New England Common Assessment Program

Science
2014–15
Technical Report

December 2015

100 EDUCATION WAY, DOVER, NH 03820 (800) 431-8901
WWW.MEASUREDPROGRESS.ORG

TABLE OF CONTENTS

CHAPTER 1 OVERVIEW 1
1.1 PURPOSE OF THE NEW ENGLAND COMMON ASSESSMENT PROGRAM 1
1.2 PURPOSE OF THIS REPORT 1
1.3 ORGANIZATION OF THIS REPORT 2
CHAPTER 2 CURRENT YEAR UPDATES 3
CHAPTER 3 TEST DESIGN AND DEVELOPMENT 4
3.1 SCIENCE TEST SPECIFICATIONS 4
3.1.1 Overview of Test Design 4
3.1.2 Standards (Assessment Targets) 4
3.1.3 Item Types 6
3.1.4 Test Design 6
3.1.5 Blueprint 7
3.1.6 Depth of Knowledge 7
3.1.7 Use of Calculators and Reference Sheets 8
3.2 TEST DEVELOPMENT PROCESS 9
3.2.1 Item Development 9
3.2.2 Item Reviews by Measured Progress 9
3.2.3 Item Reviews by the States 10
3.2.4 Bias and Sensitivity Review 11
3.2.5 Reviewing and Refining 11
3.2.6 Item Editing 11
3.2.7 Item Selection and Operational Test Assembly 11
3.2.8 Operational Test Draft Review 12
3.2.9 Alternative Presentations 13
CHAPTER 4 TEST ADMINISTRATION 14
4.1 RESPONSIBILITY FOR ADMINISTRATION 14
4.2 ADMINISTRATION PROCEDURES 14
4.3 PARTICIPATION REQUIREMENTS AND DOCUMENTATION 14
4.3.1 Students With Disabilities 15
4.3.2 Students With Limited English Proficiency 15
4.4 ADMINISTRATOR TRAINING 15
4.5 DOCUMENTATION OF ACCOMMODATIONS 16
4.6 TEST SECURITY 16
4.7 TEST AND ADMINISTRATION IRREGULARITIES 17
4.8 TEST ADMINISTRATION WINDOW 17
4.9 NECAP SERVICE CENTER 17
CHAPTER 5 SCORING 18
5.1 MACHINE-SCORED ITEMS 18

5.2 PERSON-SCORED ITEMS 18
5.3 INQUIRY TASK SCORING 19
5.4 SCORING LOCATION AND STAFF 19
5.4.1 Scorer Recruitment and Qualifications 20
5.4.2 Scorer Training 20
5.4.2.1 Anchor Set 21
5.4.2.2 Practice Set 21
5.4.2.3 Qualifying Set 22
5.4.2.4 Retraining 22
5.4.3 Scoring Supervisor and STL Training 22
5.4.4 Benchmarking Meetings 22
5.5 METHODOLOGY FOR SCORING CONSTRUCTED-RESPONSE ITEMS 23
5.5.1 Monitoring of Scoring Quality Control and Consistency 24
5.5.1.1 Embedded CRRs 25
5.5.1.2 Read-Behind Procedures 25
5.5.1.3 Double-Blind Scoring 26
5.5.1.4 Recalibration Sets 26
5.5.2 Scoring Reports 27
CHAPTER 6 CLASSICAL ITEM ANALYSIS 29
6.1 CLASSICAL DIFFICULTY AND DISCRIMINATION INDICES 29
6.2 DIFFERENTIAL ITEM FUNCTIONING (DIF) 31
6.3 DIMENSIONALITY ANALYSES 32
CHAPTER 7 SCALING AND EQUATING 36
7.1 ITEM RESPONSE THEORY (IRT) 37
7.2 IRT RESULTS 38
7.3 EQUATING 39
7.4 EQUATING RESULTS 40
7.5 ACHIEVEMENT STANDARDS 41
7.6 REPORTED SCALED SCORES 42
CHAPTER 8 RELIABILITY 44
8.1 RELIABILITY AND STANDARD ERROR OF MEASUREMENT 45
8.2 2014–15 SUBGROUP RELIABILITY 45
8.3 REPORTING SUBCATEGORY RELIABILITY 46
8.4 INTERRATER CONSISTENCY 46
8.5 RELIABILITY OF ACHIEVEMENT LEVEL CATEGORIZATION 47
8.5.1 Accuracy and Consistency Results 48
CHAPTER 9 SCORE REPORTING 50
9.1 PRIMARY REPORTS 50
9.2 STUDENT REPORT 50
9.3 INTERACTIVE REPORTING 51
9.3.1 Item Analysis Roster Report 51
9.3.2 Achievement Level Summary 52
9.3.3 Released Items Summary Data Report 53

