information functions [TIFs] for reasonableness), evaluation of model fit, evaluation of equating items (e.g., delta analyses, rescore analyses, and examination of *a*-plots and *b*-plots for reasonableness), and evaluation of the scaling results (e.g., parallel processing by the Data and Reporting Services [DRS] and Psychometrics and Research Departments and comparison of look-up tables to the previous year's look-up tables). An equating report, which provided complete documentation of the quality-control procedures and results, was submitted to the Maine DOE for its approval prior to production of student reports.

Table 7-1 lists items that required intervention either during item calibration or as a result of the evaluations of the equating items. For each flagged item, the table shows the reason it was flagged and what action was taken. Descriptions of the evaluations and results are included in the Item Response Theory Results and Equating Results sections found later in this chapter.

Table 7-1. 2018–19 MEA Science: Items That Required Intervention During IRT Calibration and Equating

Grade	*Item*	*Reason*	*Action*
5	138289A	c-parameter	set c = 0
	248566	c-parameter	set c = 0
	96271	delta analysis	removed from equating
8	153898	c-parameter	set c = 0
	153898	b/b analysis	removed from equating
	191064A	delta analysis	removed from equating
11	166387A	c-parameter	set c = 0
	191140A	c-parameter	set c = 0
	191140A	delta analysis	removed from equating
	191147A	c-parameter	set c = 0
	191163A	c-parameter	set c = 0
	550251	c-parameter	set c = 0

7.1 ITEM RESPONSE THEORY

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

For the 2018–19 MEA Science, the three-parameter logistic (3PL) model was used for dichotomous items and the graded response model (GRM) was used for polytomous items. The 3PL model for dichotomous items can be defined as

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

where
i indexes the items,
j indexes students,
a represents item discrimination,
b represents item difficulty,
c is the pseudo guessing parameter, and
D is a normalizing constant equal to 1.701.

In the GRM for polytomous items, an item is scored in $k + 1$ graded categories that can be viewed as a set of k dichotomies. At each point of dichotomization (i.e., at each threshold), a two-parameter model can be used. This implies that a polytomous item with $k + 1$ categories can be characterized by k item category threshold curves (ICTC) of the two-parameter logistic form

$$P_{ik}^*(\theta_j) = \frac{\exp[Da_i(\theta_j - b_i + d_{ik})]}{1+\exp[Da_i(\theta_j - b_i + d_{ik})]},$$

where
i indexes the items,
j indexes students,
k indexes threshold,
a represents item discrimination,
b represents item difficulty,
d represents threshold, and

D is a normalizing constant equal to 1.701.

After computing k ICTCs in the GRM, $k+1$ item category characteristic curves (ICCCs) are derived by subtracting adjacent ICTCs:

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

where
P_{ik} represents the probability that the score on item i falls in category k, and
P^*_{ik} represents the probability that the score on item i falls above the threshold k
($P^*_{i0} = 1$ and $P^*_{i(m+1)} = 0$).

The GRM is also commonly expressed as

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

Finally, the item characteristic curve (ICC) for polytomous items is computed as a weighted sum of ICCCs, where each ICCC is weighted by a score assigned to a corresponding category:

$$P_i(\theta_j) = \sum_k^{m+1} w_{ik} P_{ik}(\theta_j).$$

For more information about item calibration and determination, the reader is referred to Lord and Novick (1968), Hambleton and Swaminathan (1985), or Baker and Kim (2004).

7.2 Item Response Theory Results

The tables in Appendix J give the IRT item parameters of all common items on the 2018–19 MEA Science tests by grade. In addition, Appendix K shows graphs of the TCCs and TIFs, which are defined below.

TCCs display the expected (average) raw score associated with each θ_j value between -4.0 and 4.0. Mathematically, the TCC is computed by summing the ICCs of all items that contribute to the raw score. Using the notation introduced in Section 7.1, the expected raw score at a given value of θ_j is

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

where
i indexes the items (and n is the number of items contributing to the raw score),
j indexes students (here, θ_j runs from -4.0 to 4.0), and
$E(X|\theta_j)$ is the expected raw score for a student of ability θ_j.

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

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

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

Compared to the tails, TIFs are often higher near the middle of the distribution where most students are located and where most items are sensitive by design.

Table 7-1, presented at the beginning of this chapter, shows that a few items were flagged based on the quality-control checks implemented during the calibration process on each test. Two items in grade 5, one item in grade 8, and five items in grade 11 were flagged because the guessing parameter (*c*-parameter) was poorly estimated. Difficulty in estimating the *c*-parameter is not at all unusual and is well documented in psychometric literature, especially when an item's discrimination is below 0.50. Fixing the *c*-parameter to zero resulted in reasonable and stable item parameter estimates and improved model fit in some cases. One item in grade 5 was flagged because the delta analysis of item performance showed statistical evidence of having changed significantly from previous usage. One item in grade 8 was flagged because the delta analysis and b/b analysis of item performance showed statistical evidence of having changed significantly from previous usage. One item in grade 11 was flagged because the delta analysis of item performance showed statistical evidence of having changed significantly from previous usage. These items were removed from equating.

The number of Newton cycles required for convergence for each grade during the IRT analysis can be found in Table 7-2. For all grades, the number of cycles required fell within acceptable ranges.

Table 7-2. 2018–19 MEA Science: Number of Newton Cycles Required for Convergence

Grade	*Cycles*
5	28
8	31
11	60

7.3 EQUATING

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

The 2018–19 administration of MEA Science used a raw-score-to-theta equating procedure in which test forms were equated to the theta scale established on the reference form (i.e., the form used in the most recent standard setting). This is accomplished through the chained linking design, in which every new form is equated back to the theta scale of the previous year's test form. It can therefore be assumed that the theta scale of every new test form is the same as the theta scale of the reference form, since this is where the chain originated.

The groups of students who took the equating items on the 2018–19 MEA Science tests are not equivalent to the groups who took them in the reference year. IRT is particularly useful for equating scenarios that involve nonequivalent groups (Allen & Yen, 1979). Equating for MEA Science uses the anchor-test-nonequivalent-groups design described by Petersen, Kolen, and Hoover (1989). In this equating design, no assumption is made about the equivalence of the examinee groups taking different test forms (i.e., naturally occurring groups are assumed). Comparability is instead evaluated by utilizing a set of anchor items (also called equating items). However, the equating items are designed to mirror the common test in terms of item types and distribution of emphasis.

Item parameter estimates for 2018–19 were placed on the base-year scale by using the method of Stocking and Lord (1983), which is based on the IRT principle of item parameter invariance. According to this principle, the equating items for both the 2015–16 and the 2018–19 MEA Science tests should have the same item parameters. After the item parameters for each 2018–19 test were estimated using PARSCALE (Muraki & Bock, 2003), the Stocking and Lord method was employed to find the linear transformation (slope and intercept) that adjusted the equating items' parameter estimates such that the 2018–19 TCC was as close as possible to that of 2015–16.

A standard validation was conducted for grade 11 science in 2015. This process was described in the *Maine Educational Assessment High School Science Standard Validation Report* (August 6, 2015).

7.4 EQUATING RESULTS

Prior to calculating the Stocking and Lord (1983) transformation constants, a variety of evaluations of the equating items were conducted. Two items were flagged for evaluation as a result of these procedures.

Appendix L presents the results from the delta analysis. This procedure was used to evaluate adequacy of equating items; one item in each of grades 5, 8, and 11 had "true" in the discard status, indicating these items were flagged as inappropriate for use in equating.

Also presented in Appendix L are the results from the rescore analysis. With this analysis, 200 random papers from the previous year were interspersed with this year's papers to evaluate scorer consistency from one year to the next. All effect sizes were below the criterion value for excluding an item as an equating item: 0.50 in absolute value.

Finally, a-plots and b-plots, which show IRT parameters for 2018–19 plotted against their previous values, are presented in Appendix M. One item in grades 8 was identified as an outlier in the b/b plots and was removed as an equating item. This grade 8 item was also flagged in the delta analysis.

Once all evaluations of the equating items were complete, the Stocking and Lord (1983) method of equating was used to place the item parameters onto the previous year's scale, as described previously. The Stocking and Lord transformation constants are presented in Table 7-3.

Table 7-3. 2018–19 MEA Science: Stocking and Lord (1983) Transformation Constants

Grade	*α-slope*	*β-intercept*
5	1.06	0.10
8	1.07	0.14
11	0.92	0.11

The next administration of MEA Science (2019–20) will be scaled to the 2018–19 administration using the same equating method described above.

7.5 ACHIEVEMENT STANDARDS

MEA Science standards achievement-level cut scores in grades 5 and 8 were set in May 2009. The standard setting meeting and results were discussed in the 2009 technical report and standard setting report provided at that time. As alluded to in the discussion of equating in this chapter, the scale was established during that base year, and the forms serve as the reference for subsequent equating. The θ metric cut scores that emerged from the standard setting meeting will remain fixed throughout the assessment program unless standards are reset for any reason.

In addition, MEA Science standards achievement-level cut scores in grade 11 were validated in August 2015. The need for standard validation arose from the fact that MEA Science had a change in scoring rules in 2015. This change in scoring rules was reflected in a change in student behavior leading to a decrease in item nonresponse rates. The primary goal of the standard validation was to reconfirm the knowledge, skills, and abilities (KSAs) that are necessary for students to demonstrate in order to be classified into each of the achievement levels.

A standard validation process was implemented for the MEA Science rather than a full standard setting. Neither the assessment nor the related standards have changed since previous administrations. However, the change in scoring rules did prompt a noticeable change in nonresponse rates. The mean nonresponse rate dropped from 4.1% in 2014 to 1.4% in 2015, which represented a potentially significant change in student behavior on the assessment. Consequently, existing standards needed to be validated to account for the recent changes in the scoring rules for the high school science assessment. The θ metric cut scores that emerged from the standard validation meeting will remain fixed throughout the assessment program unless standards are reset for any reason. This process was described in the *Maine Educational Assessment High School Science Standard Validation Report* (August 6, 2015).

7.6 Reported Scaled Scores

7.6.1 Description of Scale

Because the θ scale used in IRT calibrations is not readily understood by most stakeholders, reporting scales were developed for the MEA Science tests. The reporting scale is a simple linear transformation of the underlying θ scale used in the IRT calibrations. The scales were developed such that the scaled scores for grade 5 range from 500 to 580, for grade 8 from 800 through 880, and for grade 11 from 1100 through 1180. Likewise, for each grade level, the Below State Expectations/At State Expectations and the At State Expectations/Above State Expectations cuts were placed at "*x*42" and "*x*62," respectively, where *x* denotes the grade level. (At the student level, scaled scores were reported as even numbers only.)

By providing information that is more specific about the position of a student's results, scaled scores supplement achievement-level scores. School- and SAU-level scaled scores are calculated by computing the average of student-level scaled scores. Students' raw scores (i.e., total number of points) on the MEA Science tests were translated to scaled scores using the data analytic process known as *scaling*. Scaling simply converts from one scale to another. In the same way that a given temperature can be expressed on either Fahrenheit or Celsius scales, or the same distance can be expressed in either miles or kilometers, student scores on the MEA Science tests can be expressed in raw or scaled scores.

It is important to note that converting from raw scores to scaled scores does not change students' achievement-level classifications. Given the relative simplicity of raw scores, it is fair to question why scaled scores for MEA are reported instead of raw scores. Scaled scores make the reporting of results consistent. To illustrate, standard setting typically results in different *raw* cut scores across grades. The raw cut score between Below State Expectations and At State Expectations could be, for example, 35 in grade 5 but 33 in grade 8, yet both raw scores would be transformed to scaled scores of *x*42. It is this uniformity across *scaled scores* that facilitates the understanding of student performance. The psychometric advantage of scaled scores

over raw scores comes from there being linear transformations of θ. Since the θ scale is used for equating, scaled scores are comparable from one year to the next where raw scores are not.

7.6.2 Calculations

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

$$SS = \alpha\hat{\theta} + \beta,$$

where
m is the slope and
b is the intercept.

A separate linear transformation was used for each grade. Each line is determined by fixing the "x42" and "x62" values (e.g., 542 and 562, respectively, for grade 5). Because only two points within the θ scaled score space were fixed, the cut scores between Well Below State Expectations and Below State Expectations varied across grades. Table 7-4 presents the cuts on the theta scale and the reporting scale for each grade (i.e., the minimum scaled score for getting into the next achievement level). It is important to repeat that the values in Table 7-4 do not change from year to year, because the cut scores along the θ scale do not change unless standards are reset. Also, in a given year, it may not be possible to attain a particular scaled score, but the scaled score cuts will remain the same.

Table 7-4. 2018–19 MEA Science: Cut Scores on the Theta and Reporting Scales by Grade

Grade	*Θ Cuts*			*Minimum*	*Scaled Score Cuts*			*Maximum*
	WBSE/BSE	*BSE/ASE*	*ASE/Above SE*		*WBSE/BSE*	*BSE/ASE*	*ASE/Above SE*	
5	-1.275	-0.1229	1.7747	500	530	542	562	580
8	-1.3906	-0.3913	1.0164	800	828	842	862	880
11	-0.6779	0.2042	1.6887	1100	1130	1142	1162	1180

WBSE = Well Below State Expectations; BSE = Below State Expectations; ASE = At State Expectations, Above SE = Above State Expectations

Table 7-5 shows the slope and intercept terms used to calculate the scaled scores. Note that the values in Table 7-5 will not change unless the standards are reset.

Table 7-5. 2018–19 MEA Science: Scaled Score Transformation Constants by Grade

Grade	*α-slope*	*β -intercept*
5	10.53962901	543.2953204
8	14.20757264	847.5594232
11	13.47254968	1139.248905

Appendix N contains raw score to scaled score look-up tables for this year and last year for all grades. These are the actual tables that were used to determine student scaled scores, error bands, and achievement levels.

7.6.3 Distributions

Graphs of the scaled score cumulative frequency distributions for the last two years are presented in Appendix O. Note that the graphs show the percentage of students at or below each scaled score; thus, the lowest line in a given graph depicts the highest performing group. Appendix O also shows, in Table O-1, achievement-level distributions by grade in terms of percentages. Results are shown for the 2016–17, 2017–18, and 2018–19 administrations.

CHAPTER 8 RELIABILITY

Although an individual item's performance is an important focus for evaluation, a complete evaluation of an assessment must also address the way items function together and complement one another. Tests that function well provide a dependable assessment of the student's level of ability. Unfortunately, no test can do this perfectly. A variety of factors can contribute to a given student's score being either higher or lower than his or her true ability. For example, a student may misread an item or mistakenly fill in the wrong bubble when he or she knew the answer. Collectively, extraneous factors that impact a student's score are referred to as measurement error. Any assessment includes some amount of measurement error; that is, no measurement is perfect. This is true of all academic assessments—some students will receive scores that underestimate their true ability, and other students will receive scores that overestimate their true ability. When tests have a high amount of measurement error, student scores are very unstable. Students with high ability may get low scores or vice versa. Consequently, one cannot reliably measure a student's true level of ability with such a test. Assessments that have less measurement error (i.e., errors made are small on average and student scores on such a test will consistently represent their ability) are described as reliable.

There are a number of ways to estimate an assessment's reliability. One possible approach is to give the same test to the same students at two different points in time. If students receive the same scores on each test, then the extraneous factors affecting performance are small and the test is reliable. (This is referred to as "test-retest reliability.") A potential problem with this approach is that students may remember items from the first administration or may have gained (or lost) knowledge or skills in the interim between the two administrations. A solution to the "remembering items" problem is to give a different but parallel test at the second administration. If student scores on each test correlate highly, the test is considered reliable. (This is known as "alternate forms reliability" because an alternate form of the test is used in each administration.) This approach, however, does not address the problem that students may have gained (or lost) knowledge or skills in the interim between the two administrations. In addition, the practical challenges of developing and administering parallel forms generally preclude the use of parallel form reliability indices. One way to address the latter two problems is to split the test in half and then correlate students' scores on the two half-tests; this, in effect, treats each half-test as a complete test. By doing this, the problems associated with an intervening time interval and with creating and administering two parallel forms of the test are alleviated. This is known as a "split-half estimate of reliability." If the two half-test scores correlate highly, items on the two half-tests must be measuring very similar knowledge or skills. This is evidence that the items complement one another and function well as a group. This also suggests that measurement error will be minimal.

The split-half method requires psychometricians to select items that contribute to each half-test score. This decision may have an impact on the resulting correlation, since each different possible split of the test into halves will result in a different correlation. Another problem with the split-half method of calculating

reliability is that it underestimates reliability, because test length is cut in half. All else being equal, a shorter test is less reliable than a longer test. Cronbach (1951) provided a statistic, α (alpha), that eliminates the problem of the split-half method by comparing individual item variances to total test variance. Cronbach's α was used to assess the reliability of the 2018–19 MEA Science:

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

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

8.1 Reliability and Standard Errors of Measurement

Table 8-1 presents descriptive statistics, Cronbach's α coefficient, and raw score standard errors of measurement (SEMs) for each grade.

Table 8-1. 2018–19 MEA Science: Raw Score Descriptive Statistics, Cronbach's Alpha, and Standard Errors of Measurement (SEM) by Grade

Grade	*Number of Students*	*Raw Score*			*Alpha*	*SEM*
		Maximum	*Mean*	*Standard Deviation*		
5	13,009	48	28.15	8.09	0.84	3.27
8	13,066	56	33.34	9.94	0.88	3.47
11	12,038	56	27.29	10.37	0.89	3.40

Since different grades have different test designs (e.g., the number of items varies by test), it is inappropriate to make inferences about the quality of one test by comparing its reliability to that of another test from a different grade.

8.2 Subgroup Reliability

The reliability coefficients discussed in the previous section were based on the overall population of students who took the 2018–19 MEA Science tests. Appendix P presents reliabilities for various subgroups of interest. Subgroup Cronbach's α's were calculated using the formula defined above based only on the members of the subgroup in question in the computations; values are calculated only for subgroups with 10 or more students.

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

8.3 SUBCATEGORY RELIABILITY

Of even more interest are reliabilities for the reporting subcategories within MEA Science. Cronbach's α coefficients for subcategories were calculated via the same formula defined previously using just the items of a given subcategory in the computations. Results are presented in Appendix P. Once again, as expected, because they are based on a subset of items rather than the full test, computed subcategory reliabilities were lower (sometimes substantially so) than were overall test reliabilities, and interpretations should take this into account. The subcategory reliabilities were lower than those based on the total test and approximately to the degree one would expect based on classical test theory. Qualitative differences between subtests once again preclude valid inferences about the quality of the full test based on statistical comparisons among subtests.

8.4 RELIABILITY OF ACHIEVEMENT-LEVEL CATEGORIZATION

While related to reliability, the accuracy and consistency of classifying students into achievement categories are even more important statistics in a standards-based reporting framework (Livingston & Lewis, 1995). After the achievement levels were specified and students were classified into those levels, empirical analyses were conducted to determine the statistical accuracy and consistency of the classifications. For the MEA Science tests, students are classified into one of four achievement levels: Well Below State Expectations, Below State Expectations, At State Expectations, or Above State Expectations. This section of the report explains the methodologies used to assess the reliability of classification decisions, and results are given.

Accuracy refers to the extent to which decisions based on test scores match decisions that would have been made if the scores did not contain any measurement error. Accuracy must be estimated, because errorless test scores do not exist. Consistency measures the extent to which classification decisions based on test scores match the decisions based on scores from a second, parallel form of the same test. Consistency can be evaluated directly from actual responses to test items if two complete and parallel forms of the test are

given to the same group of students. In operational test programs, however, such a design is usually impractical. Instead, techniques have been developed to estimate both the accuracy and the consistency of classification decisions based on a single administration of a test. The Livingston and Lewis (1995) technique was used for the 2018–19 MEA Science tests because it is easily adaptable to all types of testing formats, including mixed-format tests.

The accuracy and consistency estimates reported in Appendix Q make use of "true scores" in the classical test theory sense. A true score is the score that would be obtained if a test had no measurement error. Of course, true scores cannot be observed and so must be estimated. In the Livingston and Lewis (1995) method, estimated true scores are used to categorize students into their "true" classifications.

For the 2018–19 MEA Science tests, after various technical adjustments (described in Livingston & Lewis, 1995), a four-by-four contingency table of accuracy was created for each grade, where cell [*i, j*] represented the estimated proportion of students whose true score fell into classification *i* (where *i* = 1 to 4) and observed score into classification *j* (where *j* = 1 to 4). The sum of the diagonal entries (i.e., the proportion of students whose true and observed classifications matched) signified overall accuracy.

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

Another way to measure consistency is to use Cohen's (1960) coefficient κ (kappa), which assesses the proportion of consistent classifications after removing the proportion of consistent classifications that would be expected by chance. It is calculated using the following formula:

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

where

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

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

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

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

8.4.1 Accuracy and Consistency

The accuracy and consistency analyses described on the previous page are provided in Table Q-1 of Appendix Q. The table includes overall accuracy and consistency indices, including kappa. Accuracy and consistency values conditional on achievement level are also given. For these calculations, the denominator is the proportion of students associated with a given achievement level. For example, if hypothetically the conditional accuracy value is 0.78 for Well Below State Expectations for grade 5, then this figure indicates that among the students whose true scores placed them in this classification, 78% would be expected to be in this classification when categorized according to their observed scores. Similarly, a hypothetical consistency value of 0.68 indicates that 68% of students with observed scores in the Well Below State Expectations level would be expected to score in this classification again if a second, parallel test form were used.

For some testing situations, the greatest concern may be decisions around cut scores. For example, in testing done for Every Student Succeeds Act (ESSA) accountability purposes, the primary concern is distinguishing between students who are at state expectations and those who are not yet at state expectations. In this case, the accuracy of the Below State Expectations/At State Expectations threshold is of greatest interest. Table Q-2 in Appendix Q provides accuracy and consistency estimates at each cut, as well as false positive and false negative decision rates for the 2018–19 MEA Science tests. (A false positive is the proportion of students whose observed scores were above the cut and whose true scores were below the cut. A false negative is the proportion of students whose observed scores were below the cut and whose true scores were above the cut.)

Note that, as with other methods of evaluating reliability, decision accuracy and consistency (DAC) statistics calculated based on small groups can be expected to be lower than those calculated based on larger groups. For this reason, the values presented in Appendix Q should be interpreted with caution. In addition, it is important to remember that it is inappropriate to compare DAC statistics between grades.

8.5 INTERRATER CONSISTENCY

Chapter 4 of this report describes in detail the processes that were implemented to monitor the quality of the hand-scoring of student responses for constructed-response items. One of these processes was double-blind scoring: Approximately 10% of student responses were randomly selected and scored independently by two different scorers. Results of the double-blind scoring were used during the scoring process to identify scorers who required retraining or other intervention and are presented here as evidence of the reliability of the MEA Science tests. A summary of the interrater consistency results is presented in Table 8-2. Results in the table are collapsed across the hand-scored items by grade. This same information is provided at the item level in Appendix G.

Table 8-2. 2018–19 MEA Science: Summary of Interrater Consistency Statistics Collapsed Across Items by Grade

Grade	*Number of*		*Percent*		*Correlation*	*Percent of Third Scores*
	Score Categories	*Included Scores*	*Exact*	*Adjacent*		
5	5	5,096	60.46	33.69	0.79	5.83
8	5	5,172	72.35	24.07	0.85	3.58
11	5	4,455	66.24	30.35	0.80	3.43

CHAPTER 9 VALIDITY

Because interpretations of test scores, and not a test itself, are evaluated for validity, the purpose of this report is to describe several technical aspects of the MEA Science tests in support of score interpretations (AERA et al., 2014). Each chapter contributes an important component in the investigation of score validation: test development and design; test administration; scoring, scaling, and equating; item analyses; reliability; and score reporting.

Standards for Educational and Psychological Testing (AERA, et al., 2014) provides a framework for describing sources of evidence that should be considered when constructing a validity argument. The evidence around test content, response processes, internal structure, relationship to other variables, and consequences of testing each speak to different aspects of validity, but those aspects are not in themselves distinct *types* of validity. Instead, each contributes to a body of evidence about the comprehensive validity of score interpretations.

Evidence on test content validity is meant to determine how well the assessment tasks represent the curriculum and standards for each content area and grade level. Content validation is informed by the item development process, including how the test blueprints and test items align to the curriculum and standards. Viewed through the lens provided by the standards, evidence based on test content was extensively described in Chapters 2 and 3. Item alignment with MEA Science content standards; item bias, sensitivity, and content appropriateness review processes; adherence to the test blueprint; use of multiple item types; use of standardized administration procedures, with accommodated options for participation; and appropriate test administration training are all components of validity evidence based on test content. As discussed earlier, all questions are aligned by MEA Science educators to the *2007 Maine Learning Results* (MLRs) and undergo several rounds of review for content fidelity and appropriateness. Items are presented to students in multiple formats (constructed-response and multiple-choice). Finally, tests are administered according to state-mandated standardized procedures with allowable accommodations.

The scoring information in Chapter 4 describes the steps taken to train and monitor hand-scorers, as well as quality-control procedures related to scanning and machine scoring. To speak to student response processes, however, additional studies would be helpful and might include an investigation of students' cognitive methods using think-aloud protocols.

Evidence based on internal structure is presented in Chapters 6 through 8 in great detail in the discussions of item analyses, scaling and equating, and reliability. Technical characteristics of the internal structure of the assessments are presented in terms of classical item statistics (item difficulty, item-test correlation), differential item functioning (DIF) analyses, dimensionality analyses, reliability, standard errors of measurement (SEM), and item response theory (IRT) parameters and procedures. Each test is equated to the same grade and content-area test from the prior year to preserve the meaning of scores over time. In

general, item difficulty and discrimination indices were in acceptable and expected ranges. Very few items were answered correctly at near-chance or near-perfect rates. Similarly, the positive discrimination indices indicate that most items were assessing consistent constructs, and students who performed well on individual items tended to perform well overall.

Evidence based on the consequences of testing is addressed in the scaled score information in Chapter 7 and the quality assurance of data in Chapter 5. Each of these chapters speaks to the efforts undertaken to provide accurate and clear information to the public regarding test scores. Scaled scores offer the advantage of simplifying the reporting of results across grade levels and subsequent years. Achievement levels provide users with reference points for mastery at each grade level, which is another useful and simple way to interpret scores. Several different standard reports are provided to stakeholders. Additional evidence of the consequences of testing could be supplemented with broader investigation of the impact of testing on student learning.

9.1 QUESTIONNAIRE DATA

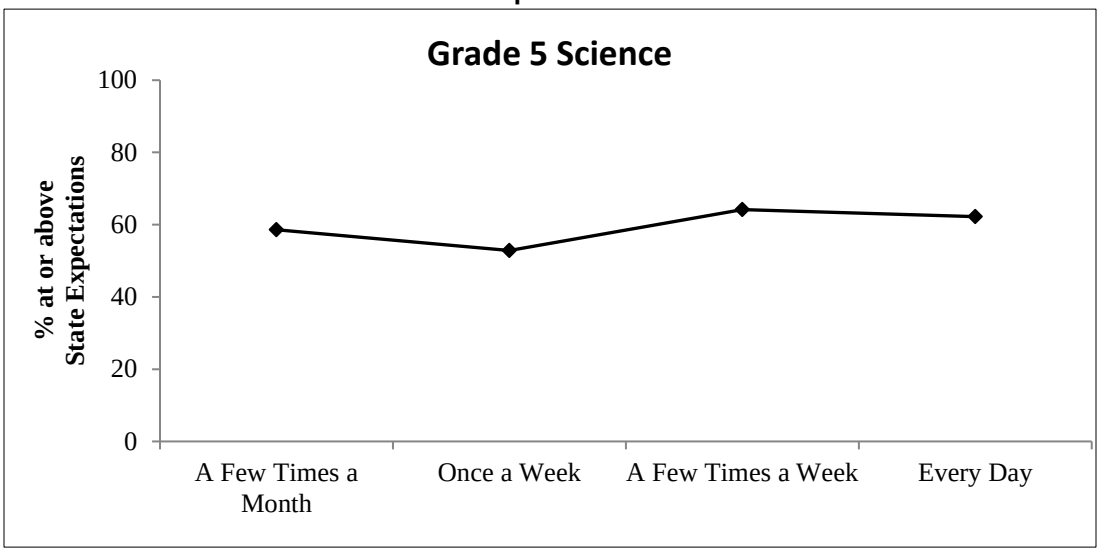
External validity of the 2018–19 MEA Science assessment is conveyed by the relationship of test scores and situational variables such as time spent patterns, self-image, and attitude toward content matter. These situational variables were all based on student questionnaire data collected during the administration of the 2018–19 MEA Science test. Note that no inferential statistics are included in the results presented on the following pages; however, because the numbers of students are quite large, differences in average scores may be statistically significant.

9.1.1 Time Spent

Examinees in grade 5 were asked how often they had science classes. Figure 9-1 shows that students who had science classes every day or a few times a week did better on the MEA Science than those who had science class once a week or a few times a month, but the difference was slight.

Question: How often do you do science in class?

Figure 9-1. 2018–19 MEA Science: Questionnaire Responses—Time Spent Grade 5



9.1.2 Self-Image

Examinees in grades 5, 8, and 11 were asked how they would rate themselves as a student in science. Figures 9-2 to 9-4 indicate a positive relationship between self-image as a student and MEA Science scores.

Question: Which of the following best describes how you rate yourself as a student in science?

Figure 9-2. 2018–19 MEA Science: Questionnaire Responses—Self-Image Grade 5

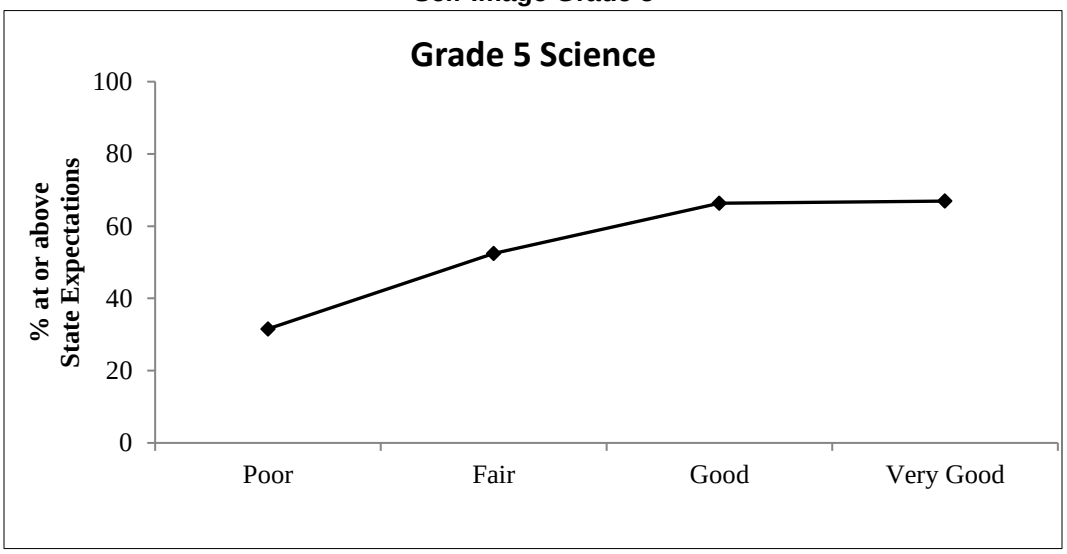


Figure 9-3. 2018–19 MEA Science: Questionnaire Responses—Self-Image Grade 8

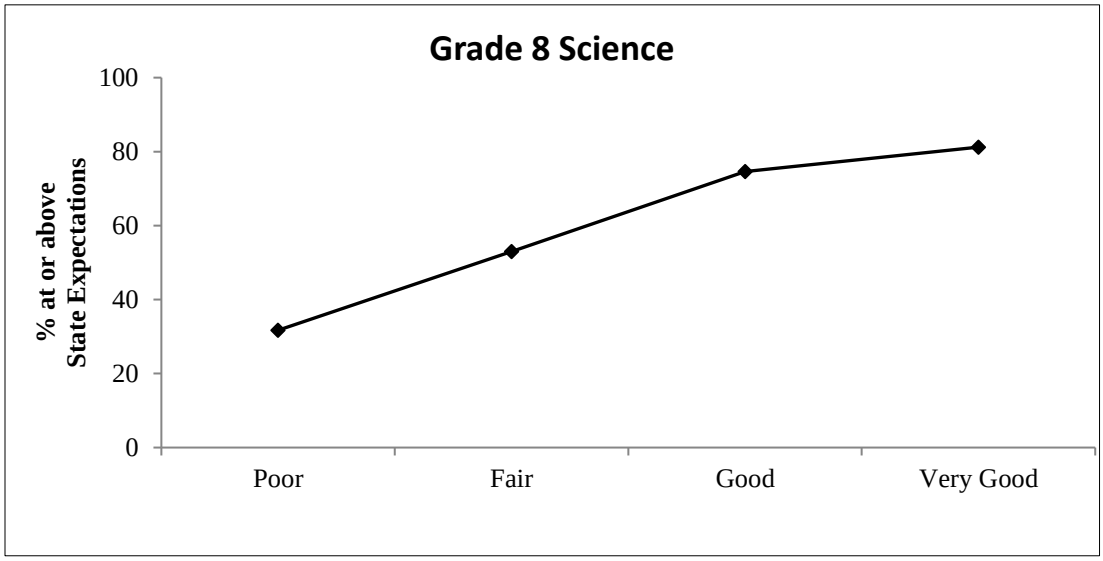
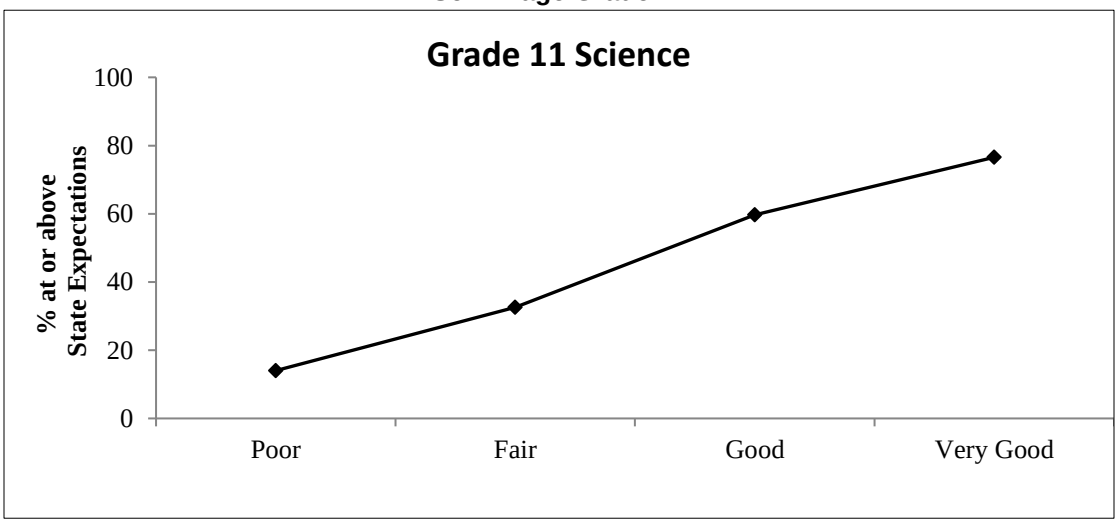


Figure 9-4. 2018–19 MEA Science: Questionnaire Responses—Self-Image Grade 11

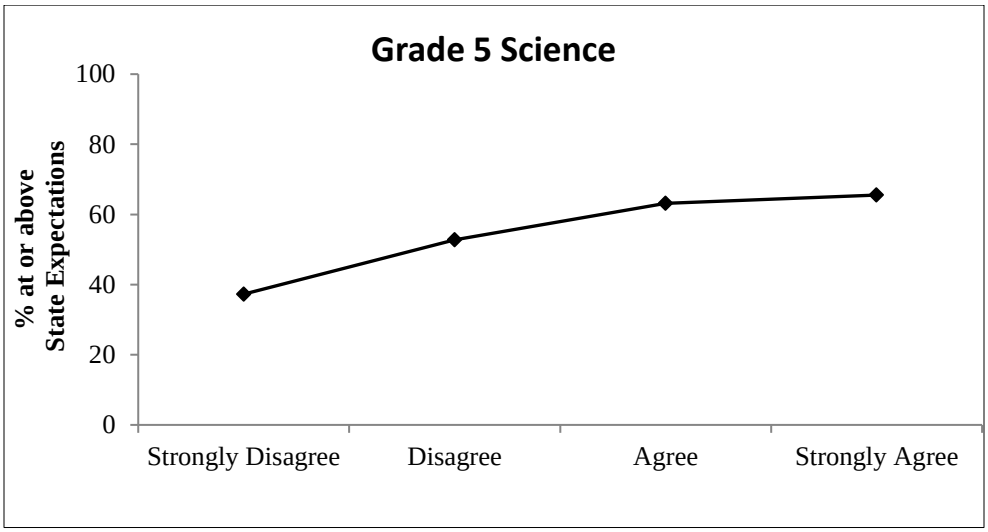


9.1.3 Attitude Toward Content Area

Questionnaire items related to examinees' attitudes toward science were administered to students in grades 5, 8, and 11. For grade 5, students were asked how they felt about the statement "Science is interesting and fun." For grades 8 and 11, students were asked how they felt about the statement "My knowledge of science will be useful to me as an adult." Figures 9-5 to 9-7 indicate that students' attitudes toward science are related positively to MEA Science scores.

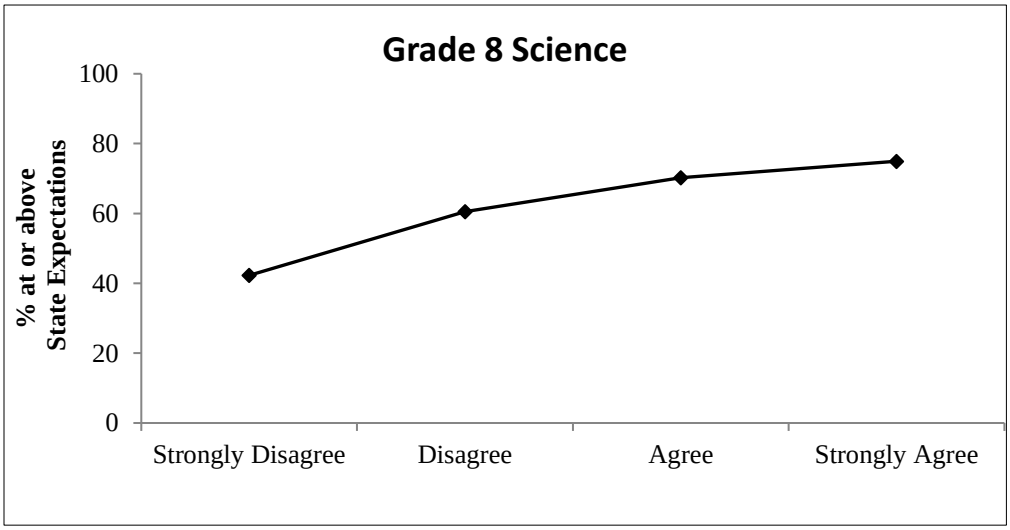
Question: How do you feel about the following statement? *Science is interesting and fun.*

Figure 9-5. 2018–19 MEA Science: Questionnaire Responses—Attitude Toward Content Grade 5



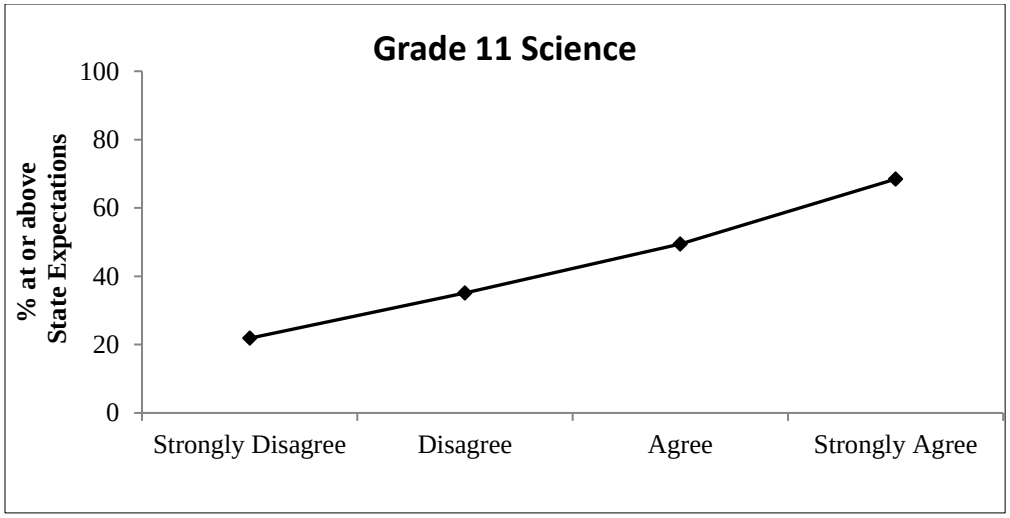
Question: How do you feel about the following statement? *My knowledge of science will be useful to me as an adult.*

Figure 9-6. 2018–19 MEA Science: Questionnaire Responses—Attitude Toward Content Grade 8



Question: How do you feel about the following statement? *My knowledge of science will be useful to me as an adult.*

Figure 9-7. 2018–19 MEA Science: Questionnaire Responses—Attitude Toward Content Grade 11



9.1.4 Match of Questions to What Is Learned in School

Students in grades 5, 8, and 11 were asked how well the questions on the MEA Science test matched what they had learned in school about science. Figures 9-8 to 9-10 indicate that there is a positive relationship between how well students feel the questions match what they have learned in science and MEA Science scores.

Question: How well do the questions that you have just been given on this MEA test match what you have learned in school about science?

Figure 9-8. 2018–19 MEA Science: Questionnaire Responses—Match of Questions to What Is Learned in School Grade 5

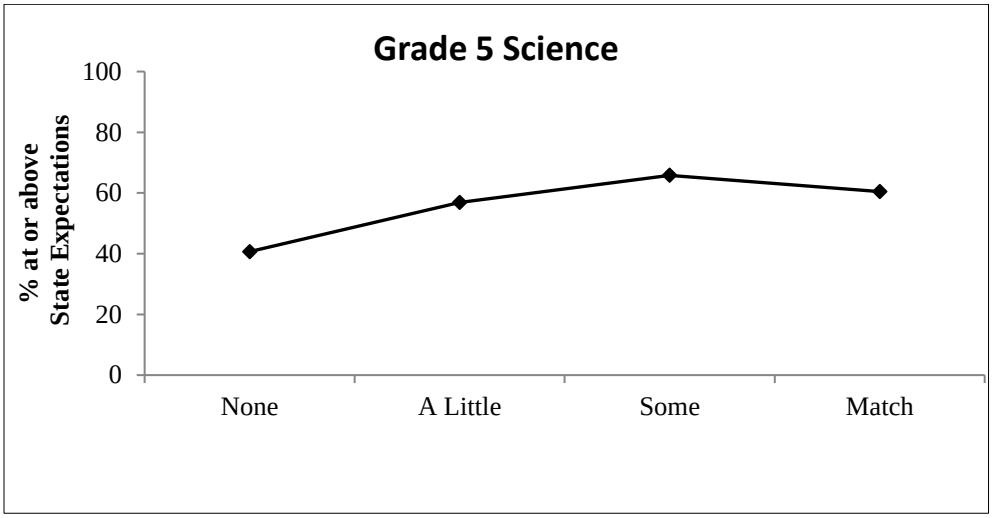


Figure 9-9. 2018–19 MEA Science: Questionnaire Responses—Match of Questions to What Is Learned in School Grade 8

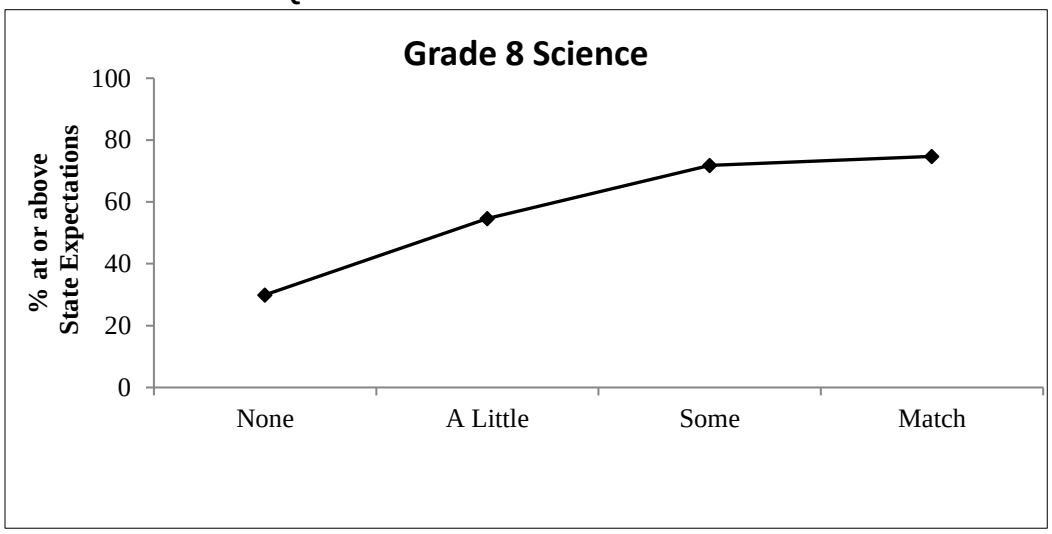
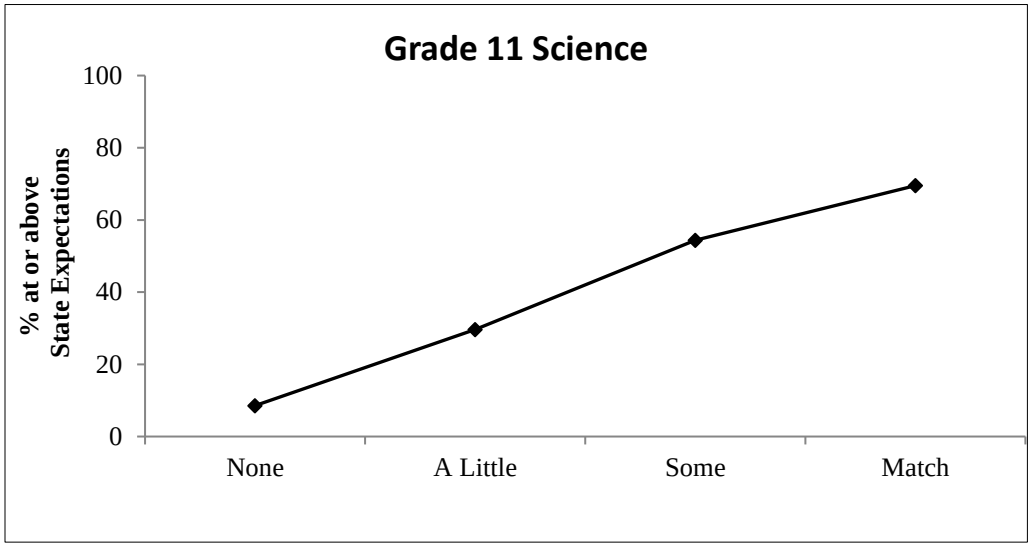


Figure 9-10. 2018–19 MEA Science: Questionnaire Responses—Match of Questions to What Is Learned in School Grade 11



9.1.5 Difficulty of Assessment

For grades 5, 8, and 11 students were asked how difficult they found the test. Figures 9-11 to 9-13 indicate that there is a modest negative relationship between how difficult the students felt the items were and overall MEA Science scores (i.e., students who found the test more difficult received lower scores than students who found the test easier) for grades 8 and 11. The pattern for grade 5 was relatively flat.

Question: How difficult was this science test?