9.3.4 Longitudinal Data Report 53
9.4 School, District, and State Grade-Level Results Reports 53
9.5 District and State Summary Reports 56
9.6 Decision Rules 56
9.7 Quality Assurance 56
CHAPTER 10 VALIDITY 59
10.1 Questionnaire Data 60
10.1.1 Difficulty of Assessment 60
10.1.2 Content 62
10.1.3 Homework 64
10.1.4 Performance in Science Class 65
REFERENCES 67
APPENDICES 69
APPENDIX A GUIDELINES FOR THE DEVELOPMENT OF SCIENCE INQUIRY TASKS
APPENDIX B NECAP SCIENCE COMMITTEE MEMBERS
APPENDIX C PARTICIPATION RATES
APPENDIX D ACCOMMODATION FREQUENCIES
APPENDIX E NECAP TABLE OF STANDARD ACCOMMODATIONS
APPENDIX F ITEM-LEVEL CLASSICAL STATISTICS
APPENDIX G ITEM-LEVEL SCORE POINT DISTRIBUTIONS
APPENDIX H DIFFERENTIAL ITEM FUNCTIONING RESULTS
APPENDIX I ITEM RESPONSE THEORY CALIBRATION RESULTS
APPENDIX J TEST CHARACTERISTIC CURVE AND TEST INFORMATION FUNCTION CHARTS
APPENDIX K DELTA ANALYSES AND RESCORE ANALYSES
APPENDIX L α-PLOTS AND b-PLOTS
APPENDIX M ACHIEVEMENT LEVEL DISTRIBUTIONS
APPENDIX N RAW TO SCALED SCORE LOOKUP TABLES
APPENDIX O SCALED SCORE DISTRIBUTIONS
APPENDIX P CLASSICAL RELIABILITIES
APPENDIX Q INTERRATER CONSISTENCY
APPENDIX R DECISION ACCURACY AND CONSISTENCY RESULTS
APPENDIX S SAMPLE REPORTS
APPENDIX T INTERACTIVE REPORTS
APPENDIX U ANALYSIS AND REPORTING DECISION RULES

CHAPTER 1 OVERVIEW

1.1 PURPOSE OF THE NEW ENGLAND COMMON ASSESSMENT PROGRAM

The New England Common Assessment Program (NECAP) is the result of collaboration among New Hampshire, Rhode Island, and Vermont to build a set of tests for grades 3 through 8 and 11 to meet the requirements of the No Child Left Behind (NCLB) Act. The specific purposes of the NECAP Science tests are (1) to provide data on student achievement in science at grades 4, 8, and 11 to meet NCLB requirements; (2) to provide information to support program evaluation and improvement; and (3) to provide information to parents/guardians and the public on the performance of students and schools. The tests are constructed to meet rigorous technical criteria, to include universal design elements and accommodations so that students can access test content, and to gather reliable student demographic information for accurate reporting. School improvement is supported by

- providing a transparent test design through the NECAP Science assessment targets, distributions of emphasis, and practice tests;
- reporting results by science domain, released items, and subgroups; and
- hosting test interpretation workshops and webinars to foster understanding of results.

Student-level results are provided to schools and families to be used as one element among all the collected evidence about progress and learning that occurred on the assessment targets for the respective grade span (K–4, 5–8, 9–11). The results are a status report of a student's performance against the assessment targets, and they should be used cautiously in concert with local data.

1.2 PURPOSE OF THIS REPORT

The purpose of this report is to document the technical aspects of the 2014–15 NECAP Science tests. Students in grades 4, 8, and 11 participated in the eighth operational administration of NECAP Science in May 2015. This report provides evidence on the technical quality of those tests, including descriptions of the processes used to develop, administer, and score the tests, as well as the processes used to analyze the results. This report is intended to serve as a guide for replicating or improving the procedures in subsequent years.

Though some parts of this technical report may be used by educated laypeople, it is intended for experts in psychometrics and educational research. The report assumes a working knowledge of measurement concepts, such as reliability and validity, and statistical concepts, such as correlation and central tendency. In some chapters, the reader is presumed also to have basic familiarity with advanced topics in measurement and statistics.