Figure 9-11. 2018–19 MEA Science: Questionnaire Responses—Difficulty of Assessment Grade 5

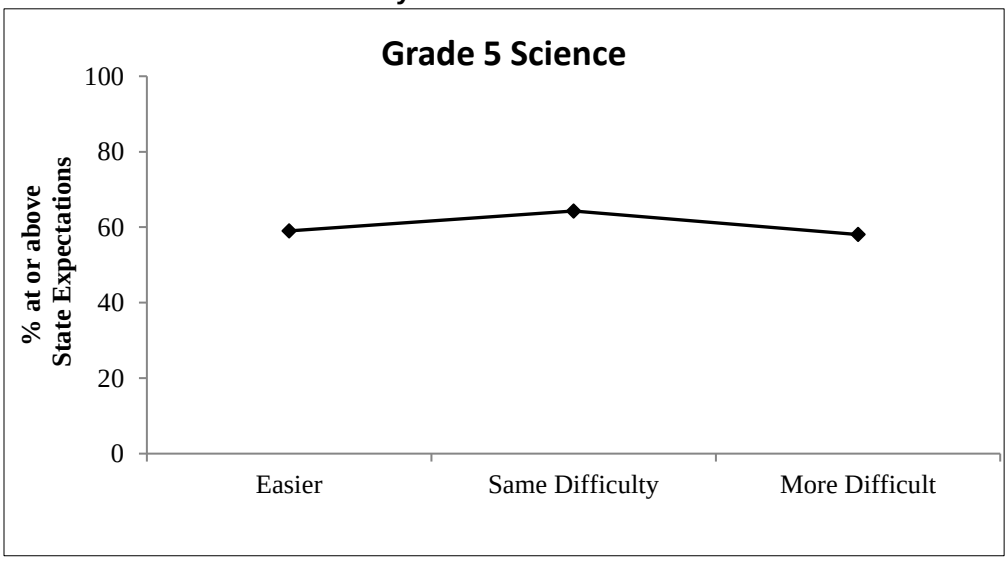


Figure 9-12. 2018–19 MEA Science: Questionnaire Responses—Difficulty of Assessment Grade 8

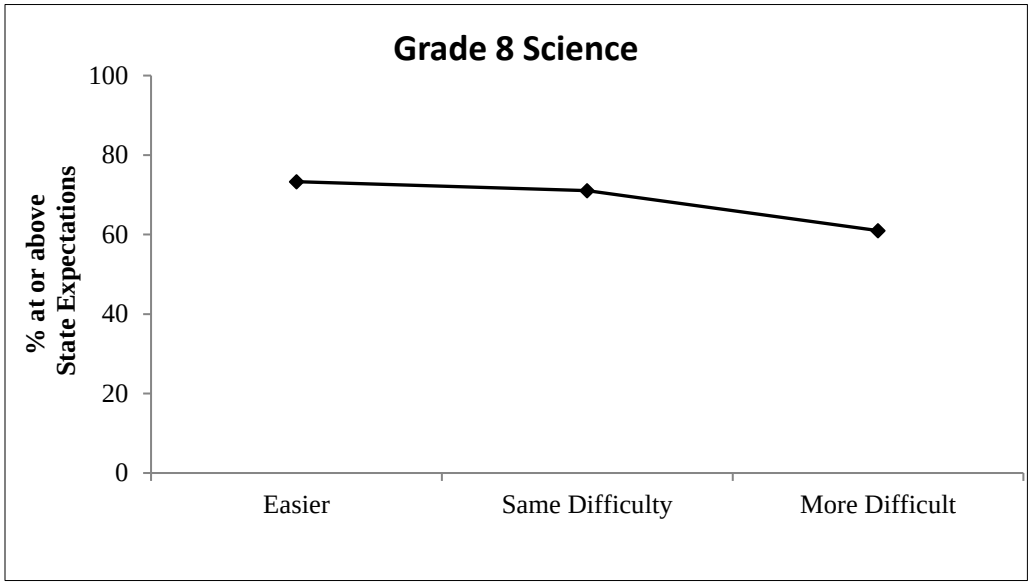
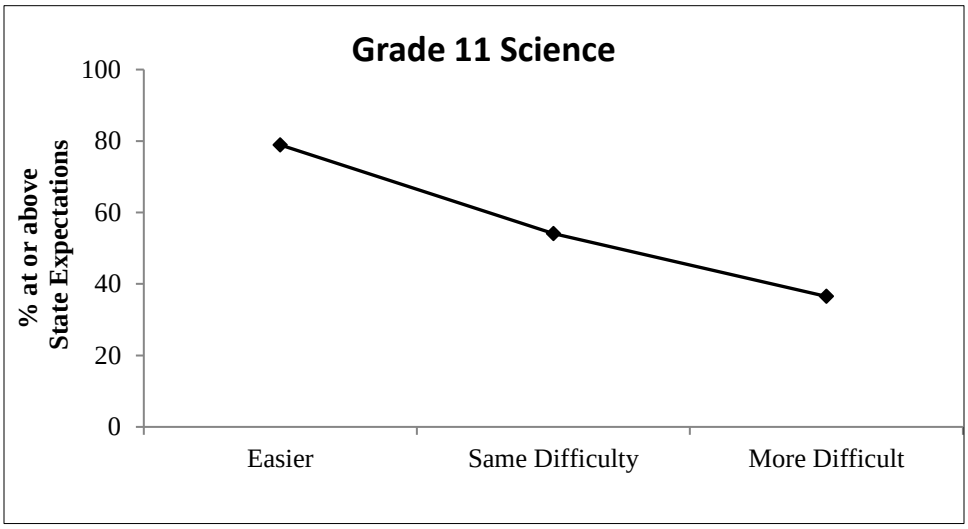


Figure 9-13. 2018–19 MEA Science: Questionnaire Responses—Difficulty of Assessment Grade 11



The evidence presented in this report supports inferences of student achievement on the content represented in the MLRs and grade-level expectations for science for the purposes of program and instructional improvement and as a component of school accountability.

REFERENCES

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

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

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

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

Chicago Manual of Style (16th ed.). (2010). Chicago: University of Chicago Press.

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

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

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

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

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

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

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

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

Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education*. Washington, D.C.: National Council on Measurement in Education.

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

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

Muraki, E. & Bock, R. D. (2003). *PARSCALE 4.1*. Lincolnwood, IL: Scientific Software International.

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

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262).

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

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

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

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

APPENDICES

APPENDIX A—MAINE TECHNICAL ADVISORY COMMITTEE MEMBERS

Table A-1. 2018–19 MEA Science: Technical Advisory Committee Members

Brian Gong	Executive Director of Center for Assessment, NCIEA
Nathan Dadey	Postdoctoral Fellow, NCIEA
Martha Thurlow	Director, National Center on Educational Outcomes
Betsy Webb	Superintendent, Bangor Public Schools
April Zenisky	Research Associate Professor, Department of Educational Policy, Research & Administration, University of Massachusetts Amherst

APPENDIX B—ITEM REVIEW AND BIAS AND SENSITIVITY REVIEW COMMITTEE MEMBERS

Table B-1. 2018–19 MEA Science: Bias and Sensitivity Committee Members

Name	*Department*
Judy Carey	Blind/VI
Rebecca L Perez	SpEd
Melissa Adams	SpEd/EL
Sheree Granger	SpEd
Andrea Freed	SpEd
Anne Marie Boucher	SpEd
Shelby Thebodeau	SpEd
Heather Anderson	EUT
Kathy Corey	Rural: Distant
Kristy LoCicero	Deaf

Table B-2. 2018–19 MEA Science: Item Review Committee—Science Grade 5

Name	*School*
Tonya Prentice	Woodstock Elementary School
Michael Harris	York Middle School
Laurette Darling	Albert S. Hall School
Carol Coy	Elm Street School
Pamela Thompson	Retired
Mandy Cyr	Biddeford Schools
Jim Chandler	Retired
Deidre Braley	Dayton Consolidated School

Table B-3. 2018–19 MEA Science: Item Review Committee—Science Grade 8

Name	*School*
Lindsay Noyes	SeDoMoCha Middle School
Margaret Pietrak	William S. Cohen Middle School
Carolyn Vose	Williams S Cohen-Bangor
Deidre Braley	Dayton Consolidated School
Mandy Cyr	Biddeford Schools
Andrea Freed	Mall-Dale Middle School
Jacey Morrill	Greely Middle School
Pamela Easler	Lawrence Junior High
Elizabeth Ladner	Richmond Middle & High School
John Libby	Richmond Middle & High School

Table B-4. 2018–19 MEA Science: Item Review Committee—Science Grade 11

Name	*School*
Jen Briggs	Nokomis High School
Beth Chagrasulis	Nokomis High School
Wendy Curtis	Waynflete
Helen Sprague	Marshwood High School
Tom Keller	Retired
Ken Vencile	Camden Hills Regional HS

APPENDIX C—DEMOGRAPHIC DATA OF PARTICIPANTS

Table C-1. 2018–19 MEA Science: Summary of Participation by Demographic Category—Science

Description	*Number Tested*	*Percent*
All Students	38,113	100.00
Male	19,530	51.24
Female	18,571	48.73
Gender Not Reported	12	0.03
Hispanic or Latino	853	2.24
American Indian or Alaskan Native	329	0.86
Asian	584	1.53
Black or African American	1,318	3.46
Native Hawaiian or Pacific Islander	38	0.10
White (non-Hispanic)	34,100	89.47
Two or More Races (non-Hispanic)	879	2.31
Race not reported	12	0.03
Currently receiving EL services	990	2.60
Former LEP student - monitoring year 1	247	0.65
Former LEP student - monitoring year 2	251	0.66
Former LEP student - monitoring year 3	-	-
Former LEP student - monitoring year 4	-	-
LEP: All Other Students	36,625	96.10
Students with an IEP	6,631	17.40
IEP: All Other Students	31,482	82.60
Economically Disadvantaged Students	16,128	42.32
SES: All Other Students	21,985	57.68
Migrant Students	2	0.01
Migrant: All Other Students	38,111	99.99
Students receiving Title 1 Services		
Title 1: All Other Students	38,113	100.00
Plan 504	2,378	6.24
Plan 504: All Other Students	35,735	93.76

APPENDIX D—ACCOMMODATION FREQUENCIES

Table D-1. 2018–19 MEA Science: Numbers of Students Tested with Accommodations by Accommodation Type and Grade

Accommodation Code	*Grade 5*	*Grade 8*	*Grade 11*
sciAccomT1	1,521	897	316
sciAccomT2	65	44	26
sciAccomT3	1,016	546	130
sciAccomT4	37	27	10
sciAccomS1	111	43	6
sciAccomS2	15	7	2
sciAccomP1	229	140	28
sciAccomP2	1,942	1,455	450
sciAccomP3	1,052	505	92
sciAccomP4	109	41	7
sciAccomP5	108	62	4
sciAccomP6	2	5	1
sciAccomP7	9	23	9
sciAccomP8	0	0	1
sciAccomP9	3	27	27
sciAccomP10	5	2	1
sciAccomR1	503	205	33
sciAccomR2	2	4	1
sciAccomR3	74	19	2
sciAccomR4	5	10	8
sciAccomR5	0	1	1
sciAccomR6	4	7	33
sciAccomR7	248	245	4
sciAccomO1	7	0	3

APPENDIX E—ACCOMMODATIONS GUIDE

MEA Science Accommodations Tables

Updated January 2019

NOTE: The chart below contains accommodations and codes to be used for the MEA Science test and should be bubbled on the student's answer booklet after testing.

SUPPORT(S) may provide more accessibility to the test and are determined on an individual basis by an educational team such as Response to Intervention (RtI), Student Assistance Team (SAT), and/or Language Acquisition Team. Supports must be consistent with the student's normal routine during instruction. Provision of supports does not alter the construct of any test item.

ACCOMMODATIONS are changes in procedures or materials that do not alter what the test measured, and are used to increase equitable access during assessment for students for whom there is a documentation of the need on an Individualized Education Program (IEP) or 504 Plan.

Any support(s) or accommodation(s) used for the assessment of an individual student will be the result of a team decision made at the local level. All decisions regarding the use of supports/accommodations must be made on an individual student basis—not for a large group, entire class, or grade level. Supports/accommodations are available to all students on the basis of individual need regardless of disability status and should be consistent with the student's normal routine during instruction and assessment. This table is not intended to be used as a stand-alone document and should always be used in conjunction with the *Test Administrator Manuals*.

T. Timing			
Code	**Description**	**Details on Delivery**	**Type**
T1	**Extended Time** - With time extended beyond standard administration (same day).	The scheduled administration time already includes additional time and the vast majority of students complete the test session within that time period. Extended time within a single sitting may be needed by students who are unable to meet time constraints. A test session may be extended until the student can no longer sustain the activity.	Support (team documented)
T2	**Extended Time** - With time extended beyond standard administration (several days).	In rare and severe cases, the extended time (T1) may not be adequate for a student not able to complete a test session within a single day. A test session may be administered to a student as two or more "mini-sessions" if procedures are followed to maintain test security and ensure that the student has access only to the items administered on that day.	Support (team documented)
T3	**Breaks** - With multiple or frequent breaks.	Multiple or frequent breaks may be required by students whose attention span, distractibility, physical and/or medical condition require shorter working periods.	Support (team documented)
T4	**Flexible Times** - At a time of day or a day of the week most beneficial to the student and/or late start.	Individual scheduling may be used for a student whose school performance is noticeably affected by the time of day or day of the school week on which it is done. This support may not be used specifically to change the order of administration of test sessions. This support must not result in the administration of a test session to an individual student prior to the regularly scheduled administration time for that session for all students.	Support (team documented)

S. Setting			
Code	**Description**	**Details on Delivery**	**Type**
S1	**Preferential Seating**	Students may have predetermined seating locations to reduce distractions for themselves or others, or to increase physical access to special equipment.	Support (team documented)
S2	**Out-of-School Setting** - With school personnel.	Out-of-school testing may be used for a student who is hospitalized or tutored because they are unable to attend school. The test must be administered by trained school personnel familiar with test administration procedures and guidelines. Relatives/guardians of the student may not be used as the test administrator/proctor.	Support (team documented)

MEA Science Accommodations Tables (continued)

P. Presentation			
Code	**Description**	**Details on Delivery**	**Type**
P1	**Individually**	Individual or small group testing may be used to minimize distractions for students whose test is administered out of the classroom or so that others will not be distracted by accommodations being used.	Support (team documented)
P2	**Small group**		
P3	**Read Aloud/Human Reader/American Sign Language – Directions AND Test**	A reader may be used for a student whose inability to read would hinder performance on the Science test. Words must be read as written. No translations (with the exception of signed language) or explanations are allowed. Trained personnel may use sign language to administer the test. Please see MEA Science Read Aloud Protocol.	Accommodation (IEP/504)
P4	**Read Aloud/Human Reader/American Sign Language – Directions ONLY**	A reader may be used for a student whose inability to read or locate directions would hinder performance on the test. Note that most directions on the MEA test occur at the beginning of the test session and are already read aloud by the test administrator/proctor. Guidelines for what are and are not "test directions" must be followed. With the exception of sign language and the case of students enrolled in a program where the test administrator/proctor routinely presents information in a foreign language, directions may not be translated. Please see MEA Science Read Aloud Protocol.	Accommodation (IEP/504)
P5	**Movement** - With opportunity for student to move, stand, and/or pace during assessment.	This opportunity may be used in a single-student setting other than the classroom for a student who cannot focus when seated for sustained periods of time.	Support (team documented)
P6	**Alt. Aids** - Using alternative or assistive technology that is part of the student's regular instruction.	The test may be presented through his/her regular communication system to a student who uses alternative or assistive technology on a daily basis.	Support (team documented)
P7	**Large-Print** version of assessment.	A large-print version of the MEA will be supplied based on registration with this accommodation.	Accommodation (IEP/504)
P8	**Braille** version of assessment.	A Braille version of the MEA will be supplied based on registration with this accommodation.	Accommodation (IEP/504)
P9	**ELL** - English language learner student use of bilingual word translation glossary.	Word-to-word translation glossary with NO definitions as determined by Language Acquisition Committee/Team. Students assigned this accommodation should be familiar with using such a glossary routinely during classroom instruction and test-taking.	Support (team documented)
P10	**Visual or Auditory Supports**	The test may be presented using visual aids such as visual magnification devices, reduction of visual print by blocking or other techniques, acetate shields; or auditory devices such as special acoustics, amplification, noise buffers, whisper phones, or calming music.	Support (team documented)

R. Response			
Code	**Description**	**Details on Delivery**	**Type**
R1	**Scribe** - With a student dictating responses to school personnel.	A student may dictate answers to constructed-response or short-answer questions to locally trained personnel or record oral answers in an individual setting so that other students will not benefit by hearing answers or be otherwise disturbed. Policies regarding recorded answers must be followed prior to returning special test materials. Please see MEA Science Scribing Protocol.	Accommodation (IEP/504)
R2	**Alt. Aids** - Using alternative or assistive technology/ devices that are part of the student's regular instruction, including visual and/ or auditory aids.	Technology is used to permit a student to respond to the test. Policies regarding recorded answers must be followed prior to returning test materials. Visual aids include any optical or nonoptical devices used to enhance visual capability. Examples include magnifiers, special lighting, markers, filters, large-spaced paper, color overlays, etc. An abacus may also be used for a student with severe visual impairment or blindness on the Science tests. Note that the use of this accommodation still requires student responses to be recorded in a student answer booklet.	Support (team documented)
R3	**Focus** - Other assistive devices and/ or tools to minimize distractions.	Noise buffers, place markers, carrels, etc., may be used to minimize distractions for the student. This accommodation does NOT include assistive devices such as templates, graphic organizers, or other devices intended specifically to help students organize thinking or develop a strategy for a specific question.	Support (team documented)
R4	**Computer/ Alternate Paper Response**	A student may use technological or other tools (e.g., large-spaced paper) to write responses to constructed-response items. A key distinction between this accommodation and R2 is that the student using this accommodation is responding in writing rather than dictating. When using a computer, word-processing device, or other assistive technology, access to the Web must be turned off. This accommodation is intended for unique individual needs, not an entire class. Policies regarding recorded answers must be followed prior to returning special test materials.	Accommodation (IEP/504)
R5	**Braille Response** - With student use of a Braille device.	A student may use a Braille device, manual and/or electronic, to respond to questions. Responses need to be recorded in standard format by a scribe, and policies regarding recorded answers must be followed prior to returning special test materials.	Accommodation (IEP/504)
R6	**ELL** - English language learner student use of bilingual word translation glossary.	Word-to-word translation glossary with NO definitions as determined by Language Acquisition Committee/Team. Students assigned this accommodation should be familiar with using such a glossary routinely during classroom instruction and test-taking.	Support (team documented)
R7	**Clarification** - Administrator verification of student understanding following the reading of test directions.	After test directions have been read, the test administrator/ proctor may ask the student to explain what he/she has been asked to do. If directions have been misunderstood by the student, the test directions may be paraphrased or demonstrated. Test items MUST NOT be paraphrased or explained.	Support (team documented)

MEA Science Accommodations Tables (continued)

O. Other This accommodation requires DOE approval or no credit will be given.			
Code	**Description**	**Details on Delivery of Accommodations**	**Type**
O1	Using other accommodation(s) not on this list, requested by the accommodations team.	An IEP team or other appropriate accommodation team may request that a student be provided an accommodation not included on this standard list of accommodations. Like all other accommodations, these should be consistent with the student's normal routine during instruction and/or assessment. Requests should be made to the DOE when accommodation plans are being made for a student prior to testing. DOE approval must be received for the requested accommodation to be coded as an O1 accommodation. Nonapproved accommodations used during test administration will result in a nonscored test.	Accommodation (IEP/504)

APPENDIX F—PROCESSING AND REPORTING BUSINESS REQUIREMENTS

Processing and Reporting Business Requirements

Maine Education Assessment Science

161501: Spring 2019 Maine Science

Version Number	Date	Updated Content Description	Updated By Name
1.0	02/07/2019	Initial Document	Keira Nevers
1.1	02/15/2019	Client Feedback	Keira Nevers
2.0	02/25/2019	Client Feedback Changes	Keira Nevers

Glossary	
DOE	Department of Education
ELA	English Language Arts
PBT	Paper Based Test
CBT	Computer Based Test
LP	Large Print
TTS	Text to Speech
ASR	Assessed Student Review
PAAP	Personalized Alternate Assessment Portfolio
SRB	Student Scanned Booklet
SDF	Student Demographic File
SAU	School Administration Unit
ASR	Assessed Student Review

Approval

I acknowledge that I have read this Decision Rules document and been informed of the contents of this document. By entering my name, title and date approved, I certify my approval. I have received a copy of this document for my records and understand that any further changes will require additional approvals as necessary.

Version	Printed Name	Title	Date Approved
2.0	**Nancy Godfrey**	Assessment Coordinator, DOE	03/06/2019
2.0	**Varun Motay**	Assessment Data Analyst, DOE	03/06/2019
2.0	**Jimmy Hartman**	Project Management Function Project Manager, MP	03/06/2019
2.0	**Keira Nevers**	Senior Business Analyst, MP	03/06/2019
2.0	**Tyler Blouin**	Primary Data Analyst Engineer, MP	03/06/2019
2.0	**Lezlie Turner**	Primary Data Analyst, MP	03/06/2019

Table of Contents

I. Overview 4
A. Points of Contact 4
B. Document References 4
C. Assumptions 4
D. Risks 5
II. General Information 5
A. Assessments 5
B. Reporting Cycles 6
C. Receivables 6
D. Deliverables 6
E. Quality Assurance 7
III. IT Processing Pre-Test Assessment Administration 7
A. Student Roster and Test Data Preparation 7
IV. Post Test Assessment Administration 8
A. Assessed Student Review 8
B. Test Data 8
C. Scan Paper Delivery and Data Denotation 8
D. Discrepancy Processing 8
E. Scoring Data 9
V. Participation and Exclusions 10
A. School and District Summary 10
B. Test Attempt Rules 11
C. Not Tested Reasons 11
D. Test Invalidations 12
E. Student Participation Status 12
F. Special Circumstances (by subject) 12
G. Student Participation Status Hierarchy (by subject) 13
VI. Calculations 14
A. Rounding Table 14
B. Raw Scores 15
C. Released Item Data 15
D. Item Scores 15
E. Item Averages and Percentages (on the Item Analysis Report) 15

F. Students included in calculations based on participation status 15
G. Average Points Earned Students At State Expectation (Range) 15
H. Content Standards 15
I. Questionnaire 16
J. Scaling: Assignment of Scaled Score and Achievement Level 16
VII. Data File Rules 17
A. State Raw Data 17
B. State Scored Data 17
C. State Standard Deviations & Average Scaled Scores 17
D. Invalidated Students Original Score 17
E. Department Chair, Teacher and Principal Questionnaire Raw Data 18
F. Department Chair, Teacher and Principal Questionnaire Frequency Distribution 18
G. Minimally Statistically Significant Differences for Scaled Scores 18
H. Standard Error of Measurement 18
I. Released Item Percent Responses Data 18
J. Data File Table (YYYY indicates year, GG indicates grade, SSS indicates subject) 18
VIII. Non-Functional Requirements 19
A. A. Operational Requirements 19
B. Approval 20

I. Overview

This document will describe the Information Technology Processing and Reporting requirements for the 2018-2019 Maine Science testing window for the Maine Department of Education in support of providing assessment reporting for Maine Science final student test results.

A. Points of Contact

Title	*Name*	*Contact Email*
Client Services Program Manager	Jaimie Rissin	Rissin.Jaimie@measuredprogress.org
Project Management Functional Project Manager	Jimmy Hartman	Hartman.Jimmy@measuredprogress.org
Assessment Coordinator	Nancy Godfrey	Nancy.Godfrey@maine.gov
Assessment Data Analyst	Varun Motay	Varun.Motay@maine.gov
Project Manager Information Technology	Tony Scavuzzo	Scavuzzo.Tony@measuredprogress.org
Manager Information Technology Processing & Reporting	Sanjay Iyer	Iyer.Sanjay@measuredprogress.org
Manager Information Technology Software Quality Assurance	Scott Duquette	Duquette.Scott@measuredprogress.org
Senior Business Analyst Information Technology	Keira Nevers	Nevers.Keira@measuredprogress.org
Primary Processing Developer	Lezlie Turner	Turner.Lezlie@measuredprogress.org
Primary Data Analyst Engineer	Tyler Blouin	Blouin.Tyler@measuredprogress.org
Principal Data & Reporting Architect	Andrea Hebert	Hebert.Andrea@measuredprogress.org

B. Document References

Name the documents that were sources or results of this document that must be referenced to deliver the requirements as set forth in this document. Examples are File Layouts, and any referenced Technical Documents,

Pertains to the referenced documents and the location where the documents can be found. Each document shall be stored in the SharePoint site and have a Change Control audit log.

1) Final Test Results Layout and Specification
2) 2018-2019eMetricPostAdminTransfer.xls
3) DP Layout.xls
4) Student Demographic Layout File Specifications
5) ASR Layout Specifications

C. Assumptions

In order to commit to delivering Data and Printed reports to the client, the following assumptions must be assumed.

D. Risks

Any risks shall be identified and recorded in their respective repositories. All stakeholders shall be notified of any risks associated to their responsible area's and be engaged as necessary.

1) Reporting issues from last Administration shall be mitigated
   a) The Student Demographic File was transmitted from DOE to MP IT Processing with incorrect values for Active Status field
      i) The DOE expected Labels for students, yet were not indicated as Active value '2'
      ii) Grade 11 students are considered Third Year High School and shall be considered for producing Labels, yet shall not be considered in Aggregate Reporting

Maine Student Demographic Layout				
20	Active	1	Active Status	1=Yes, 0=No, 2 = 3rd Yr HS

II. General Information

A. Assessments

1) Assessments

Contract	*Test Administration Open*	*Test Administration End*
Maine Science High School	April 1, 2019	April 12, 2019
Maine Science Grade 05 & 08	April 22, 2019	May 3, 2019

2) Test Forms Maine Science
   a) Grades 5, 8 and HS
   b) Content Area Science
   c) Forms
      i) Grade 5: 4
      ii) Grade 8: 5
      iii) Grade HS: 5
   d) Accommodated Forms
      i) PBT: Large Print and Braille
3) Test Administration Names
   a) Maine 2018-2019
   b) Test Name MEA 2018-19 SCI GR[xs]
4) Test Blueprints

Grade	*Session 1*	*Session 2*	*Session 3*	*TOTAL*		*Reporting Categories*		
				MC	*CR*	*RepCat*	*Standard*	*Indicator*
5	8 MC, 2 CR	6 MC, 1 CR	18 MC, 1 CR	32	4	Cat3	Cat3	Cat4
8	11 MC, 2 CR	11 MC, 1 CR	18 MC, 1 CR	40	4	Cat3, Cat4	Cat3	Cat4
11	16 MC, 2 CR	24 MC, 2 CR	N/A	40	4	Cat3, Cat4	Cat3	Cat4
MC = multiple-choice; CR = constructed-response								

5) Distribution of Score Points

Science Standards	Grade 5	Grade 8	Grade 11
D1–D2 (Earth / Space)	12	17	12
D3–D4 (Physical)	12	17	22
E1–E5 (Life)	24	22	22
Total Score Points	48	56	56

B. Reporting Cycles

1) Pre-Assessment Student Roster
2) Pre-Assessment Labels to Schools for Paper Tests
   a) Pre-Assessment Labels to Districts for Paper Test for Private Schools (8000 ID schools)
3) Demographic Clean up Post Test Assessment
4) Final Reporting

C. Receivables

Data Files received shall pass all validation rules and formats based on the layout and specification document. All documents must be complete and accurate upon receipt.

1) Answer Documents
2) Questionnaire Survey Result Documents
   a) Principal/Test Coordinator
   b) Teacher Questionnaire for Grade 5 and Grade 8
   c) Department Chair Questionnaire for Grade 11
3) PreID File shall be delivered to Cognia from DOE
   a) PreID file and Demographic files are received for particular scheduled cycles
4) ASR File shall be sent to MP after review and edits are complete from DOE
   a) ASRFileLayout.xlsx
5) MP Reporting group shall have access to the Test Items to pull the Content Bank Metadata
6) MP Reporting group shall receive Scanned Paper results from Scanning
7) MP Reporting group shall receive Scored Data for Hand-scored items and Machine-scored items from Scoring
8) MP Reporting group shall have access to retrieve Test Results Data

D. Deliverables

Data Files delivered shall pass all validation rules and formats based on the layout and specification document.

1) Maine Science
   a) State Raw Data Files (by grade)
   b) State Scored Data Files (by grade)
   c) State Student Released Item Files (by grade)
   d) State Standard Deviations & Average Scaled Scores (by grade)
   e) Department Chair, Teacher, and Principal Questionnaire Raw Data
   f) Department Chair, Teacher, and Principal Questionnaire Frequency Distribution

g) Minimally Statistically Significant Differences for Scaled Scores (by grade and subject)
h) Standard Error of Measurement (by grade and subject)
i) Raw Score Ranges
j) Scaled Score Lookup
k) Released Item Percent Responses Data (For Program Management)
l) MEA School List for Mailing (For Program Management)

E. Quality Assurance

All datafiles and reports identified as a deliverable to the Client shall pass internal Quality Assurance measures. The Software Quality Assurance (SQA) team works together with the data processing and data analysis teams to ensure quality data is captured and delivered accurately. Quality control checks are being performed by the data processors and data analysts as the data is handed off via multiple internal software tools. Included in the final execution, the Software Quality Assurance team executes test cases validating student printed reports and student labels for accuracy in comparison to the previously agreed upon report design specifications.

III. IT Processing Pre-Test Assessment Administration

Pre-Test Assessment Administration activities shall be completed prior to the Test Assessment Administration window. The Pre-Administration window shall allow for the client to gather the student and testing subject data to provide Measrued Progress and all other vendors the information to administer the Test Assessments.

A. Student Roster and Test Data Preparation

1) MP shall receive the State Student Demographic file with all students from the DOE in the specified file format and layout (Maine Student Demographic Layout.xlsx)
   a) The incoming Student Demographic file shall be comma separated value file format (csv)
   b) The incoming Student Demographic file shall contain a Header row
   c) The incoming Student Demographic file shall use a comma delimiter for data fields, and shall not contain any other delimiter values
   d) The Student Demographic file shall indicate any paper booklets and any accommodated booklets using the TestType field
   e) The Student Demographic file shall be validated upon receipt
      i) Validate all school and district codes exist in MP's iCore Database
      ii) Date of Birth values must be numeric, else MP will set to blank
      iii) First Name or Last Name fields shall contain no special characters, MP will set special characters to blank
      iv) Student ID's must be unique across all Schools and Districts
      v) MP shall compare Student ID's with prior year and report all name differences
      vi) MP shall compare Student ID's with prior year grade and report all students that remained the same grade
      vii) MP shall report the count of students by grade and test type

IV. Post Test Assessment Administration

The Test Assessment Administration window shall be defined and closed prior to processing and reporting for Reporting. The commencement of the testing window shall initiate activity to complete all Results and Reporting to the Client.

The Test Assessment Administration window shall be defined and closed prior to processing. The commencement of the testing window shall initiate activity to complete all deliverables to the Client.

A. Assessed Student Review

1) The client shall be provided a Post Test Clean Up five-day time frame which shall be used for additional post-test clean-up of Student Participation and Demographic record modifications
2) A specified window shall be determined where the client shall have the opportunity to update all allowable edits necessary to the Post Test records between June 10 and June 14, 2019
3) DOE shall use the ASR file layout and specifications during the clean-up window
   a) The ASR File Specification has specific fields identified that may be edited as well as not
   b) Any field that does not allow edits will be ignored during processing

B. Test Data

1) Every imported Test record shall have an associated Student record
2) Paper Test Items shall have an associated BarcodeID assigned with the initial Student Demographic file which Cognia will use to link the Test metadata including, but not limited to Test Item, Grade, Subject and Session
3) Duplicate test records shall be merged/resolved prior to processing
   a) All duplicate records shall be reviewed and updated accordingly based on DOE feedback as necessary
   b) In the event the Student has a Test record with no items attempted, Cognia shall not supress any records unless specifically directed to do so as part of the duplicate resolution
4) Processing of Test Records for Data Analysis shall begin no later than July 1, 2019

C. Scan Paper Delivery and Data Denotation

Each Paper Booklet test shall be scanned and delivered immediately to IT Processing team. At the time of receipt, IT Processing shall perform procedures to accurately identify inaccuracies in the data. IT Processing shall receive final raw scored data at time of delivery.

1) The data file shall be formatted as specified in the Scan Delivery Layout Format
2) Any and all discrepancies with the Scan File shall be resolved accordingly
3) IT Processing shall validate each field for each Student, Test and Item metadata to ensure the data is acceptable
4) IT Processing shall provide a report of any and all discrepancies back to the Scanning department for research and/or re-scanning
5) IT Processing shall validate each Label Barcode and update as necessary

D. Discrepancy Processing

Identification of Discrepancies in records shall have a defined process to determine the record of source for reporting purposes. Duplicates may exist where there are more than one data record with

the same Student ID. Other discrepancies may be defined within the data where the DOE shall have the opportunity to review prior to scoring test records.

1) Duplicate Test records with the same Student ID, Subject and Language shall be combined or otherwise suppressed
   a) If there are duplicate tests associated to one Student ID, the test which has attempted items shall be processed, and all other tests that do not have valid attempted items shall be suppressed
   b) If there are duplicate tests associated to one Student ID, and the duplicate tests do not have any attempted items, all but one test record shall be suppressed from processing
   c) If there are duplicate tests associated to one Student ID, and the duplicate tests have different sessions attempted, the data shall be merged into one record with the Last booklet ID (Last session attempted) as the source of record
      i) The data from the booklet with the first attempted items shall be incorporated into the book of record
   d) If there are duplicate tests associated to one Student ID, and the duplicate books have some of the same items attempted, the data shall be identified as requiring Client and Program Management intervention for resolution
   e) All unresolved duplicate student and or test records shall require Client and Program Manager participation for resolution
2) Duplicate Cross Grade tests are identified as tests where the same student has taken a test not within their actual grade
   a) If a student has multiple grade tests, compare the Test Grade to the Student grade
   b) Should the Student have no work in the off-grade book and work in the matching grade book, suppress the off-grade book responses
   c) If both books have responses, send a report to Program Management for research and resolution
3) Braille Validation
   a) Send PM a report for confirmation of booklets with Student with Disability Accommodation bubbled for any subject
   b) Should PM determine student is not Braille, clear Student with Disability Accommodation as required
   c) Otherwise, set Accommodation for all subjects for the identified Student
4) Login Discrepancy shall be identified when the Login School or District is different than what was submitted during student registration and reported to Program Management for resolution
5) Program Management and DOE shall participate in resolution of any and all discrepancies remaining after initial processing of student and test records

E. Scoring Data

Scoring division shall provide IT Processing with the scores from all tests.

1) Every score record shall contain a valid score value
   a) A validation of score values shall be performed
   b) In the event that a score value is found to be invalid, resolution shall be attempted with the Scoring Division

2) Each score record shall be associated with a Booklet ID
   a) If a score record is received without an associated Booklet ID, resolution shall be attempted with the Scoring Division
3) All unresolved scoring records shall be included in a report to the Scoring Division, as well as the Program Manager for research and resolution

V. Participation and Exclusions

A. School and District Summary

1) School types are identified in groups for reporting purposes
   a) Public School
   b) Charter School
   c) Public Special Purposes
   d) Public Special Education
   e) Private School (with ≥60% or more publicly funded)
   f) Private Special Purpose
2) All student results data shall report based on Testing District Code and School Code
   a) School Administration Unit data shall be blank for students tested at Private School or Private Special Purpose school designations
   b) Tested Year shall always be present
3) A datafile shall be reported for each school with at least one student enrolled, regardless of School Type designation
   a) SAU data will be blank for Private School or Private Special Purpose school designations
   b) Tested Year shall always be present
4) A datafile shall be reported for each SAU

SchType	**Source: ICORE SubTypeID**	**Description**
'PUB'	1	Public
'CHA'	11	Charter School
'PSP'	19	Public Special Purpose
'PSE'	15	Public Special Ed
'BIG'	6	Private with 60% or more Publicly Funded (Big 11)
'PSN'	23	Private Special Purpose

School Type impact on Data Analysis and Reporting		
Level	**Impact on Analysis**	**Impact on Reporting**
Student	n/a	Report students based on testing discode and schcode. SAU data will be blank for students tested at BIG or PSN schools. Always print tested year state data.
School	Do not exclude any students based on school type using	Generate a report for each school with at least one student enrolled using the tested school aggregate denominator.

School Type impact on Data Analysis and Reporting		
Level	**Impact on Analysis**	**Impact on Reporting**
	testing school code for aggregations	SAU data will be blank for BIG and PSN schools. Always print tested year state data.
SAU	For BIG and PSN schools, aggregate using the sending SAU. If BIG or PSN student does not have a sending SAU, do not include in aggregations.	Generate a report for each SAU with at least one student enrolled using the tested SAU aggregate denominator. Always report tested year state data.
State	Include all students regardless of schtype.	Always report testing year state data.

B. Test Attempt Rules

Test Attempted indicates that a student has answered a minimum of one (1) operational multiple-choice test item within a content area, regardless number of sessions. Each of the items must not be indicated as flawed or otherwise not scorable.

1) A valid multiple-choice response shall only include A, B, C, D or multiple of such indicated with an asterisk
2) A student shall be considered attempted when
   a) The student provided a valid response to at least one multiple choice item
   b) The student has a non-blank score value for a common open response item

C. Not Tested Reasons

Not Tested Reasons shall be indicated as separate categories. Not Tested Reasons shall be defined by the DOE.

1) If a student has more than one reason for not participating on the test, we will assign one participation code using the following hierarchy
   a) Special Consideration
      i) If a student links to the demographic data file has content area “Not Tested State Approved Special Consideration” indicated, then the student is identified as “Not Tested State Approved Special Consideration”
   b) Alternate Assessment
      i) Students are identified as participating in the MEA Alternate Assessment based on the MEA Alternate Assessment Decision Rules for each subject assessed at the grade level in Synergy
   c) Not Tested Other
      i) If content area test was not attempted, the student is identified as ‘Not Tested Other’

D. Test Invalidations

Item invalidation codes shall be provided by the DOE during the ASDR clean up process defined. Student records shall maintain the continuity of the set values from the DOE in final reporting.

1) Each student shall be identified with a valid tested participation when there is no Invalidated Test Reason to report
   a) Tested Damaged SRB shall be identified when the Student Response Booklet is damaged and unable to be scored for reporting purposes
   b) Tested Incomplete shall be identified when the

E. Student Participation Status

Student Participation Status reflects the participation of the test assessment performed by an individual student.

StuStatus	Description
1	Homeschooled
2	Privately Funded
3	Bureau of Indian Education
4	Excluded State
0	Publicly Funded

StuStatus impact on Data Analysis and Reporting		
Level	**Impact on Analysis**	**Impact on Reporting**
Student	n/a	School and SAU data will be blank for students with a StuStatus value of 1,2 or 3. Always print tested year state data. For StuStatus values of 1, 2, and 3 print the description from the table above for the school and SAU names.
School	Exclude all students with a StuStatus value of 1or 2.	Students with a StuStatus value of 1, 2 or 3 are excluded from Interactive Reporting.
SAU	Exclude all students with a StuStatus value of 1or 2.	n/a
State	Exclude all students with a StuStatus value of 1, 2 or 4.	n/a.

F. Special Circumstances (by subject)

1) Item invalidation flags are provided to the DOE during data processing test clean up. The item invalidation flag variables are initially set using the rules below. The final values used for reporting are provided back to Cognia by the DOE and used in reporting
2) A student is identified as content area tested if the student does not have any content area not tested reasons identified. Tested students are categorized in one of the four tested participation statuses: “Tested Damaged SRB”, “Tested with Non-Standard Accommodations”, “Tested Incomplete”, and “Tested”

a) Students with a common item response of 'X' are identified as "Tested Damaged SRB".
b) Students identified as content area tested, are not identified as "Tested Damaged SRB", and have at least one of the content area invalidation session flags marked will be identified as "Tested with Non-Standard Accommodations".
c) Students identified as content area tested, are not identified as "Tested Damaged SRB", and not identified as "Tested with Non-Standard Accommodations" and did not attempt all sessions in the test are considered to be "Tested Incomplete."
d) All other tested students are identified as "Tested".

3) For students identified as "Tested Damaged SRB", the content area subcategories with at least one damaged item will not be reported. These students are excluded from all raw score aggregations (item, subcategory, and total raw score). They are included in participation, achievement level, and scaled score aggregations.
4) For students identified as "Tested with Non-Standard Accommodations" the content area sessions item responses which are marked for invalidation will be treated as a non-response.
5) Students identified as tested in a content area will receive released item scores, scaled score, scale score bounds, achievement level, raw total score, and subcategory scores.
6) Students identified as not tested in a content area will not receive a scaled score, scaled score bounds, or achievement level. They will receive released item scores, raw total score, and subcategory scores.
7) Item scores for students with an invalidation flag marked and have a not tested status will be blanked out based on the invalidation flag. For example, if the student is identified as "Not Tested: State Approved Alternate Assessment" for science and has SciInvSes1 marked, then all science session 1 item responses will be reported as a blank.
8) For students identified as any of the four tested participation categories they will be considered to be either tested with or without accommodations for participation purposes. A student is tested with accommodations if any of the accommodations were marked.

G. Student Participation Status Hierarchy (by subject)

1) Not Tested State Approved Special Consideration
2) Not Tested State Approved Alternate Assessment
3) Not Tested Other
4) Tested Damaged SRB
5) Tested with Non-Standard Accommodations
6) Tested Incomplete
7) Tested

Student Participation Summary (by subject)

Participation Status	Description	Raw Score (*)	Scaled Score	Ach. Level	Student Report Ach. Level Text	Roster Ach. Level Text
Z	Tested Damaged SRB (**)	✓	✓	✓	Well Below State Expectations, Below State Expectations, At State Expectations, or Above State Expectations	1,2,3, or 4
A	Tested	✓	✓	✓	Well Below State Expectations, Below State Expectations, At State	1,2,3, or 4

Participation Status	Description	Raw Score (*)	Scaled Score	Ach. Level	Student Report Ach. Level Text	Roster Ach. Level Text
					Expectations, or Above State Expectations	
C	Tested with Non-Standard Accommodations (%%)	✓	✓	✓	Well Below State Expectations, Below State Expectations, At State Expectations, or Above State Expectations	1,2,3, or 4
B	Tested Incomplete (%)	✓	✓	✓	Well Below State Expectations, Below State Expectations, At State Expectations, or Above State Expectations	1,2,3, or 4
D	Not Tested State Approved Alternate Assessment	✓			Alternate Assessment	A
H	Not Tested State Approved Special Consideration	✓			Special Consideration	S
I	Not Tested Other	✓			Not Tested	N

(*) Raw scores are not printed on student report for students with a not tested status.

(**) Raw scores for Tested damaged SRB students will be reported based on the set of non-damaged items. Subcategory scores will not be reported if it includes a damaged item.

(%) Tested incomplete students will be identified on the student report with a footnote.

(%%) Tested with Non-standard accommodations students will be identified on student report with a footnote. The invalidated items will be stored as a '-' for item analysis.

VI. Calculations

A. Rounding Table

Report	Calculation	Rounded (to the nearest)
Summary of Scores	Average Scaled Score	Whole value
Results	Percent at each achievement level, Percent of points possible, Percent of students in each Category, Scaled Score	Whole value
	Cluster Average Points Attained (Number & Percent) Round Cluster Average Points Attained before calculating the percent.	Tenth
Item Analysis Report	Multiple Choice Percent Correct	Whole value
	Open Response Average Score	Tenth
	Content Standard Earned Averages	Tenth
All Grade Summary	Percent at each achievement level, Mean Scaled Score	Whole value

B. Raw Scores

1) Raw scores are based on the scores for common items

C. Released Item Data

1) The data for the released items are provided by Program Management or exist in IABS.
2) Details on how the standard/performance indicators are derived can be found later in the document under the Content Standards section

D. Item Scores

1) For all analysis, non-response for an item by a tested student is treated as a score of 0.
2) For multiple choice released item data store a '+' for correct response, or A,B,C,D,* or blank
3) For open response released items, store the student score
   a. If the score is not numeric ('B')
   b. else store it as blank
4) For students identified as content area tested with non-standard accommodations, then store the released item score as '-' for invalidated items

E. Item Averages and Percentages (on the Item Analysis Report)

1) Multiple-choice item averages are multiplied by 100 and rounded to the nearest whole number
2) Open-response item averages are rounded to the nearest tenth

F. Students included in calculations based on participation status

1) For number and percent of students enrolled, tested, and not tested categories include all students not excluded by other decision rules.
2) For number and percent at each achievement level, average scaled score, subcategories average points earned, percent/correct average score for each released item include all tested students not excluded by other decision rules.
3) Students identified as Tested Damaged SRB are excluded from all raw score aggregations (item, subcategory, and total raw score). They are included in participation, achievement level, and scaled score aggregations.

G. Average Points Earned Students At State Expectation (Range)

1) Select all students across the states with Y40 scaled score, where Y=grade. Average the content area subcategories across the students and round to the nearest tenth. Add and subtract one standard error of measurement to get the range.

H. Content Standards

1) Science

a. The standard/performance indicator is calculated by concatenating the third and fourth portion of content framework. This is stored in tblIref in the Standard column.
b. Grade 5
    i. RepCat1: D. The Physical Setting
        1. This is defined in IABS when standard ='D' .
    ii. RepCat2: E. The Living Environment
        1. This is defined in IABS when standard = 'E' .
c. Grade 8 and Grade 11
    i. RepCat1: D. The Physical Setting
        1. This is defined in IABS when standard = 'D' .
    ii. RepCat2: D1/D2 Earth/Space
        1. This is defined in IABS when standard = 'D' and performance indicator = '1' or '2'.
    iii. RepCat3: D3/D4 Matter/Energy/Force/Motion
        1. This is defined in IABS when standard = 'D' and performance indicator = '3' or '4'.

I. Questionnaire

1) Only tested students will be included in the calculations.
2) Percent of students in this category is computed by the number of tested students that selected that response/number of tested students with a single response for the question *100. Students are considered to have a single response, if their response is not blank or '*'.
3) Achievement level data will be suppressed based on minimum N rules

J. Scaling: Assignment of Scaled Score and Achievement Level

1) Scaling is accomplished by defining the unique set of test forms for the grade/subject. This is accomplished as follows:
    a. Translate each form and position into the unique item number assigned to the form/position.
    b. Order the items by
        i. Type – multiple-choice, short-answer, constructed- response, extended-response, writing prompt.
        ii. Form – common, then by ascending form number.
        iii. Position
    c. If an item number is on a form, then set the value for that item number to '1', otherwise set to '.'. Set the Exception field to '0' to indicate this is an original test form.
    d. If an item number contains an 'X' (item is not included in scaling) then set the item number to '.'. Set the Exception field to '1' to indicate this is not an original test form.
    e. Compress all of the item numbers together into one field in the order defined in step II to create the test for the student.
    f. Select the distinct set of tests from the student data and order them by the exception field and the descending test field.
    g. Check to see if the test has already been assigned a scale form by looking in the tblScaleForm table. If the test exists then assign the existing scale form.

Otherwise assign the next available scale form number. All scale form numbering starts at 01 and increments by 1 up to 99.

2) Scaled Score Assignment
    a. Psychometrics provides data analysis with a lookup table for each scale form. The lookup table contains the raw score and the resulting scaled score.
    b. Scaled Scores are rounded to even integers.
3) Achievement level coding:
    a. 1 = Well Below State Expectations
    b. 2 = Below State Expectations
    c. 3 = At State Expectations
    d. 4 = Above State Expectations

VII. Data File Rules

A. State Raw Data

1) Only students from 'PUB','CHA', 'PSP', or 'PSE' schools are included, or if they have a sending SAU. Students with all participation statuses are included.
2) There are two files per grade; one with names and one without.
3) Field test item responses are not displayed.
4) Data is ordered by SAU code, School code, last name, first name.

B. State Scored Data

1) Only students from 'PUB','CHA','PSP', or 'PSE' schools are included, or if they have a sending SAU. Students with all participation statuses are included
2) There are two files per grade; one with names and one without
3) Field test item responses are not displayed
4) Data is ordered by subject, SAU code, School code, last name, first name

C. State Standard Deviations & Average Scaled Scores

1) Exclude students with not tested participation status.
2) The data reported in these files are the number of students tested, the average scaled score, and the standard deviations for the following subgroups:

a. Identified Disability, No Identified Disability
b. LEP (1st year or 2nd year and beyond), Not LEP
c. Economically Disadvantaged, Not Economically Disadvantaged
d. Migrant, Not Migrant
e. Gender
f. Ethnicity
g. Title 1, Not Title 1
h. Total (All students)

D. Invalidated Students Original Score

1) A CSV file will be created for each grade
2) Original raw scores for students whose responses were invalidated for reporting will be provided

E. Department Chair, Teacher and Principal Questionnaire Raw Data

1) One CSV file will be created containing raw Department Chair data
2) One CSV file will be created containing raw Teacher Questionnaire data
3) One CSV file will be created containing raw Principal Questionnaire data

F. Department Chair, Teacher and Principal Questionnaire Frequency Distribution

1) One CSV file will be created containing the distribution of responses of Department Chair, Principal, and Teacher Questionnaire raw data

G. Minimally Statistically Significant Differences for Scaled Scores

1) The data reported in this file are the number of scaled score points denoting minimally statistically significant differences for average school/SAU results. This is calculated by the psychometricians

H. Standard Error of Measurement

1) The data reported in this file are the number of students tested, the number of possible raw score points, the minimum, maximum, mean, standard deviation, reliability, and the SEM of the raw score

I. Released Item Percent Responses Data

1) One excel file will be produced for each grade
2) Will contain average score and percent at each response for the released items

J. Data File Table (YYYY indicates year, GG indicates grade, SSS indicates subject)

Data File	Layout	File Name
State Raw Data	MEAYYYYStateStudentRawDataLayout.xls	**YYYYMEAYYYYStateStudentRawDataGGade[GG].csv**
State Scored Data	MEAYYYYStateStudentScoredDataLayout.xls	**YYYYMEAYYYYStateStudentScoredDataGGade[GG].csv**
State Student Released Item	MEAYYYYStudentReleasedItemLayout.xls	**MEAYYYYStateStudentReleasedItem[GG].csv**
State Standard Deviations & Average Scaled Scores	MEAYYYYStateStandardDeviationsLayout.xls	**MEAYYYYStandardDeviation[GG].xls**

Data File	Layout	File Name
Minimally Significant Differences for Scaled Scores	N/A	**MEAYYYYSignificantDifferenceChart.xls**
Standard Error of Measurement	N/A	**MEAYYYYSEM.xls**
Raw Score Ranges	N/A	**MEAYYYYScoreRanges.xls**
Scaled Score Lookup	MEAYYYYScaledScoreLookupLayout.xls	**MEAYYYYScaledScoreLookup.xls**
MEA School Mailing List	N/A	**MEAYYYYSchDisList.xls**
Invalidated Students Original Score	MEAYYYYStateInvalidatedStudentOriginalScoredLayout.xls	**MEAYYYYStateInvalidatedStudentOriginalScored.csv**
Dept. Chair,Teacher and Principal Questionnaire Raw Data	MEAYYYYTeacherQuestionnaireRawLayout.xls MEAYYYYPrincipalQuestionnaireRawLayout.xls MEAYYYYDepartmentChairQuestionnaireRawLayout.xlsx	**MEAYYYYTeacher05QuestionnaireRaw.csv** **MEAYYYYTeacher08QuestionnaireRaw.csv** **MEAYYYYPrincipal0508QuestionnaireRaw.csv** **MEAYYYYPrincipal11QuestionnaireRaw.csv** **MEAYYYYDepartmentChair11QuestionnaireRaw.csv**
Dept. Chair, Teacher and Principal Questionnaire Frequency Distribution	MEAYYYYDepartmentChair_Teacher_PrincipalQuestionnaireFreqLayout.xls	**MEAYYYYDepartmentChairTeacherPrincipalQuestionnaireFreq.csv**
Released Item Percent Responses Data	**MEAYYYYReleasedItemPercentResponses Layout.xls**	**ReleasedItem[GG].xls**

VIII. Non-Functional Requirements

A. A. Operational Requirements

1) Vendor system
   a) Performance shall be satisfactory
   b) Availability shall be uninhibited during the open windows
   c) Security measured shall be in place for protection of data and transfers
   d) Usability of the system must be satisfactory
   e) Integrity of the system shall be adequate
2) Carrier vendor timeliness
   a) Material receipt is on time
   b) Material delivery is on time
3) Training
   a) Training is performed
   b) Training is available and delivered adequately
4) Systems support, and maintenance is available
5) Schedules are adhered to (include handoff schedule to and from reporting groups)
   a) Scheduled dates are agreed to and adhered to
6) Resources
   a) Availability of personnel must be adequate and permit capacity

b) Accessibility of systems and shall be available for processing and reporting

B. Approval

MEA Science
Business Rules (Decisi

APPENDIX G—INTERRATER AGREEMENT

Table G-1. 2018–19 MEA Science: Item-Level Interrater Consistency Statistics by Grade