1.3 ORGANIZATION OF THIS REPORT

The organization of this report is based on the conceptual flow of a test's life span. The report begins with the initial test specifications and addresses all intermediate steps that lead to final score reporting. Chapters 1 through 5 give a description of NECAP Science by covering the test design and development process, the administration of the test, and scoring. Chapters 6 through 8 provide statistical and psychometric information, including chapters on scaling and equating, item analysis, and reliability. Chapter 9 is devoted to NECAP Science score reporting, and Chapter 10 is devoted to discussions on validity. Finally, the references cited throughout the report are provided, followed by the report appendices.

CHAPTER 2 CURRENT YEAR UPDATES

There were two changes in the NECAP Science assessment in 2015. First, the NECAP state assessment directors expanded the testing window from 14 days to 18 days. This was done to provide more flexibility to testing schools, as this was the first year NECAP Science testing overlapped with PARCC and Smarter Balanced testing for mathematics and English/language arts. Second, the New Hampshire department of education made changes to the student identification system for NECAP Science testing. In 2015, all New Hampshire test takers were assigned a new, unique testing ID, which was different from the state-assigned student ID (SASID) that was used for previous test administrations and exists across the years the student is enrolled in a participating school. As a result of this change, certain historical reporting data, such as longitudinal data across multiple NECAP test administrations, were no longer available in the NECAP analysis and reporting system.

CHAPTER 3 TEST DESIGN AND DEVELOPMENT

3.1 SCIENCE TEST SPECIFICATIONS

3.1.1 Overview of Test Design

The NECAP Science test is a criterion-referenced test. Items on the test are developed specifically for NECAP and are directly aligned to NECAP's Science assessment targets. These assessment targets guide the development of test items and are the basis for the reporting categories.

The 2014–15 NECAP Science test was administered in spring 2015 at grades 4, 8, and 11. The test consisted of four forms per grade. Each form included common items, equating items, and embedded field-test items. Common items are those that appear on every form of the test and are used to determine a student's test score. Each equating item appears on one form only, and because these items have been on previous tests, they are used by psychometricians to keep the test scores on the same scale from year to year. This design provides reliable and valid results at the student level (the common items) and breadth of science coverage for school results (the common plus equating items) while minimizing testing time.

The NECAP Science test included an embedded field test. Each embedded field-test item generally appears on only one of the four forms. The field-test items were distributed equally among the forms. Since students do not know which items count for their test score, embedding field-test items into the operational test ensures that students take these items seriously. Because each field-test form is taken by approximately one-fourth of the NECAP students, the sample size is large enough to produce reliable field-test data. The embedded field test yields a pool of replacement items, which are needed due to the release of approximately twenty-five percent of the common items every year.

Each form of the test has three sessions. Physical Science, Earth Space Science, and Life Science are assessed in Sessions 1 and 2. These sessions contain common, equating, and embedded field-test items. Scientific Inquiry is assessed in Session 3 by an inquiry task. Session 3 contains only common items, as the inquiry tasks go through a separate field test (rather than an embedded field test).

3.1.2 Standards (Assessment Targets)

Although the NECAP Science assessment targets are unique for each grade, the assessment targets across the grades are classified under the same statements of enduring knowledge or broad areas of inquiry

that make up the four reporting categories of Physical Science, Earth Space Science, Life Science, and Scientific Inquiry.

Life Science

1. Survival of Organisms—All living organisms have identifiable structures and characteristics that allow for survival (organisms, populations, and species).
2. Matter and Energy in Ecosystems—Matter cycles and energy flows through an ecosystem.
3. Organisms Change Over Time—Groups of organisms show evidence of change over time (structures, behaviors, and biochemistry).
4. Humans Are Similar Yet Unique—Humans are similar to other species in many ways and yet are unique among Earth's life forms.

Physical Science

1. Properties and Structure of Matter—All living and nonliving things are composed of matter having characteristic properties that distinguish one substance from another (independent of size or amount of substance).
2. Energy—Energy is necessary for change to occur in matter. Energy can be stored, transferred, and transformed, but it cannot be destroyed.
3. Forces and Motion—The motion of an object is affected by forces.

Earth Space Science

1. Earth and Earth Materials—The Earth and its materials as we know them today have developed over long periods of time, through continual change processes.
2. Solar System—The Earth is part of a solar system made up of distinct parts that have temporal and spatial interrelationships.
3. Universe and Galaxies—The origin and evolution of galaxies and the universe demonstrate fundamental principles of physical science across vast distances and time.