Grade	*Item Number*	*Number of*		*Percent*		*Correlation*	*Percent of Third Scores*
		Score Categories	*Responses Scored Twice*	*Exact*	*Adjacent*		
5	165328	5	1,260	55.79	36.19	0.62	8.02
	191018	5	1,288	53.80	37.34	0.75	8.77
	191047	5	1,260	71.35	26.03	0.87	2.62
	199466	5	1,288	61.02	35.09	0.81	3.88
8	469354	5	1,304	81.37	17.02	0.87	1.61
	469905	5	1,299	64.13	28.41	0.81	7.47
	470376	5	1,265	72.65	25.45	0.86	1.90
	65117	5	1,304	71.24	25.46	0.87	3.30
11	164367	5	1,135	70.22	26.08	0.80	3.70
	258242	5	1,073	64.03	32.99	0.71	2.98
	546738	5	1,145	64.37	32.58	0.78	3.06
	550207	5	1,102	66.24	29.85	0.84	3.99

APPENDIX H—CLASSICAL ITEM ANALYSIS

Table H-1. 2018–19 MEA Science: Item-Level Classical Test Theory Statistics—Science Grade 5

Item		*Difficulty*	*Discrimination*	*Percent Omitted*
Number	*Type*			
96271	MC	0.77	0.28	0
96274	MC	0.83	0.25	1
96275	MC	0.75	0.33	1
96581	MC	0.73	0.40	1
97064	MC	0.54	0.27	1
138205	MC	0.70	0.36	1
138266	MC	0.63	0.35	1
138268	MC	0.60	0.37	1
138269	MC	0.71	0.22	1
138289	MC	0.81	0.22	0
164093	MC	0.44	0.30	2
165111	MC	0.63	0.24	1
165321	MC	0.51	0.23	1
165328	OR	0.37	0.42	3
166403	MC	0.54	0.35	1
166842	MC	0.40	0.18	1
166843	MC	0.67	0.40	1
169211	MC	0.85	0.32	1
169229	MC	0.68	0.33	1
191013	MC	0.42	0.21	1
191015	MC	0.70	0.51	1
191018	OR	0.44	0.50	4
191027	MC	0.86	0.35	0
191032	MC	0.58	0.39	1
191043	MC	0.64	0.34	2
191047	OR	0.58	0.44	2
228134	MC	0.68	0.40	1
228155	MC	0.72	0.17	1
248521	MC	0.82	0.31	1
248566	MC	0.59	0.22	1
466901	MC	0.62	0.36	1
467311	MC	0.67	0.51	1
467335	MC	0.70	0.36	0
467356	MC	0.50	0.22	1
468930	OR	0.44	0.57	1
547304	MC	0.50	0.30	1

Table H-2. 2018–19 MEA Science: Item-Level Classical Test Theory Statistics—Science Grade 8

Item		*Difficulty*	*Discrimination*	*Percent Omitted*
Number	*Type*			
63442	MC	0.68	0.47	0
64657	MC	0.86	0.38	1
65117	OR	0.47	0.52	2
96613	MC	0.71	0.36	1
97788	MC	0.56	0.26	0
138322	MC	0.66	0.35	1
138367	MC	0.47	0.37	0
138455	MC	0.67	0.38	1
153664	MC	0.85	0.44	0
153851	MC	0.73	0.40	1
153898	MC	0.77	0.21	1
164358	MC	0.63	0.47	1
165644	MC	0.62	0.34	0
165948	MC	0.67	0.34	0
165954	MC	0.52	0.29	1
187328	MC	0.81	0.29	0
191063	MC	0.50	0.38	1
191064	MC	0.66	0.37	1
191069	MC	0.54	0.38	0
191074	MC	0.81	0.46	0
191076	MC	0.73	0.40	0
191077	MC	0.82	0.37	0
191086	MC	0.64	0.39	0
191090	MC	0.54	0.26	1
191101	MC	0.51	0.32	1
191107	MC	0.72	0.23	0
191112	MC	0.65	0.44	1
228096	MC	0.67	0.34	1
228125	MC	0.77	0.36	1
250949	MC	0.51	0.33	1
258114	MC	0.72	0.14	0
258140	MC	0.58	0.34	0
258144	MC	0.65	0.44	0
258147	MC	0.71	0.35	0
469354	OR	0.45	0.47	1
469789	MC	0.73	0.36	1
469859	MC	0.52	0.33	1
469905	OR	0.41	0.51	4
470097	MC	0.40	0.27	1
470179	MC	0.41	0.30	1
470220	MC	0.64	0.31	0
470376	OR	0.51	0.52	3
470447	MC	0.71	0.36	0
541774	MC	0.63	0.42	1

Table H-3. 2018–19 MEA Science: Item-Level Classical Test Theory Statistics—Science Grade 11

Item		*Difficulty*	*Discrimination*	*Percent Omitted*
Number	*Type*			
138230	MC	0.41	0.43	0
138232	MC	0.72	0.43	1
138280	MC	0.44	0.21	1
138282	MC	0.70	0.48	1
138295	MC	0.50	0.33	1
153427	MC	0.65	0.39	0
164111	MC	0.70	0.37	1
164363	MC	0.63	0.50	1
164367	OR	0.23	0.61	6
166387	MC	0.59	0.22	0
167546	MC	0.47	0.34	0
167605	MC	0.51	0.38	1
167607	MC	0.54	0.34	1
191117	MC	0.69	0.31	0
191118	MC	0.43	0.33	0
191120	MC	0.23	0.15	1
191123	MC	0.67	0.40	0
191140	MC	0.78	0.34	0
191142	MC	0.49	0.30	0
191147	MC	0.68	0.40	1
191149	MC	0.60	0.38	1
191151	MC	0.46	0.37	1
191152	MC	0.58	0.40	1
191154	MC	0.44	0.40	0
191161	MC	0.69	0.41	1
191163	MC	0.74	0.31	1
191166	MC	0.50	0.31	1
196922	MC	0.54	0.36	1
228234	MC	0.53	0.22	1
248451	MC	0.58	0.44	0
258239	MC	0.59	0.39	1
470892	OR	0.26	0.59	11
471225	MC	0.36	0.31	0
473541	MC	0.58	0.29	0
473557	MC	0.69	0.40	0
473562	MC	0.48	0.40	1
473639	MC	0.43	0.36	1
473753	MC	0.71	0.43	1
473802	MC	0.58	0.39	1
477967	MC	0.62	0.38	1
477988	MC	0.53	0.42	1
546738	OR	0.34	0.58	7
550207	OR	0.33	0.59	9
550251	MC	0.57	0.23	0

Table H-4. 2018–19 MEA Science: Item-Level Score Distributions for Constructed-Response Items—by Subject and Grade

Grade	Item Number	Total Possible Points	Percent of Students at Score Point				
			0	*1*	*2*	*3*	*4*
5	165328	4	18.92	23.00	41.89	12.62	1.00
	191018	4	17.00	21.87	28.16	21.17	8.13
	191047	4	8.14	16.09	26.01	27.53	19.86
	468930	4	15.45	24.64	29.92	22.28	6.54
8	65117	4	14.95	18.05	37.21	14.86	12.86
	469354	4	12.08	11.29	61.43	8.35	5.66
	469905	4	15.08	39.19	15.13	13.84	13.08
	470376	4	10.64	11.66	40.26	24.47	9.50
11	164367	4	40.16	26.65	19.05	6.99	1.42
	470892	4	23.54	35.22	24.23	5.20	0.82
	546738	4	16.95	34.52	28.68	9.61	3.70
	550207	4	22.79	26.23	24.37	13.03	4.79

APPENDIX I—DIFFERENTIAL ITEM FUNCTIONING RESULTS

Table I-1. 2018–19 MEA Science: Number of Items Classified as "Low" or "High" DIF Overall and by Group Favored

Grade	Group		Item Type	Number of Items	Number "Low"			Number "High"		
					Total	Favoring		Total	Favoring	
	Reference	Focal				Reference	Focal		Reference	Focal
5	Male	Female	MC	32	6	2	4	0	0	0
			CR	4	0	0	0	0	0	0
	No Disability	Disability	MC	32	5	5	0	0	0	0
			CR	4	0	0	0	0	0	0
	Non-EconDis	EconDis	MC	32	0	0	0	0	0	0
			CR	4	0	0	0	0	0	0
	Non-LEP	LEP	MC	32	6	6	0	1	1	0
			CR	4	0	0	0	0	0	0
	White	Black	MC	32	3	3	0	1	1	0
			CR	4	2	2	0	0	0	0
		Hispanic	MC	32	4	2	2	0	0	0
			CR	4	0	0	0	0	0	0
8	Male	Female	MC	40	9	7	2	1	1	0
			CR	4	1	0	1	1	0	1
	No Disability	Disability	MC	40	1	1	0	0	0	0
			CR	4	2	2	0	0	0	0
	Non-EconDis	EconDis	MC	40	0	0	0	0	0	0
			CR	4	0	0	0	0	0	0
	Non-LEP	LEP	MC	40	11	11	0	1	1	0
			CR	4	0	0	0	0	0	0
	White	Black	MC	40	6	6	0	0	0	0
			CR	4	1	0	1	0	0	0
		Hispanic	MC	40	1	1	0	0	0	0
			CR	4	0	0	0	0	0	0
11	Male	Female	MC	40	7	4	3	2	2	0
			CR	4	0	0	0	0	0	0
	No Disability	Disability	MC	40	5	3	2	0	0	0
			CR	4	2	2	0	0	0	0
	Non-EconDis	EconDis	MC	40	0	0	0	0	0	0
			CR	4	0	0	0	0	0	0
	Non-LEP	LEP	MC	40	7	4	3	3	3	0
			CR	4	2	2	0	0	0	0
	White	Asian	MC	40	7	4	3	0	0	0
			CR	4	0	0	0	0	0	0
		Black	MC	40	3	3	0	0	0	0
			CR	4	1	1	0	0	0	0
		Hispanic	MC	40	1	1	0	0	0	0
			CR	4	0	0	0	0	0	0

MC = Multiple-Choice; CR = Constructed-Response.

APPENDIX J—ITEM RESPONSE THEORY PARAMETERS

Table J-1. 2018–19 MEA Science: IRT Parameters for Dichotomous Items—Science Grade 5

Item Number	*a*	*S.E. (a)*	*b*	*S.E. (b)*	*c*	*S.E.(c)*
96271	0.58517	0.03399	-0.60055	0.14264	0.37326	0.03844
165111A	0.65931	0.04362	0.69597	0.07348	0.40960	0.01834
191027A	0.74349	0.02938	-1.49377	0.10552	0.21099	0.04746
467335	0.64195	0.02778	-0.52102	0.08732	0.19719	0.03109
169229A	0.64004	0.03083	-0.24218	0.08754	0.25905	0.02824
138269A	0.31128	0.01594	-1.37405	0.21540	0.11125	0.04849
165321A	0.34390	0.02376	0.44287	0.15859	0.11166	0.03721
138205A	0.60080	0.02391	-0.70672	0.08847	0.12874	0.03259
138266A	0.79691	0.03453	0.17408	0.05090	0.27282	0.01776
191013A	0.51127	0.04046	1.52898	0.06724	0.22161	0.01846
138289A	0.34020	0.01311	-2.66206	0.10033	0.00000	0.00000
166842A	0.52173	0.04664	1.86015	0.07015	0.24578	0.01647
228155	0.27596	0.02212	-1.24099	0.43805	0.22001	0.07909
248566	0.27637	0.01061	-0.76055	0.04912	0.00000	0.00000
96275	0.54363	0.02271	-1.08534	0.11945	0.13706	0.04237
96581	0.64877	0.01809	-0.93904	0.04953	0.04670	0.01904

Item Number	*a*	*S.E. (a)*	*b*	*S.E. (b)*	*c*	*S.E.(c)*
191032A	0.70258	0.02618	-0.01288	0.05000	0.11984	0.01908
228134	0.84706	0.03253	-0.20383	0.05098	0.24352	0.01990
166403A	0.89289	0.03736	0.55572	0.03550	0.24688	0.01281
466901	0.60111	0.02408	-0.26998	0.07150	0.11206	0.02554
169211A	0.60034	0.01893	-1.88833	0.08246	0.07503	0.03293
248521	0.55015	0.02187	-1.59560	0.13228	0.13996	0.05014
96274	0.44430	0.02051	-1.89583	0.20142	0.16633	0.06396
97064	0.70794	0.03994	0.82810	0.05228	0.30458	0.01556
547304	0.54436	0.02975	0.60285	0.07043	0.16080	0.02194
467356	0.36540	0.02885	0.71789	0.15281	0.14901	0.03676
138268A	0.65896	0.02624	-0.06613	0.05949	0.13276	0.02180
166843A	0.65105	0.02177	-0.57932	0.06020	0.07960	0.02356
191015A	1.19408	0.03498	-0.38640	0.02831	0.16001	0.01450
467311	1.27260	0.03748	-0.23254	0.02517	0.17591	0.01281
164093A	0.81803	0.03892	1.01234	0.03431	0.21378	0.01131
191043A	0.57300	0.02592	-0.30054	0.08956	0.15388	0.03006

Table J-2. 2018–19 MEA Science: IRT Parameters for Polytomous Items—Science Grade 5

Item Number	*a*	*S.E. (a)*	*b*	*S.E. (b)*	*D0*	*S.E. (D0)*	*D1*	*S.E. (D1)*	*D2*	*S.E. (D2)*	*D3*	*S.E. (D3)*	*D4*	*S.E. (D4)*
165328A	0.51778	0.00470	1.60514	0.01859	3.23391	0.02579	1.80150	0.02152	-0.89136	0.03005	-4.14403	0.09999	0.00000	0.00000
191018A	0.65620	0.00590	0.54187	0.01437	1.96392	0.02129	0.77596	0.01757	-0.56601	0.01881	-2.17387	0.02995	0.00000	0.00000
191047A	0.52558	0.00474	-0.41118	0.01773	2.24911	0.03316	0.78648	0.02345	-0.69129	0.02104	-2.34431	0.02591	0.00000	0.00000
199466	0.80881	0.00715	0.47043	0.01181	1.96378	0.01936	0.69849	0.01488	-0.55277	0.01590	-2.10951	0.02726	0.00000	0.00000

Table J-3. 2018–19 MEA Science: IRT Parameters for Dichotomous Items—Science Grade 8

Item Number	*a*	*S.E. (a)*	*b*	*S.E. (b)*	*c*	*S.E.(c)*
258114	0.19413	0.01327	-2.27946	0.40662	0.12817	0.05784
470220	0.54576	0.02809	-0.08875	0.10074	0.20961	0.03050
258147	0.69916	0.03152	-0.26019	0.07687	0.28405	0.02575
470447	0.72031	0.03153	-0.26304	0.07230	0.27838	0.02489
191086A	0.89167	0.03515	0.13399	0.04311	0.27431	0.01621
165644A	0.53532	0.02372	-0.21106	0.08800	0.12318	0.02877
97788	0.47039	0.03077	0.52780	0.10988	0.21521	0.02925
187328	0.47424	0.01944	-1.66524	0.14896	0.12707	0.04989
63442	1.07547	0.03424	-0.20981	0.03266	0.20083	0.01529
191077A	0.75099	0.03056	-1.01750	0.08969	0.25533	0.03586
191063A	0.72180	0.02922	0.53073	0.04116	0.14734	0.01495
191069A	0.95550	0.03759	0.54860	0.03237	0.24417	0.01214
191107A	0.33449	0.01529	-1.36243	0.17220	0.09424	0.04149
138367A	0.82140	0.03225	0.69196	0.03293	0.16113	0.01215
258140	0.65313	0.03053	0.32060	0.06157	0.21588	0.02030
165948A	0.57874	0.02741	-0.28565	0.09554	0.20264	0.03082
258144	0.90186	0.03048	-0.12959	0.03942	0.18062	0.01686
153664	0.91907	0.02495	-1.40364	0.04315	0.05696	0.02106
191074A	1.13520	0.03825	-0.78511	0.04311	0.25383	0.02176
228125	0.58413	0.01774	-1.26373	0.06755	0.05974	0.02502
191076A	0.67668	0.02267	-0.80767	0.06547	0.09618	0.02676
153851	0.67919	0.02331	-0.80387	0.06791	0.10645	0.02774
138322A	0.57293	0.02469	-0.39018	0.08857	0.14517	0.03023
64657	0.73974	0.02036	-1.65986	0.05207	0.04844	0.02116
96613	0.69916	0.02996	-0.36671	0.07568	0.24872	0.02672
228096	0.57318	0.02524	-0.42752	0.09307	0.15896	0.03150
153898	0.29814	0.01202	-2.44194	0.10146	0.00000	0.00000
470179	0.68392	0.03523	1.19165	0.04064	0.17605	0.01279
164358A	0.94614	0.02937	-0.12579	0.03333	0.13823	0.01497
165954A	0.43280	0.02409	0.39694	0.10177	0.10800	0.02852
541774	0.90538	0.03278	0.04928	0.03958	0.22187	0.01600
469789	0.55676	0.01879	-0.97938	0.07823	0.07448	0.02827
470097	0.48937	0.02994	1.21981	0.06299	0.12231	0.01903
191101A	0.47925	0.02291	0.37272	0.07679	0.08739	0.02342
191112A	1.07191	0.03771	0.06205	0.03308	0.26280	0.01404
138455A	0.65740	0.02670	-0.35282	0.07274	0.17299	0.02653
250949	0.88095	0.03978	0.80636	0.03563	0.26789	0.01202
191090A	0.35781	0.01793	0.07094	0.11809	0.07316	0.02919
469859	0.71343	0.03315	0.65739	0.04668	0.21637	0.01568
191064A	0.57372	0.02244	-0.46813	0.07942	0.10585	0.02807

Table J-4. 2018–19 MEA Science: IRT Parameters for Polytomous Items—Science Grade 8

Item Number	*a*	*S.E. (a)*	*b*	*S.E. (b)*	*D0*	*S.E. (D0)*	*D1*	*S.E. (D1)*	*D2*	*S.E. (D2)*	*D3*	*S.E. (D3)*	*D4*	*S.E. (D4)*
469354	0.63122	0.00564	0.57159	0.01659	2.57348	0.02577	1.70982	0.02067	-1.62883	0.02499	-2.65448	0.03636	0.00000	0.00000
469905	0.70203	0.00648	0.58506	0.01363	2.03573	0.02088	0.11687	0.01652	-0.61213	0.01803	-1.54046	0.02320	0.00000	0.00000
470376	0.68220	0.00602	0.13080	0.01402	1.96401	0.02361	1.15956	0.01907	-0.73599	0.01756	-2.38759	0.02705	0.00000	0.00000
65117	0.68107	0.00615	0.30958	0.01391	1.90982	0.02204	0.84977	0.01760	-0.88292	0.01844	-1.87668	0.02400	0.00000	0.00000

Table J-5. 2018–19 MEA Science: IRT Parameters for Dichotomous Items—Science Grade 11

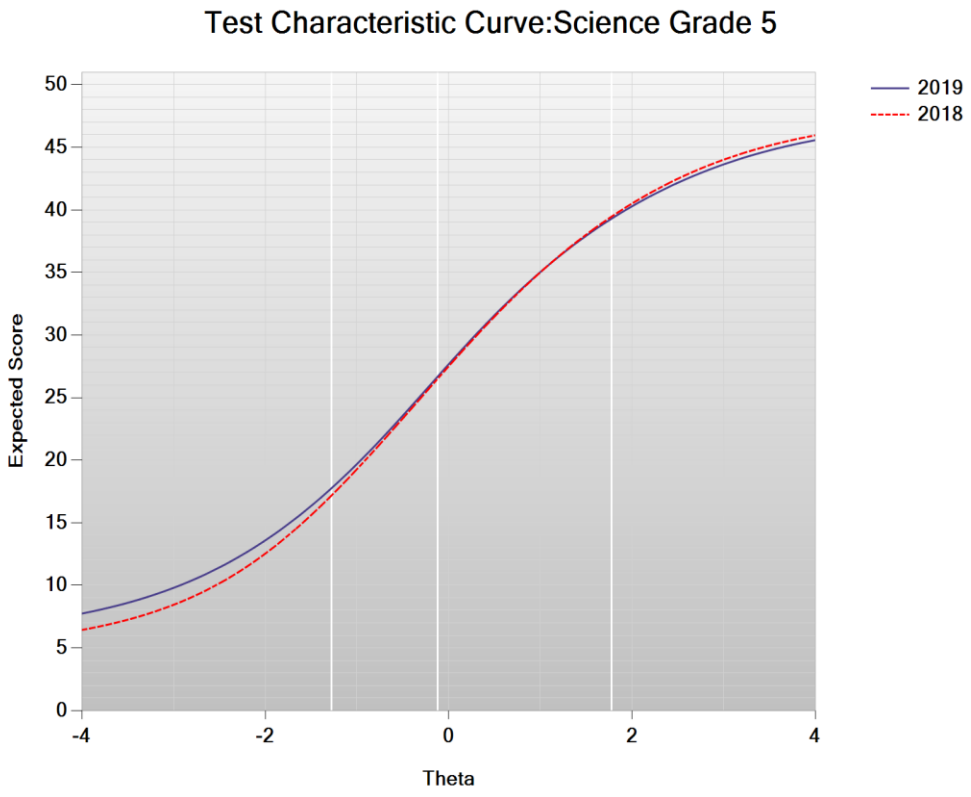
Item Number	*a*	*S.E. (a)*	*b*	*S.E. (b)*	*c*	*S.E.(c)*
191117A	0.69129	0.03961	-0.10091	0.09513	0.30706	0.03171
167546A	0.61562	0.03035	0.48499	0.05658	0.09361	0.02123
138230A	1.34529	0.04976	0.74670	0.01904	0.16894	0.00817
191140A	0.66313	0.01757	-1.25119	0.03289	0.00000	0.00000
473541	0.77027	0.04580	0.57018	0.05795	0.31868	0.01921
550251	0.34423	0.01287	-0.43964	0.03719	0.00000	0.00000
471225	1.65116	0.07686	1.15646	0.01905	0.23181	0.00636
166387A	0.33362	0.01286	-0.61579	0.04184	0.00000	0.00000
153427	1.12776	0.04818	0.19240	0.03634	0.32701	0.01525
473557	0.81799	0.03203	-0.40608	0.05998	0.14878	0.02771
248451	1.17765	0.04450	0.25684	0.02855	0.23629	0.01329
191142A	0.88804	0.04901	0.88969	0.03791	0.27238	0.01341
191154A	1.11804	0.04626	0.75268	0.02460	0.19286	0.01016
191123A	0.81482	0.03041	-0.39590	0.05558	0.12104	0.02607
191163A	0.54616	0.01573	-1.19510	0.03735	0.00000	0.00000
196922	1.03204	0.04836	0.58975	0.03381	0.28941	0.01316
228234	0.36357	0.02543	0.29520	0.16210	0.10694	0.04079
191118A	1.09667	0.05357	0.98984	0.02715	0.24230	0.00980
191120A	1.94864	0.12534	1.72729	0.02490	0.17740	0.00444
138232A	0.89477	0.02838	-0.61196	0.04514	0.08191	0.02342
477967	0.72698	0.03061	-0.15185	0.06207	0.12867	0.02612
138280A	0.77331	0.06017	1.43727	0.04834	0.30403	0.01307
258239	0.73373	0.03029	-0.04912	0.05599	0.11607	0.02365
477988	0.98027	0.03804	0.34168	0.03200	0.16939	0.01433
138282A	1.23558	0.03989	-0.31961	0.03076	0.16969	0.01742
473753	0.92152	0.03025	-0.56968	0.04586	0.09928	0.02403
167607A	0.75423	0.03830	0.43174	0.05274	0.20925	0.01997
164363A	1.12269	0.03384	-0.19808	0.02787	0.09702	0.01507
191161A	0.88373	0.03396	-0.34915	0.05329	0.16792	0.02522
164111A	0.72193	0.02826	-0.57857	0.07198	0.11585	0.03209
473802	0.73233	0.02889	-0.04532	0.05229	0.09646	0.02231
191166A	0.68654	0.03923	0.64823	0.05650	0.20220	0.02030
473562	1.37626	0.05553	0.70690	0.02130	0.24557	0.00905
138295A	0.95473	0.04774	0.74764	0.03494	0.26776	0.01310
191147A	0.74335	0.01709	-0.63940	0.02061	0.00000	0.00000
473639	1.14034	0.05028	0.86462	0.02483	0.21658	0.00970
167605A	1.17893	0.05076	0.64895	0.02657	0.26366	0.01092
191152A	1.01737	0.04241	0.29108	0.03570	0.24592	0.01532
191151A	0.93736	0.04272	0.72605	0.03173	0.19729	0.01270

Table J-6. 2018–19 MEA Science: IRT Parameters for Polytomous Items—Science Grade 11

Item Number	*a*	*S.E. (a)*	*b*	*S.E. (b)*	*D0*	*S.E. (D0)*	*D1*	*S.E. (D1)*	*D2*	*S.E. (D2)*	*D3*	*S.E. (D3)*	*D4*	*S.E. (D4)*
164367A	1.13202	0.01219	1.42618	0.00989	1.45494	0.01160	0.56284	0.01246	-0.44759	0.01883	-1.57020	0.04152	0.00000	0.00000
258242	1.00026	0.01018	1.53869	0.01061	1.97356	0.01331	0.73170	0.01347	-0.69827	0.02397	-2.00698	0.05964	0.00000	0.00000
546738	0.93734	0.00886	0.94768	0.01083	1.86819	0.01548	0.58378	0.01331	-0.70700	0.01824	-1.74497	0.03167	0.00000	0.00000
550207	0.96746	0.00953	0.90678	0.01057	1.45813	0.01387	0.54501	0.01294	-0.45316	0.01598	-1.54997	0.02741	0.00000	0.00000

APPENDIX K—TEST CHARACTERISTIC CURVES AND TEST INFORMATION FUNCTIONS

Figure K-1. 2018–19 MEA Science: Grade 5 Plots
Top: Test Characteristic Curve Bottom: Test Information Function



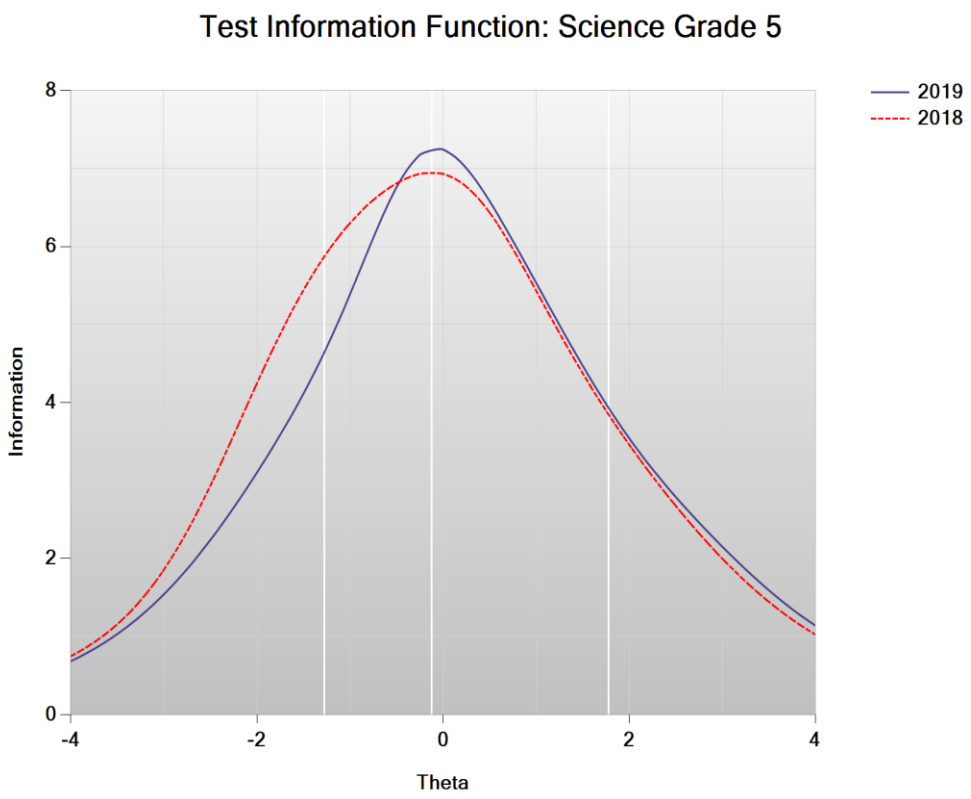
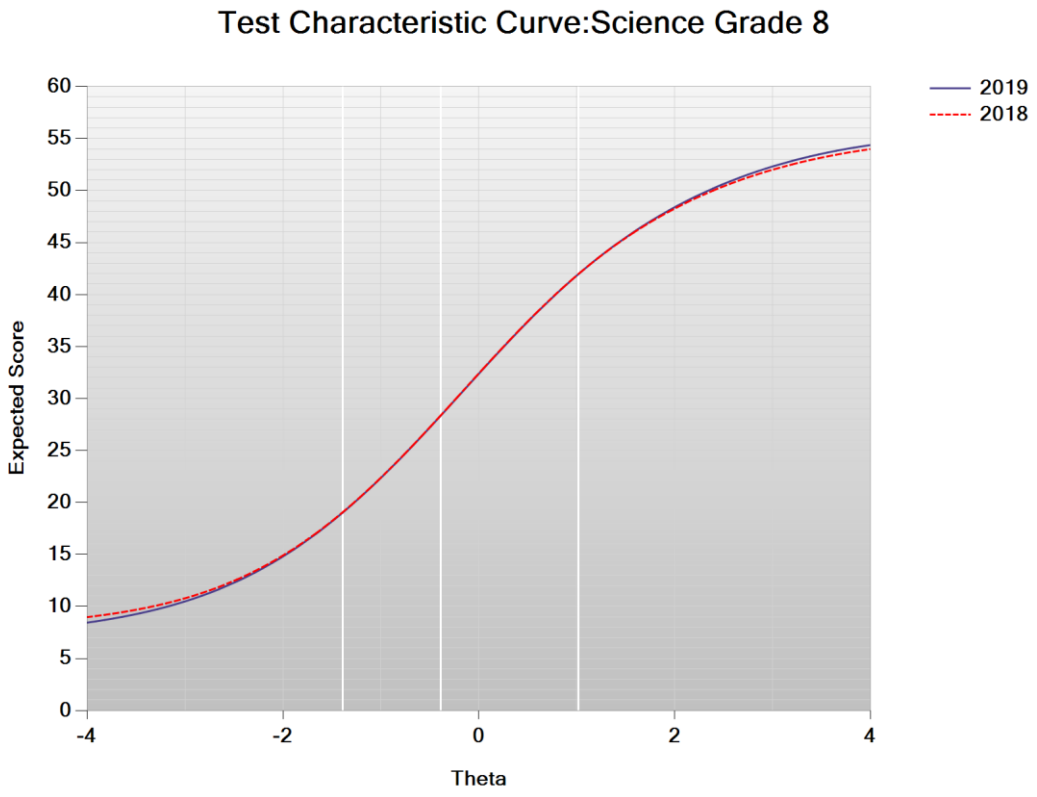


Figure K-2. 2018–19 MEA Science: Grade 8 Plots
Top: Test Characteristic Curve Bottom: Test Information Function



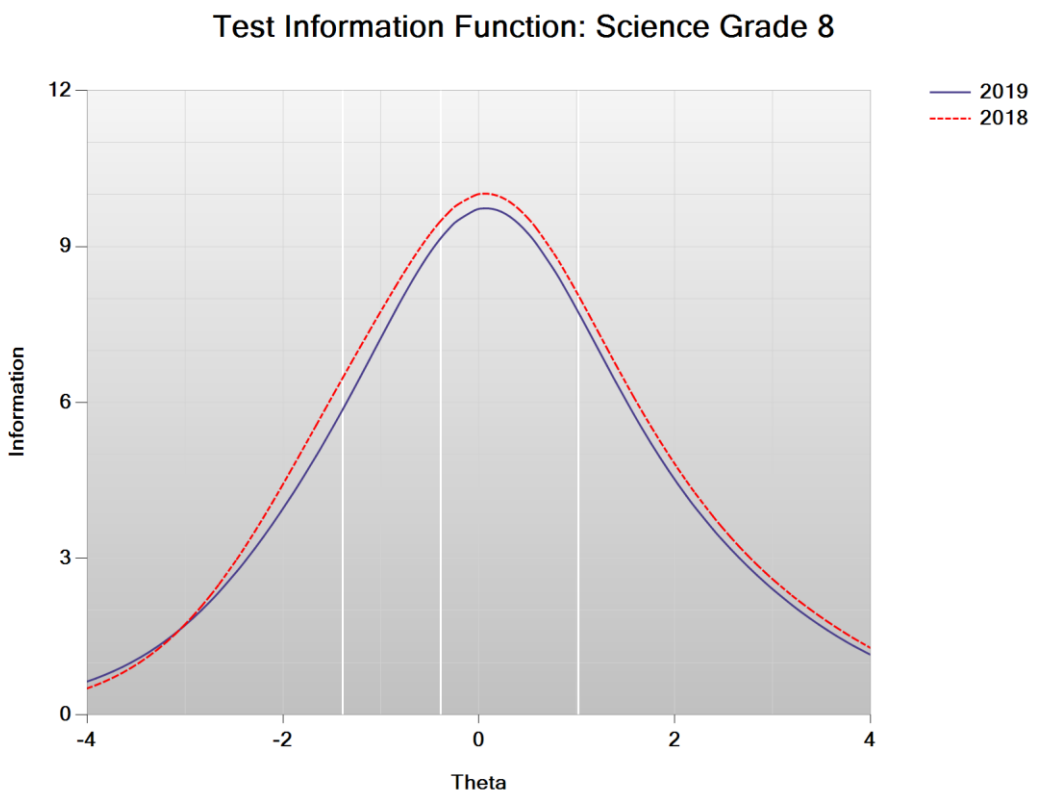
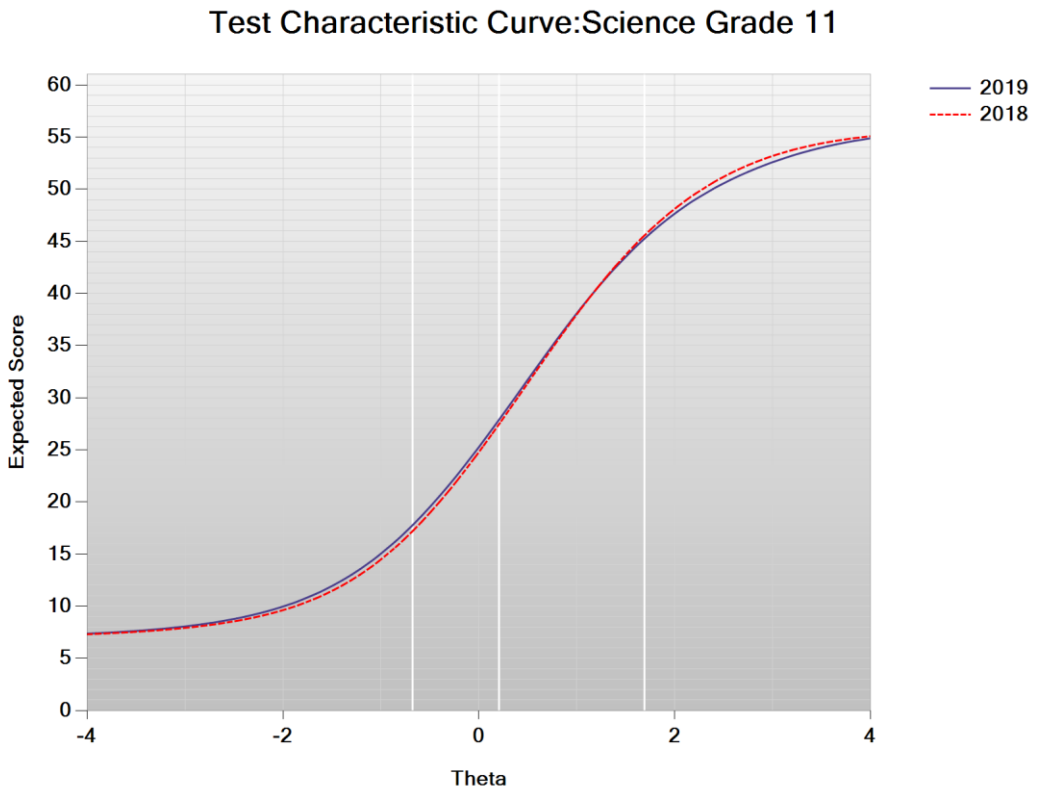
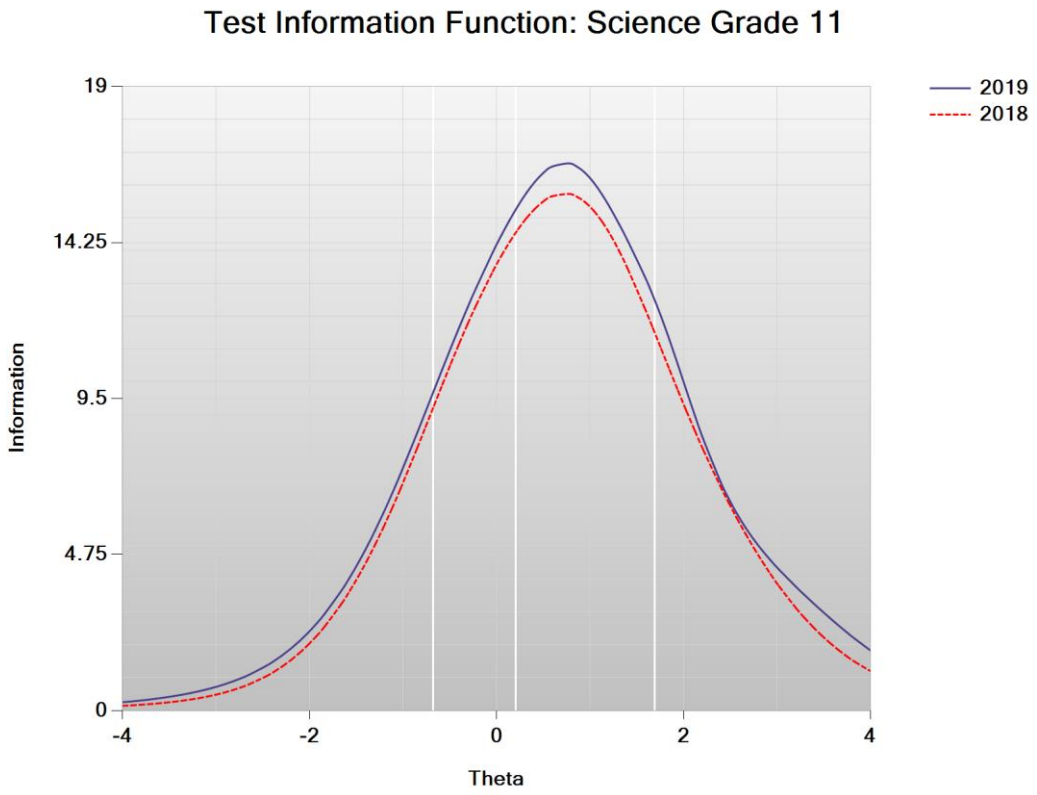


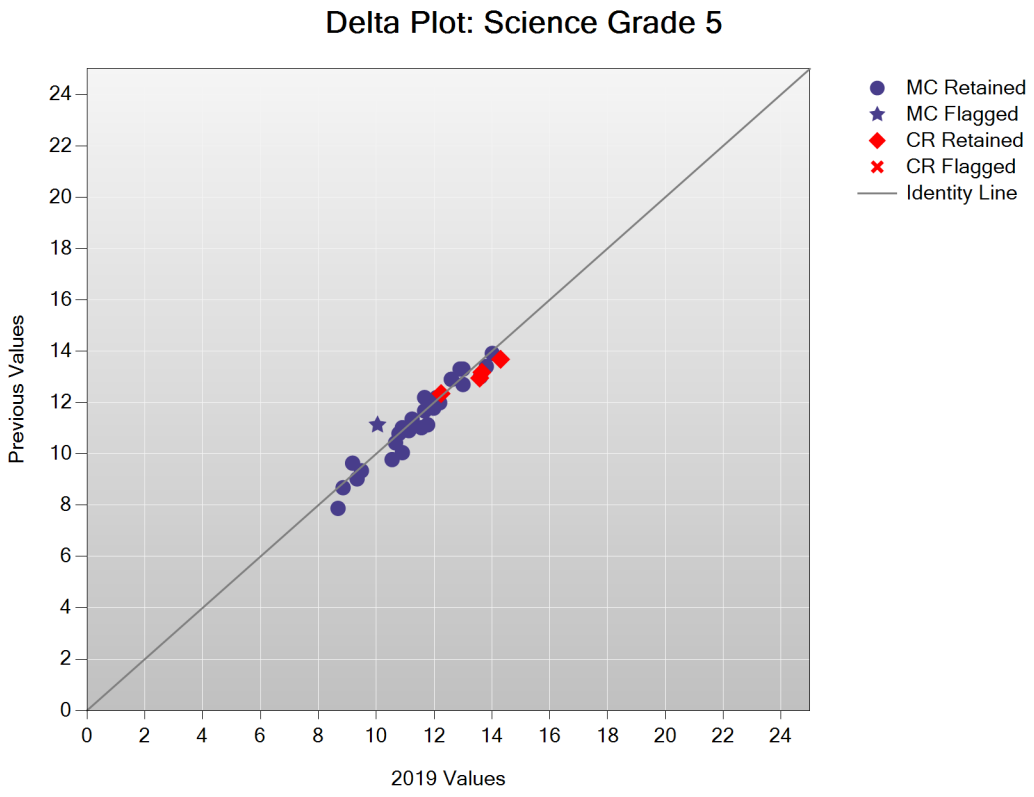
Figure K-3. 2018–19 MEA Science: Grade 11 Plots
Top: Test Characteristic Curve Bottom: Test Information Function





APPENDIX L—DELTA AND RESCORE ANALYSIS

Figure L-1. 2018–19 MEA Science: Delta Analysis Plots—Science
Top: Grade 5 Bottom: Grade 8



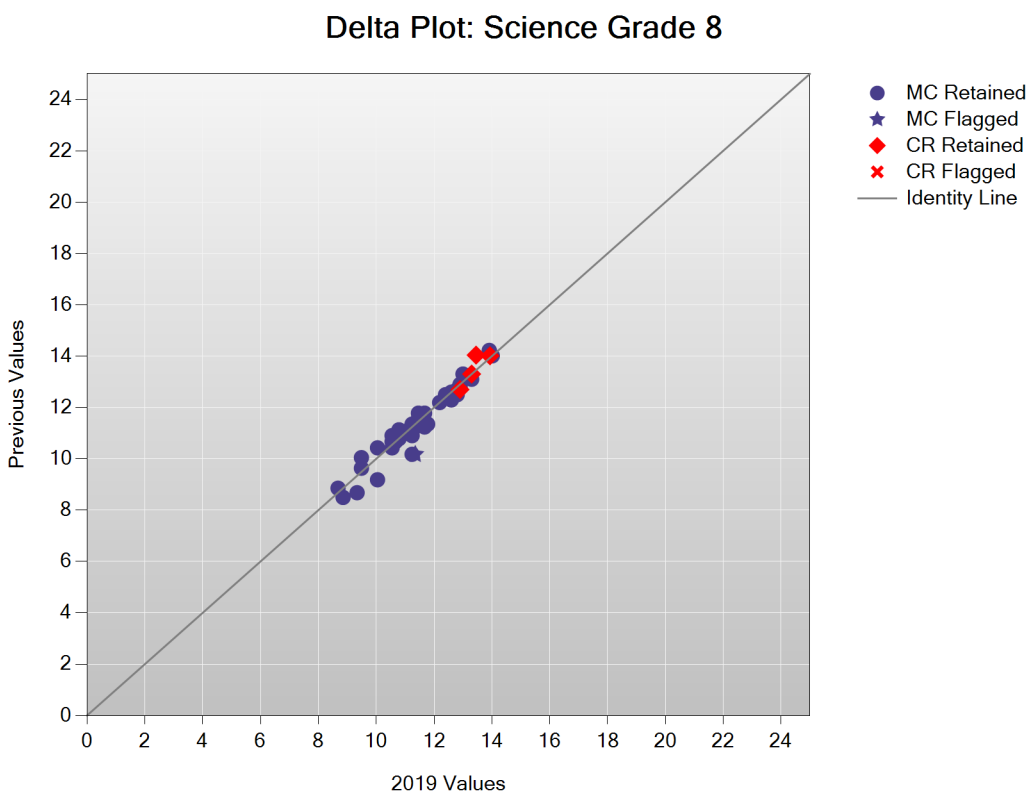


Figure L-2. 2018–19 MEA Science: Delta Analysis Plots—Science Grade 11

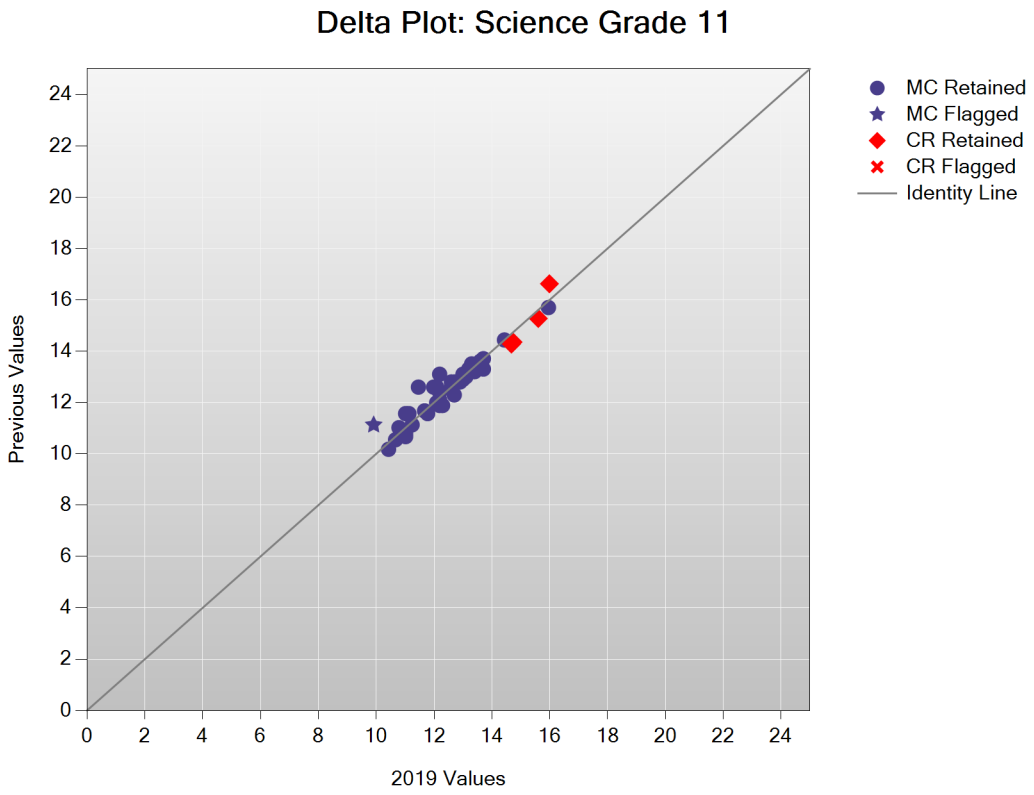


Table L-1. 2018–19 MEA Science: Delta Analysis Results—Science Grade 5

Item Number	*Mean*		*Delta*		*Discard*	*Standardized Difference*
	Old	*New*	*Old*	*New*		
138205A	0.70000	0.70000	10.90240	10.90240	False	-0.72417
138266A	0.63000	0.63000	11.67259	11.67259	False	-0.75551
138268A	0.62000	0.60000	11.77808	11.98661	False	-0.97821
138269A	0.71000	0.71000	10.78646	10.78646	False	-0.71946
138289A	0.82000	0.81000	9.33854	9.48841	False	-1.22554
164093A	0.50000	0.44000	13.00000	13.60388	False	0.56186
165111A	0.58000	0.63000	12.19243	11.67259	False	1.18301
165321A	0.47000	0.51000	13.30108	12.89972	False	0.69124
165328A	0.43250	0.37250	13.68005	14.30096	False	0.65373
166403A	0.51000	0.54000	12.89972	12.59827	False	0.33099
166842A	0.41000	0.40000	13.91018	14.01339	False	-1.23563
166843A	0.66000	0.67000	11.35015	11.24035	False	-0.32847
169211A	0.86000	0.85000	8.67872	8.85427	False	-1.22869
169229A	0.69000	0.68000	11.01660	11.12920	False	-1.15332
191013A	0.46000	0.42000	13.40173	13.80757	False	-0.16835
191015A	0.77000	0.70000	10.04461	10.90240	False	1.39879
191018A	0.48250	0.43500	13.17552	13.65463	False	0.09867
191027A	0.90000	0.86000	7.87379	8.67872	False	1.11120
191032A	0.60000	0.58000	11.98661	12.19243	False	-0.97998
191043A	0.69000	0.64000	11.01660	11.56616	False	0.27642
191047A	0.56500	0.57500	12.34537	12.24353	False	-0.39898
199466	0.50500	0.44250	12.94987	13.57854	False	0.65328
228134	0.70000	0.68000	10.90240	11.12920	False	-0.94496
228155	0.74000	0.72000	10.42662	10.66863	False	-0.90699
248521	0.84000	0.82000	9.02217	9.33854	False	-0.68383
248566	0.58000	0.59000	12.19243	12.08982	False	-0.38986
466901	0.68000	0.62000	11.12920	11.77808	False	0.65536
467335	0.69000	0.70000	11.01660	10.90240	False	-0.29831
467356	0.53000	0.50000	12.69892	13.00000	False	-0.59187
547304	0.47000	0.50000	13.30108	13.00000	False	0.31323
96271	0.68000	0.77000	11.12920	10.04461	True	3.35527
96274	0.80000	0.83000	9.63352	9.18334	False	1.02452
96581	0.79000	0.73000	9.77432	10.54875	False	1.07357
97064	0.51000	0.54000	12.89972	12.59827	False	0.33099
138205A	0.70000	0.70000	10.90240	10.90240	False	-0.72417

Table L-2. 2018–19 MEA Science: Delta Analysis Results—Science Grade 8

Item Number	Mean		Delta		Discard	Standardized Difference
	Old	New	Old	New		
138322A	0.67000	0.66000	11.24035	11.35015	False	-0.94235
138367A	0.49000	0.47000	13.10028	13.30108	False	-0.10976
138455A	0.70000	0.67000	10.90240	11.24035	False	-0.06577
153664	0.87000	0.85000	8.49444	8.85427	False	-0.55683
153851	0.70000	0.73000	10.90240	10.54875	False	0.81944
153898	0.74000	0.77000	10.42662	10.04461	False	1.05374
164358A	0.62000	0.63000	11.77808	11.67259	False	-0.43503
165644A	0.66000	0.62000	11.35015	11.77808	False	0.42062
165948A	0.66000	0.67000	11.35015	11.24035	False	-0.31332
165954A	0.55000	0.52000	12.49735	12.79939	False	0.16950
187328	0.77000	0.81000	10.04461	9.48841	False	1.87797
191063A	0.47000	0.50000	13.30108	13.00000	False	0.01786
191064A	0.76000	0.66000	10.17479	11.35015	True	3.27582
191069A	0.54000	0.54000	12.59827	12.59827	False	-1.07483
191074A	0.80000	0.81000	9.63352	9.48841	False	0.25058
191076A	0.72000	0.73000	10.66863	10.54875	False	-0.10595
191077A	0.86000	0.82000	8.67872	9.33854	False	0.74795
191086A	0.65000	0.64000	11.45872	11.56616	False	-0.89937
191090A	0.57000	0.54000	12.29450	12.59827	False	0.12766
191101A	0.51000	0.51000	12.89972	12.89972	False	-1.00184
191107A	0.72000	0.72000	10.66863	10.66863	False	-0.60956
191112A	0.64000	0.65000	11.56616	11.45872	False	-0.37551
228096	0.76000	0.67000	10.17479	11.24035	False	2.81458
228125	0.83000	0.77000	9.18334	10.04461	False	1.71639
250949	0.52000	0.51000	12.79939	12.89972	False	-0.60464
258114	0.71000	0.72000	10.78646	10.66863	False	-0.14313
258140	0.58000	0.58000	12.19243	12.19243	False	-0.97849
258144	0.62000	0.65000	11.77808	11.45872	False	0.46338
258147	0.68000	0.71000	11.12920	10.78646	False	0.71871
469354	0.39750	0.45500	14.03929	13.45215	False	1.04080
469789	0.74000	0.73000	10.42662	10.54875	False	-1.06401
469859	0.53000	0.52000	12.69892	12.79939	False	-0.62843
469905	0.40000	0.40750	14.01339	13.93592	False	-1.05763
470097	0.40000	0.40000	14.01339	14.01339	False	-0.73222
470179	0.38000	0.41000	14.22192	13.91018	False	-0.16028
470220	0.65000	0.64000	11.45872	11.56616	False	-0.89937
470376	0.53000	0.51000	12.69892	12.89972	False	-0.20693
470447	0.71000	0.71000	10.78646	10.78646	False	-0.63809
541774	0.67000	0.63000	11.24035	11.67259	False	0.41214
63442	0.68000	0.68000	11.12920	11.12920	False	-0.72107
64657	0.85000	0.86000	8.85427	8.67872	False	0.56712
65117	0.47000	0.47000	13.30108	13.30108	False	-0.90467
96613	0.71000	0.71000	10.78646	10.78646	False	-0.63809
97788	0.55000	0.56000	12.49735	12.39612	False	-0.62706

Table L-3. 2018–19 MEA Science: Delta Analysis Results—Science Grade 11

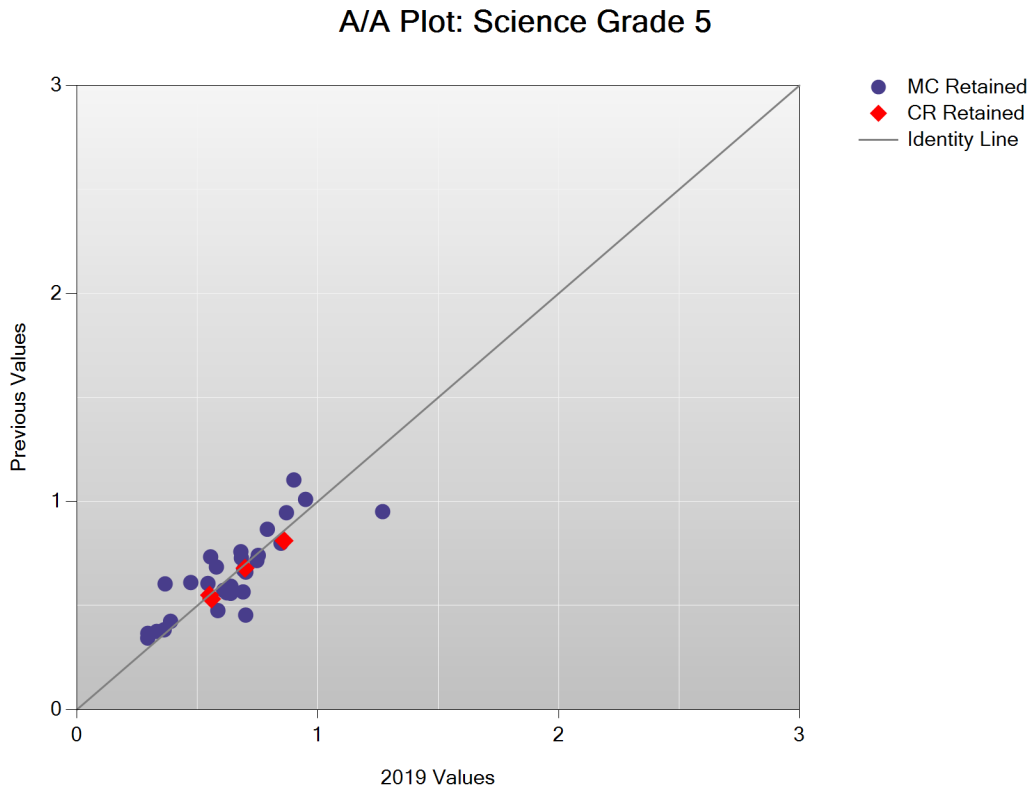
Item Number	Mean		Delta		Discard	Standardized Difference
	Old	New	Old	New		
138232A	0.73000	0.72000	10.54875	10.66863	False	-0.21556
138280A	0.44000	0.44000	13.60388	13.60388	False	-1.12001
138282A	0.71000	0.70000	10.78646	10.90240	False	-0.26554
138295A	0.51000	0.50000	12.89972	13.00000	False	-0.63617
153427	0.54000	0.65000	12.59827	11.45872	False	2.98988
164111A	0.71000	0.70000	10.78646	10.90240	False	-0.26554
164363A	0.63000	0.63000	11.67259	11.67259	False	-0.83555
164367A	0.18250	0.22750	16.62352	15.98842	False	1.67048
166387A	0.60000	0.59000	11.98661	12.08982	False	-0.49056
167546A	0.45000	0.47000	13.50265	13.30108	False	-0.43261
167605A	0.52000	0.51000	12.79939	12.89972	False	-0.62116
167607A	0.54000	0.54000	12.59827	12.59827	False	-0.97189
191117A	0.64000	0.69000	11.56616	11.01660	False	0.60136
191118A	0.47000	0.43000	13.30108	13.70550	False	0.45765
191120A	0.25000	0.23000	15.69796	15.95539	False	-0.45259
191123A	0.68000	0.67000	11.12920	11.24035	False	-0.33420
191140A	0.68000	0.78000	11.12920	9.91123	True	3.07081
191142A	0.50000	0.49000	13.00000	13.10028	False	-0.65094
191147A	0.64000	0.68000	11.56616	11.12920	False	0.17450
191149A	0.54000	0.60000	12.59827	11.98661	False	0.98874
191151A	0.48000	0.46000	13.20061	13.40173	False	-0.29821
191152A	0.49000	0.58000	13.10028	12.19243	False	2.18550
191154A	0.44000	0.44000	13.60388	13.60388	False	-1.12001
191161A	0.72000	0.69000	10.66863	11.01660	False	0.63138
191163A	0.76000	0.74000	10.17479	10.42662	False	0.33969
191166A	0.49000	0.50000	13.10028	13.00000	False	-0.87584
196922	0.52000	0.54000	12.79939	12.59827	False	-0.53788
228234	0.52000	0.53000	12.79939	12.69892	False	-0.91944
248451	0.61000	0.58000	11.88272	12.19243	False	0.30751
258239	0.54000	0.59000	12.59827	12.08982	False	0.59750
258242	0.28500	0.25750	15.27221	15.60429	False	-0.10689
471225	0.36000	0.36000	14.43384	14.43384	False	-1.05954
473541	0.56000	0.58000	12.39612	12.19243	False	-0.58751
473557	0.71000	0.69000	10.78646	11.01660	False	0.16737
473562	0.47000	0.48000	13.30108	13.20061	False	-0.84555
473639	0.43000	0.43000	13.70550	13.70550	False	-1.13498
473753	0.69000	0.71000	11.01660	10.78646	False	-0.69047
473802	0.59000	0.58000	12.08982	12.19243	False	-0.50805
477967	0.64000	0.62000	11.56616	11.77808	False	-0.01656
477988	0.57000	0.53000	12.29450	12.69892	False	0.60591
546738	0.37500	0.33750	14.27456	14.67718	False	0.30747
550207	0.36750	0.33250	14.35393	14.73208	False	0.20301
550251	0.61000	0.57000	11.88272	12.29450	False	0.69447

Table L-4. 2018–19 MEA Science: Rescore Analysis Results by Subject and Grade

Grade	*Item Number*	*Maximum*	*Mean*		*Standard Deviation*		*Effect Size*	*Discard*
			Old	*New*	*Old*	*New*		
5	165328A	4	1.67317	1.61463	1.02685	0.94598	-0.05701	False
	191018A	4	2.07843	2.13725	1.20905	1.07858	0.04865	False
	191047A	4	2.16585	2.21463	1.13406	1.11266	0.04301	False
	468930	4	2.07353	1.73529	1.06881	1.07755	-0.31646	False
8	469354	4	1.71078	1.85294	0.98245	0.97654	0.14470	False
	469905	4	1.67805	1.76585	1.10425	1.20611	0.07952	False
	470376	4	2.22927	2.32683	0.89723	1.05511	0.10874	False
	65117	4	1.94146	2.08780	1.04623	1.14279	0.13988	False
11	164367A	4	0.85854	1.09268	0.99237	1.03668	0.23595	False
	470892	4	1.16832	0.90594	1.13794	0.83818	-0.23057	False
	546738	4	1.63902	1.63902	1.10100	1.12304	0.00000	False
	550207	4	1.70244	1.60000	1.20629	1.20294	-0.08492	False

APPENDIX M–α-PLOTS AND b-PLOTS

Figure M-1. 2018–19 MEA Science: Grade 5 Plots
Top: α-Plot Bottom: b-Plot



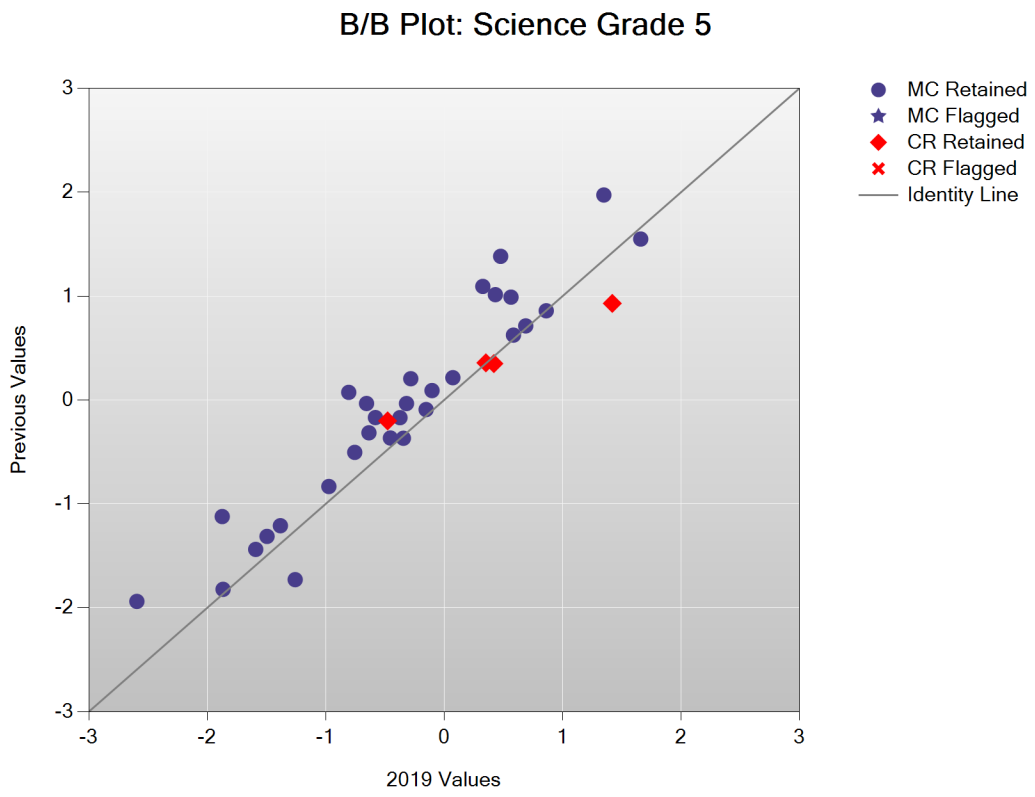
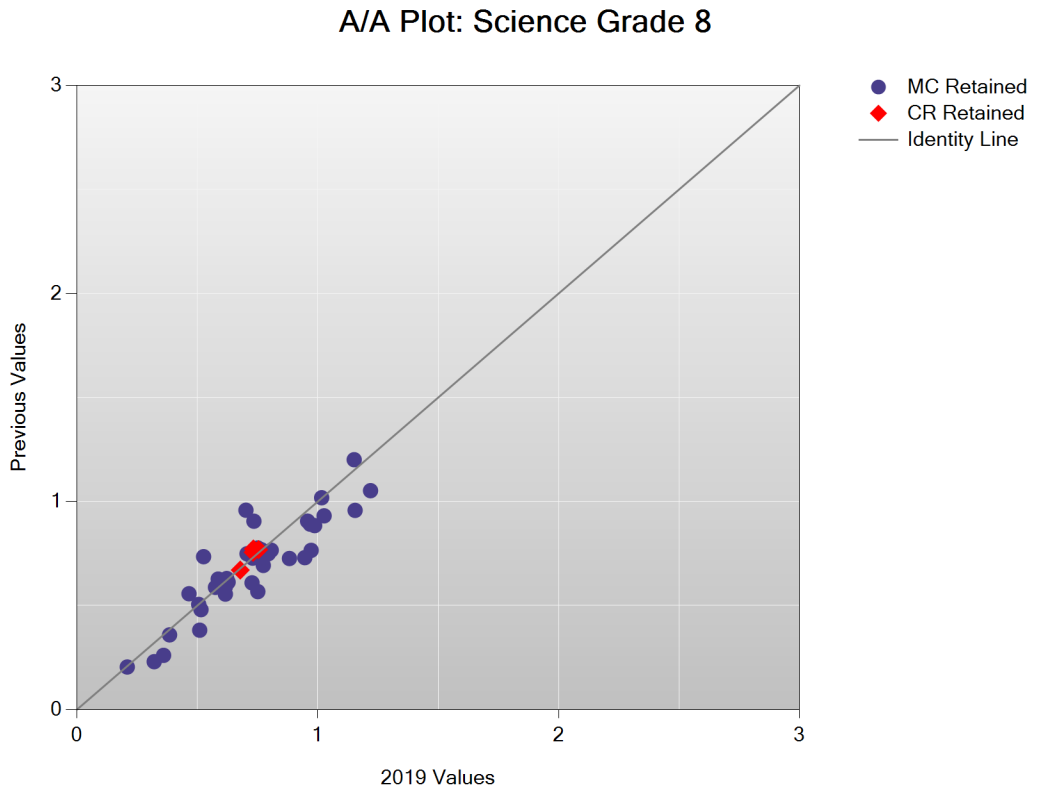


Figure M-2. 2018–19 MEA Science: Grade 8 Plots
Top: *α*-Plot Bottom: *b*-Plot



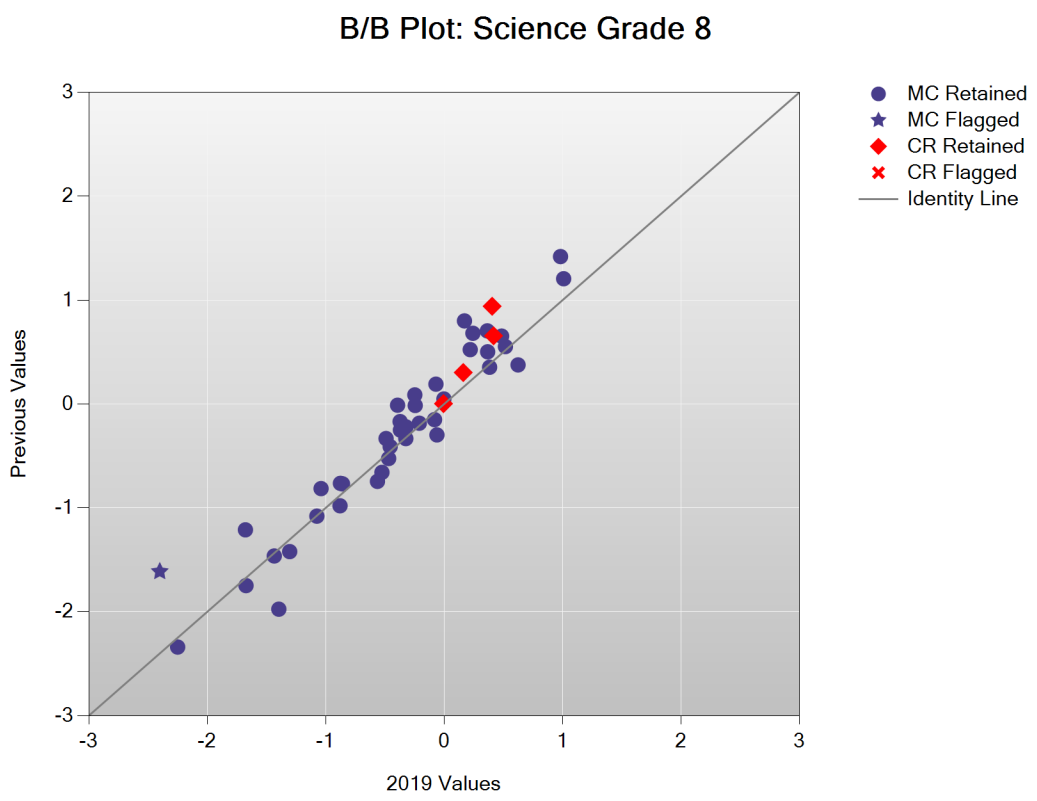
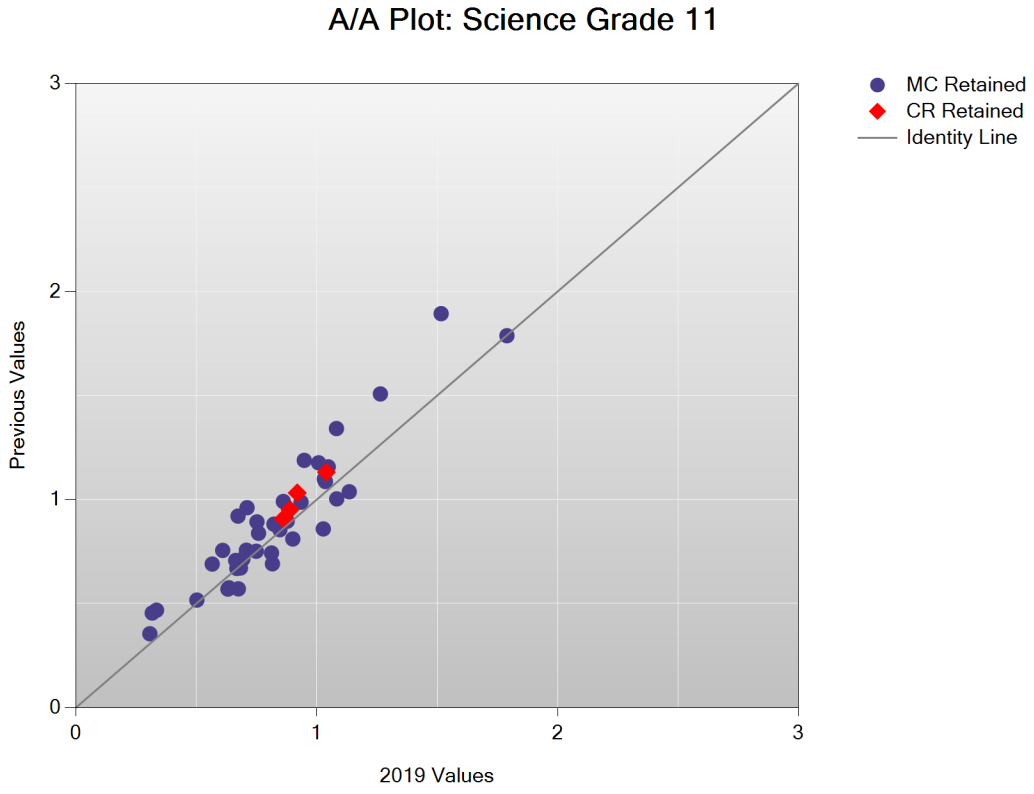
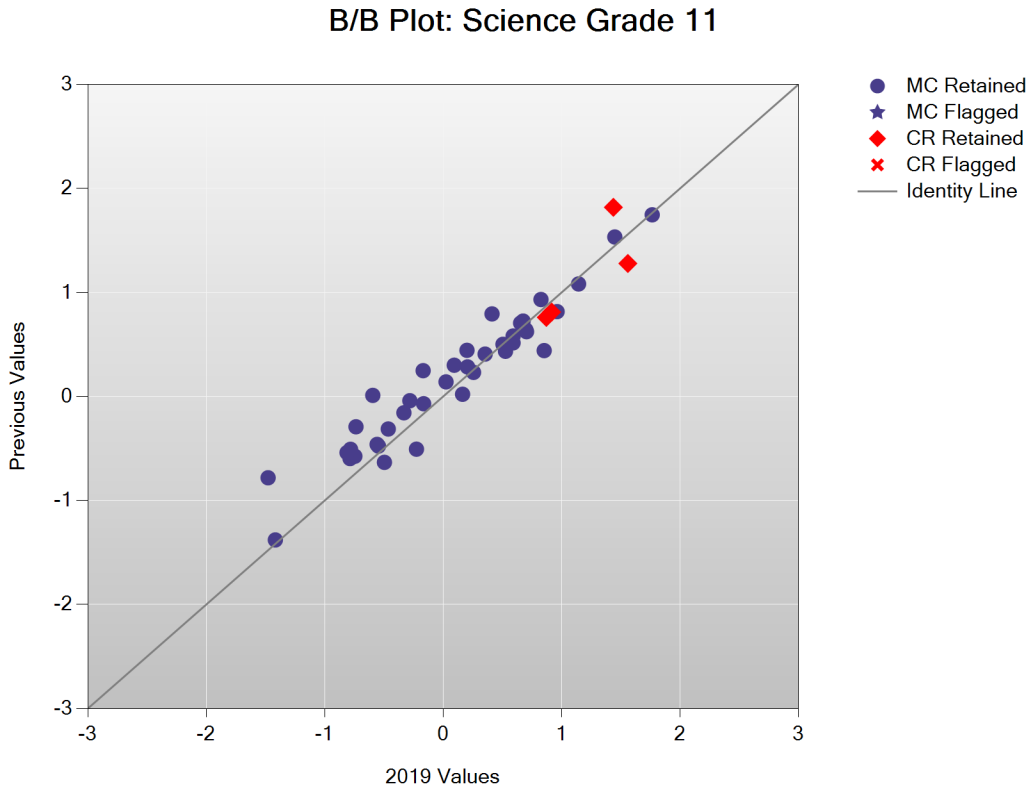


Figure M-3. 2018–19 MEA Science: Grade11 Plots
Top: *α*-Plot Bottom: *b*-Plot





APPENDIX N—RAW TO SCALED SCORE LOOK-UP TABLES

Table N-1. 2018–19 MEA Science: Raw to Scaled Score Look-Up Table—Science Grade 5

Raw Score	This Year			Last Year		
	Scaled Score	Standard Error	Performance Level	Scaled Score	Standard Error	Performance Level
0	500	10.0	1	500	10.0	1
1	500	10.0	1	500	10.0	1
2	500	10.0	1	500	10.0	1
3	500	10.0	1	500	10.0	1
4	500	10.0	1	500	10.0	1
5	500	10.0	1	500	10.0	1
6	500	10.0	1	500	10.0	1
7	500	10.0	1	504	10.0	1
8	502	10.0	1	510	8.0	1
9	508	10.0	1	514	8.0	1
10	512	8.0	1	516	6.0	1
11	516	8.0	1	518	6.0	1
12	518	6.0	1	522	6.0	1
13	520	6.0	1	524	4.0	1
14	524	6.0	1	524	4.0	1
15	526	6.0	1	526	4.0	1
16	526	6.0	1	528	4.0	1
17	528	6.0	1	528	4.0	1
18	530	4.0	2	530	4.0	2
19	532	4.0	2	532	4.0	2
20	534	4.0	2	534	4.0	2
21	534	4.0	2	536	4.0	2
22	536	4.0	2	536	4.0	2
23	538	4.0	2	538	4.0	2
24	538	4.0	2	538	4.0	2
25	540	4.0	2	540	4.0	2
26	540	4.0	2	540	4.0	2
27	542	4.0	3	542	4.0	3
28	544	4.0	3	544	4.0	3
29	546	4.0	3	546	4.0	3
30	546	4.0	3	546	4.0	3
31	548	4.0	3	548	4.0	3
32	550	4.0	3	550	4.0	3
33	550	4.0	3	550	4.0	3
34	552	4.0	3	552	4.0	3
35	554	4.0	3	554	4.0	3
36	556	4.0	3	556	4.0	3
37	558	4.0	3	558	4.0	3
38	560	4.0	3	560	6.0	3
39	560	6.0	3	560	6.0	3
40	564	6.0	4	564	6.0	4
41	566	6.0	4	566	6.0	4
42	570	6.0	4	568	6.0	4
43	572	6.0	4	572	6.0	4
44	576	8.0	4	574	8.0	4
45	580	8.0	4	580	8.0	4
46	580	10.0	4	580	10.0	4
47	580	10.0	4	580	10.0	4
48	580	10.0	4	580	10.0	4

Table N-2. 2018–19 MEA Science: Raw to Scaled Score Look-Up Table—Science Grade 8

Raw Score	This Year			Last Year		
	Scaled Score	Standard Error	Performance Level	Scaled Score	Standard Error	Performance Level
0	800	10.0	1	800	10.0	1
1	800	10.0	1	800	10.0	1
2	800	10.0	1	800	10.0	1
3	800	10.0	1	800	10.0	1
4	800	10.0	1	800	10.0	1
5	800	10.0	1	800	10.0	1
6	800	10.0	1	800	10.0	1
7	800	10.0	1	800	10.0	1
8	800	10.0	1	800	10.0	1
9	800	10.0	1	800	10.0	1
10	802	10.0	1	800	10.0	1
11	808	10.0	1	806	10.0	1
12	812	8.0	1	810	8.0	1
13	814	8.0	1	814	8.0	1
14	818	8.0	1	816	8.0	1
15	820	8.0	1	820	6.0	1
16	822	6.0	1	822	6.0	1
17	824	6.0	1	824	6.0	1
18	826	6.0	1	826	6.0	1
19	826	6.0	1	826	6.0	1
20	830	6.0	2	830	6.0	2
21	832	6.0	2	832	6.0	2
22	832	6.0	2	832	6.0	2
23	834	6.0	2	834	6.0	2
24	836	6.0	2	836	4.0	2
25	838	4.0	2	838	4.0	2
26	838	4.0	2	838	4.0	2
27	840	4.0	2	840	4.0	2
28	840	4.0	2	840	4.0	2
29	842	4.0	3	842	4.0	3
30	844	4.0	3	844	4.0	3
31	846	4.0	3	846	4.0	3
32	848	4.0	3	846	4.0	3
33	848	4.0	3	848	4.0	3
34	850	4.0	3	850	4.0	3
35	852	4.0	3	852	4.0	3
36	852	4.0	3	852	4.0	3
37	854	4.0	3	854	4.0	3
38	856	4.0	3	856	4.0	3
39	858	4.0	3	858	4.0	3
40	858	4.0	3	858	4.0	3
41	860	4.0	3	860	4.0	3
42	862	6.0	4	862	6.0	4
43	864	6.0	4	864	6.0	4
44	866	6.0	4	866	6.0	4
45	868	6.0	4	868	6.0	4

continued

Raw Score	This Year			Last Year		
	Scaled Score	Standard Error	Performance Level	Scaled Score	Standard Error	Performance Level
46	870	6.0	4	870	6.0	4
47	872	6.0	4	872	6.0	4
48	874	6.0	4	876	6.0	4
49	878	6.0	4	878	6.0	4
50	880	8.0	4	880	8.0	4
51	880	8.0	4	880	8.0	4
52	880	8.0	4	880	8.0	4
53	880	10.0	4	880	10.0	4
54	880	10.0	4	880	10.0	4
55	880	10.0	4	880	10.0	4
56	880	10.0	4	880	10.0	4

Table N-3. 2018–19 MEA Science: Raw to Scaled Score Look-Up Table—Science Grade 11

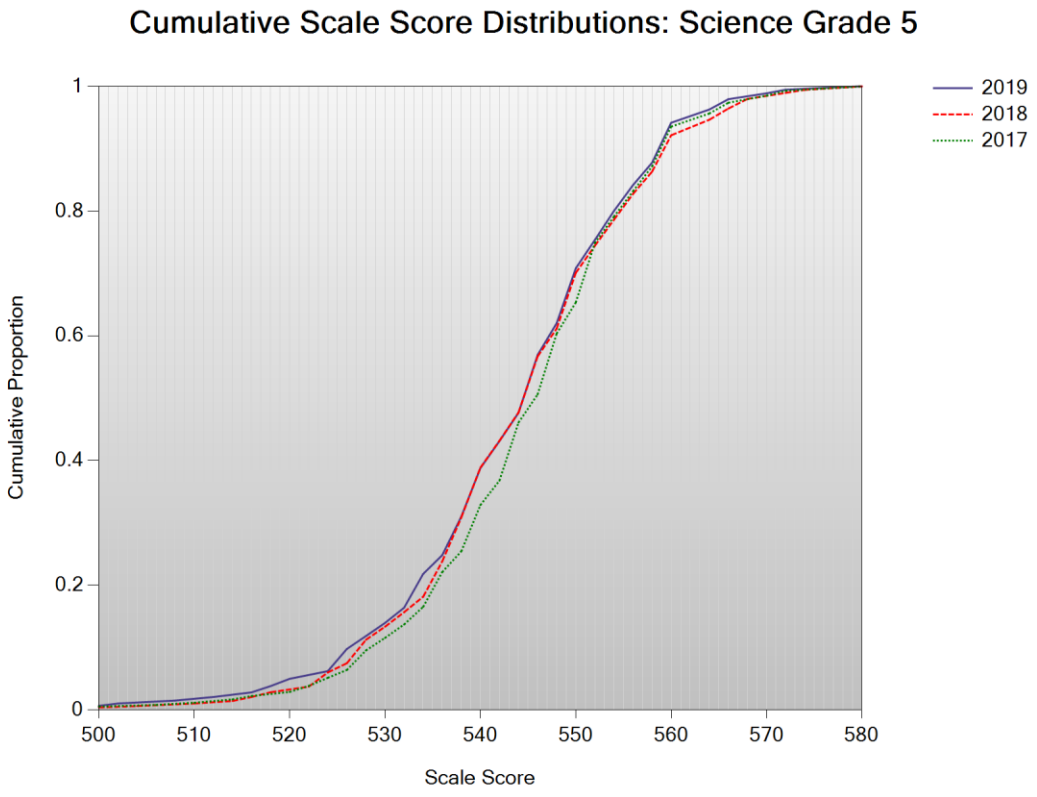
Raw Score	This Year			Last Year		
	Scaled Score	Standard Error	Performance Level	Scaled Score	Standard Error	Performance Level
0	1100	10.0	1	1,100	10.0	1
1	1100	10.0	1	1,100	10.0	1
2	1100	10.0	1	1,100	10.0	1
3	1100	10.0	1	1,100	10.0	1
4	1100	10.0	1	1,100	10.0	1
5	1100	10.0	1	1,100	10.0	1
6	1100	10.0	1	1,100	10.0	1
7	1100	10.0	1	1,100	10.0	1
8	1100	10.0	1	1,100	10.0	1
9	1108	10.0	1	1,108	10.0	1
10	1112	8.0	1	1,114	8.0	1
11	1116	8.0	1	1,118	8.0	1
12	1120	6.0	1	1,120	6.0	1
13	1122	6.0	1	1,122	6.0	1
14	1124	6.0	1	1,124	6.0	1
15	1126	4.0	1	1,126	4.0	1
16	1128	4.0	1	1,128	4.0	1
17	1128	4.0	1	1,128	4.0	1
18	1130	4.0	2	1,132	4.0	2
19	1132	4.0	2	1,132	4.0	2
20	1134	4.0	2	1,134	4.0	2
21	1134	4.0	2	1,136	4.0	2
22	1136	4.0	2	1,136	4.0	2
23	1136	4.0	2	1,138	4.0	2
24	1138	4.0	2	1,138	4.0	2
25	1138	4.0	2	1,140	4.0	2
26	1140	4.0	2	1,140	4.0	2
27	1140	4.0	2	1,140	4.0	2
28	1142	4.0	3	1,142	4.0	3
29	1144	4.0	3	1,144	4.0	3

continued

Raw Score	This Year			Last Year		
	Scaled Score	Standard Error	Performance Level	Scaled Score	Standard Error	Performance Level
30	1144	4.0	3	1,144	4.0	3
31	1146	4.0	3	1,146	4.0	3
32	1146	4.0	3	1,146	4.0	3
33	1148	4.0	3	1,148	4.0	3
34	1148	4.0	3	1,148	4.0	3
35	1150	4.0	3	1,150	4.0	3
36	1150	4.0	3	1,150	4.0	3
37	1152	4.0	3	1,152	4.0	3
38	1152	4.0	3	1,152	4.0	3
39	1154	4.0	3	1,154	4.0	3
40	1154	4.0	3	1,154	4.0	3
41	1156	4.0	3	1,156	4.0	3
42	1158	4.0	3	1,158	4.0	3
43	1158	4.0	3	1,158	4.0	3
44	1160	4.0	3	1,160	4.0	3
45	1160	4.0	3	1,160	4.0	3
46	1164	4.0	4	1,162	4.0	4
47	1164	4.0	4	1,164	4.0	4
48	1166	4.0	4	1,166	4.0	4
49	1168	4.0	4	1,168	4.0	4
50	1172	6.0	4	1,170	4.0	4
51	1174	6.0	4	1,172	6.0	4
52	1178	6.0	4	1,176	6.0	4
53	1180	6.0	4	1,178	6.0	4
54	1180	8.0	4	1,180	8.0	4
55	1180	10.0	4	1,180	10.0	4
56	1180	10.0	4	1,180	10.0	4

APPENDIX O—SCORE DISTRIBUTIONS

Figure O-1. 2018–19 MEA Science: Score Distribution Plots
Top: Grade 5 Bottom: Grade 8



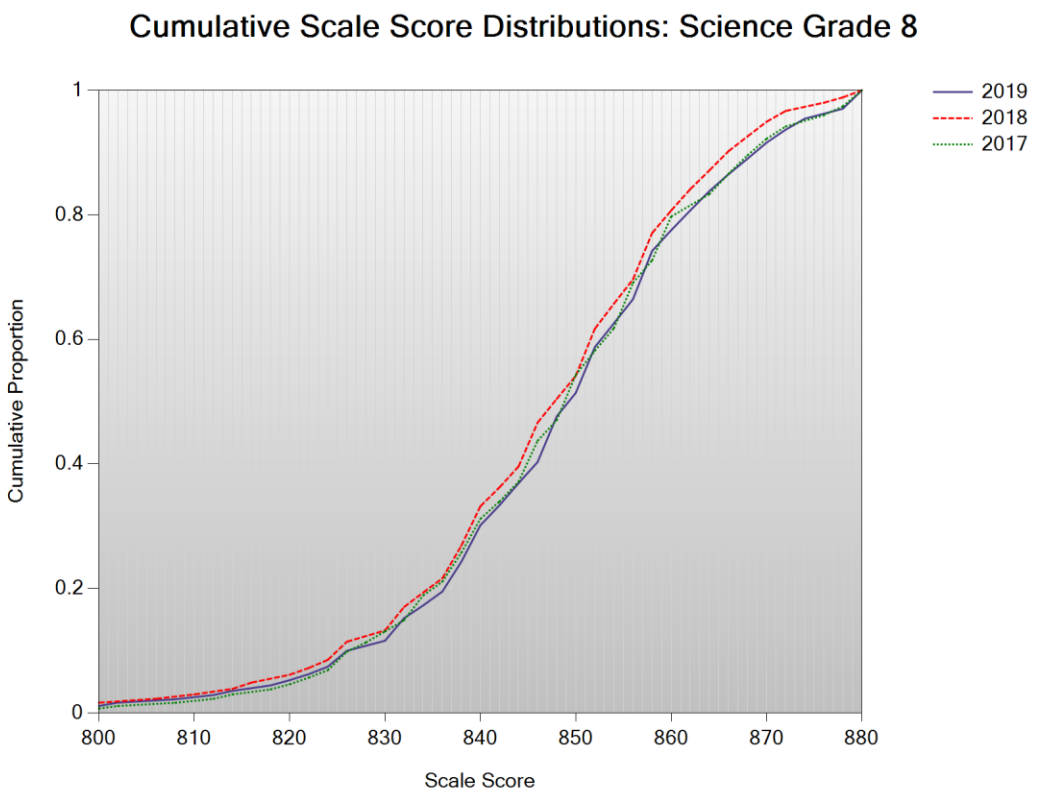


Figure O-2. 2018–19 MEA Science: Score Distribution Plots
Grade 11

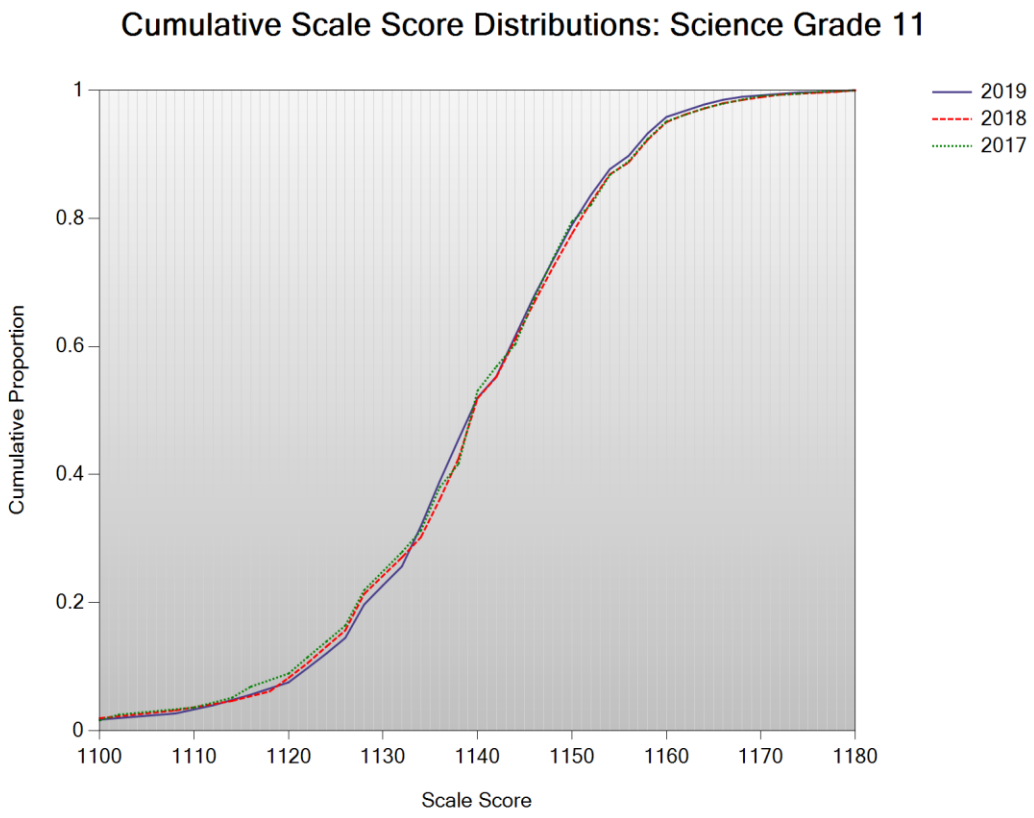


Table O-1. 2018–19 MEA Science: Achievement Level Distributions by Subject and Grade

Grade	*Performance Level*	*2018–19*	*2017–18*	*2016–17*
5	4	5.83	7.84	6.45
	3	55.42	53.35	60.73
	2	26.96	27.62	23.32
	1	11.80	11.20	9.49
8	4	22.49	19.32	20.29
	3	47.41	47.51	48.55
	2	20.16	21.76	21.36
	1	9.95	11.41	9.80
11	4	4.16	4.97	4.85
	3	43.87	43.15	42.09
	2	32.30	30.55	31.15
	1	19.67	21.33	21.91

APPENDIX P—RELIABILITY

Table P-1. 2018–19 MEA Science: Subgroup Reliabilities

Grade	Group	Number of Students	Raw Score			Alpha	SEM
			Maximum	Mean	Standard Deviation		
5	All Students	13,009	48	28.15	8.09	0.84	3.27
	Male	6,727	48	27.76	8.28	0.84	3.27
	Female	6,277	48	28.56	7.86	0.83	3.25
	Gender Not Reported	5	48				
	Hispanic or Latino	473	48	20.11	8.20	0.84	3.32
	American Indian or Alaskan Native	304	48	26.12	8.41	0.84	3.33
	Asian	119	48	25.25	8.34	0.84	3.29
	Black or African American	181	48	29.34	8.80	0.87	3.23
	Native Hawaiian or Pacific Islander	12	48	27.58	8.26	0.85	3.17
	White (non-Hispanic)	11,518	48	28.56	7.87	0.83	3.25
	Two or More Races (non-Hispanic)	397	48	27.71	8.09	0.84	3.25
	Race not reported	5	48				
	Currently receiving LEP services	389	48	16.95	6.51	0.75	3.24
	Former LEP student - monitoring year 1	161	48	27.16	7.69	0.81	3.31
	Former LEP student - monitoring year 2	56	48	30.80	7.25	0.80	3.26
	LEP: All Other Students	12,403	48	28.50	7.89	0.83	3.26
	Students with an IEP	2,594	48	22.42	7.83	0.82	3.35
	IEP: All Other Students	10,415	48	29.57	7.50	0.82	3.22
	Economically Disadvantaged Students	6,129	48	25.69	8.09	0.83	3.32
	SES: All Other Students	6,880	48	30.34	7.42	0.82	3.19
	Migrant Students	1	48				
	Migrant: All Other Students	13,008	48	28.15	8.09	0.84	3.27
	Students receiving Title 1 Services	0	48				
	Title 1: All Other Students	13,009	48	28.15	8.09	0.84	3.27
	Plan 504	620	48	28.11	7.66	0.82	3.28
	Plan 504: All Other Students	12,389	48	28.15	8.11	0.84	3.27
8	All Students	13,066	56	33.34	9.94	0.88	3.47
	Male	6,710	56	33.20	10.44	0.89	3.47
	Female	6,349	56	33.49	9.37	0.87	3.43
	Gender Not Reported	7	56				
	Hispanic or Latino	445	56	25.74	10.43	0.88	3.54
	American Indian or Alaskan Native	287	56	31.29	10.35	0.89	3.48
	Asian	117	56	29.40	10.10	0.88	3.44
	Black or African American	198	56	35.67	9.75	0.87	3.45
	Native Hawaiian or Pacific Islander	10	56	38.00	7.85	0.82	3.30
	White (non-Hispanic)	11,720	56	33.69	9.78	0.87	3.46
	Two or More Races (non-Hispanic)	282	56	32.80	9.98	0.88	3.47
	Race not reported	7	56				
	Currently receiving LEP services	314	56	20.29	8.38	0.83	3.44
	Former LEP student - monitoring year 1	47	56	31.62	7.10	0.75	3.53
	Former LEP student - monitoring year 2	120	56	33.12	8.52	0.83	3.50
	LEP: All Other Students	12,585	56	33.67	9.77	0.87	3.46

continued

Grade	Group	Number of Students	Raw Score			Alpha	SEM
			Maximum	Mean	Standard Deviation		
8	Students with an IEP	2,312	56	24.77	9.49	0.86	3.50
	IEP: All Other Students	10,754	56	35.18	9.02	0.86	3.42
	Economically Disadvantaged Students	5,658	56	29.95	9.87	0.87	3.51
	SES: All Other Students	7,408	56	35.92	9.18	0.86	3.40
	Migrant Students	0	56				
	Migrant: All Other Students	13,066	56	33.34	9.94	0.88	3.47
	Students receiving Title 1 Services	0	56				
	Title 1: All Other Students	13,066	56	33.34	9.94	0.88	3.47
	Plan 504	895	56	33.93	9.46	0.87	3.46
	Plan 504: All Other Students	12,171	56	33.29	9.97	0.88	3.47
11	All Students	12,038	56	27.29	10.37	0.89	3.40
	Male	6,093	56	27.59	10.98	0.91	3.36
	Female	5,945	56	26.98	9.70	0.88	3.42
	Gender Not Reported	0	56				
	Hispanic or Latino	400	56	19.59	8.32	0.84	3.30
	American Indian or Alaskan Native	262	56	25.27	10.17	0.89	3.38
	Asian	93	56	23.48	8.46	0.85	3.31
	Black or African American	205	56	29.51	11.36	0.91	3.44
	Native Hawaiian or Pacific Islander	16	56	27.00	9.59	0.88	3.36
	White (non-Hispanic)	10,862	56	27.64	10.32	0.89	3.40
	Two or More Races (non-Hispanic)	200	56	26.01	10.37	0.89	3.36
	Race not reported	0	56				
	Currently receiving LEP services	287	56	15.58	6.72	0.78	3.13
	Former LEP student - monitoring year 1	39	56	22.62	6.43	0.73	3.34
	Former LEP student - monitoring year 2	75	56	22.89	7.14	0.77	3.41
	LEP: All Other Students	11,637	56	27.62	10.30	0.89	3.40
	Students with an IEP	1,725	56	18.68	8.47	0.85	3.23
	IEP: All Other Students	10,313	56	28.73	9.95	0.88	3.40
	Economically Disadvantaged Students	4,341	56	23.54	9.51	0.88	3.34
	SES: All Other Students	7,697	56	29.40	10.24	0.89	3.40
	Migrant Students	1	56				
	Migrant: All Other Students	12,037	56	27.29	10.37	0.89	3.40
	Students receiving Title 1 Services	0	56				
	Title 1: All Other Students	12,038	56	27.29	10.37	0.89	3.40
	Plan 504	863	56	27.04	10.01	0.89	3.39
	Plan 504: All Other Students	11,175	56	27.31	10.40	0.89	3.40

Table P-2. 2018–19 MEA Science: Reliabilities by Reporting Category

Grade	*Item Reporting Category*	*Number of Items*	*Raw Score*			*Alpha*	*SEM*
			Maximum	*Mean*	*Standard Deviation*		
5	Physical Setting	18	24	13.14	4.27	0.72	2.27
	Living Environment	18	24	15.01	4.40	0.71	2.35
8	Physical Setting	28	34	19.77	6.25	0.80	2.78
	Earth/Space	14	17	10.33	3.38	0.66	1.96
	Matter/Energy/Force/Motion	14	17	9.45	3.41	0.66	1.98
	Living Environment	16	22	13.57	4.25	0.76	2.06
11	Physical Setting	28	34	16.34	6.37	0.82	2.68
	Earth/Space	9	12	5.64	2.44	0.60	1.54
	Matter/Energy/Force/Motion	19	22	10.70	4.45	0.76	2.20
	Living Environment	16	22	10.95	4.54	0.79	2.08

APPENDIX Q—DECISION ACCURACY AND CONSISTENCY

Table Q-1. 2018–19 MEA Science: DAC on Performance Level

Grade	*Overall*	*Kappa*	*Conditional on Level*			
			Well Below State Expectations	*Below State Expectations*	*At State Expectations*	*Above State Expectations*
5	0.79 (0.70)	0.50	0.79 (0.67)	0.72 (0.60)	0.82 (0.79)	0.57 (0.33)
8	0.78 (0.69)	0.54	0.80 (0.68)	0.67 (0.55)	0.80 (0.74)	0.80 (0.70)
11	0.78 (0.70)	0.54	0.82 (0.73)	0.74 (0.64)	0.79 (0.75)	1.00 (0.32)

Table Q-2. 2018–19 MEA Science: DAC on Cut Score

Grade	*Well Below State Expectations / Below State Expectations*			*Below State Expectations / At State Expectations*			*At State Expectations / Above State Expectations*		
	Accuracy (consistency)	*False*		*Accuracy (consistency)*	*False*		*Accuracy (consistency)*	*False*	
		Positive	*Negative*		*Positive*	*Negative*		*Positive*	*Negative*
5	0.95 (0.93)	0.02	0.03	0.89 (0.85)	0.05	0.06	0.94 (0.92)	0.05	0.01
8	0.96 (0.94)	0.02	0.02	0.91 (0.87)	0.04	0.05	0.91 (0.87)	0.05	0.04
11	0.93 (0.90)	0.03	0.03	0.89 (0.85)	0.05	0.06	0.96 (0.95)	0.04	0.00