Scientific Inquiry

1. Formulating Questions and Hypothesizing
2. Planning and Critiquing of Investigations
3. Conducting Investigations
4. Developing and Evaluating Explanations

3.1.3 Item Types

Since the beginning of the program, the goal of NECAP has been to measure what students know and are able to do by using a variety of test item types. The 2014–15 NECAP Science test consisted of standalone items and an inquiry task at each grade level. At grades 4 and 8, the inquiry task consisted of a hands-on investigation. At grade 11, the inquiry task consisted of a description of a scientific investigation.

The item types used and the functions of each are described below.

- **Multiple-choice (MC)** items were administered to provide breadth of coverage of the assessment targets. Because they require approximately one minute for most students to answer, these items make efficient use of limited testing time and allow coverage of a wide range of knowledge and skills. Each multiple-choice item was worth one point. Multiple-choice items were administered in Sessions 1 and 2 of the test in the Physical Science, Earth Space Science, and Life Science domains.
- **Short-answer (SA)** items assess students' skills and their abilities to work with brief, well-structured problems that have one solution or a very limited number of solutions. The advantage of this item type is that it requires students to demonstrate knowledge and skills by generating, rather than merely selecting, an answer. Each short-answer item was worth two points. Short-answer items require approximately two to five minutes for most students to answer. Short-answer items were employed in the Session 3 inquiry task.
- **Constructed-response (CR)** items typically require students to use higher-order thinking skills—evaluation, analysis, and summarization—in constructing a satisfactory response. Constructed-response items should take most students approximately five to ten minutes to complete. Four-point constructed-response items were administered in Sessions 1 and 2 of the test in the Physical Science, Earth Space Science, and Life Science domains. Three-point constructed-response items were administered in the Session 3 inquiry task.

At each grade, approximately twenty-five percent of the common standalone items and the entire inquiry task were released to the public. The released items are posted on a website hosted by Measured Progress and on the state agency/departments of education websites. Schools are encouraged to incorporate the use of released items in their instructional activities so that students will be familiar with them.

3.1.4 Test Design

Table 3-1 summarizes the numbers and types of common items that were used to determine student scores on the NECAP Science assessment. In addition, each test form contained 18 multiple-choice and 3 constructed-response standalone items that were used either for equating or field-testing.

Table 3-1. 2014–15 NECAP Science: Common Items

Grade	Session			Total			
	1	*2*	*3*	*MC 1 pt*	*SA 2 pt*	*CR 3 pt*	*CR 4 pt*
4	16 MC 2 CR	17 MC 1 CR	6 SA 2 CR	33	6	2	3
8	16 MC 2 CR	17 MC 1 CR	6 SA 2 CR	33	6	2	3
11	16 MC 2 CR	17 MC 1 CR	6 SA 2 CR	33	6	2	3

MC = multiple choice; SA = short answer; CR = constructed response

3.1.5 Blueprint

NECAP Science items are categorized into the following reporting categories: Physical Science, Earth Space Science, Life Science, and Scientific Inquiry. The item distribution across the reporting categories is shown in Table 3-2.

Table 3-2. 2014–15 NECAP Science: Distribution of Common Items

Domain	*MC 1 pt*	*SA 2 pt*	*CR 3 pt*	*CR 4 pt*
Physical Science	11			1
Earth Space Science	11			1
Life Science	11			1
Scientific Inquiry		6	2	
Total	33	6	2	3

MC = multiple choice; SA = short answer; CR = constructed response

Table 3-3 displays how the raw score points are distributed across the reporting categories.

Table 3-3. 2014–15 NECAP Science: Distribution of Score Points

Domain	*Points*	*Percentage of Test*
Physical Science	15	24%
Earth Space Science	15	24%
Life Science	15	24%
Scientific Inquiry	18	28%
Total	63	100%

3.1.6 Depth of Knowledge

Each item on the NECAP Science test is assigned a Depth of Knowledge (DOK) level. The DOK level reflects the complexity of mental processing students must use to answer an item. The 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 only 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. Due to the nature of this level, there are no level 4 items on the NECAP Science test.

It is important that the NECAP Science test measures a range of DOK levels. Table 3-4 lists the percentage of total score points assigned to each DOK level.

Table 3-4. 2014–15 NECAP Science: DOK Percentages

	Grade 4	*Grade 8*	*Grade 11*
DOK 1	11%	16%	11%
DOK 2	78%	67%	68%
DOK 3	11%	17%	21%

3.1.7 Use of Calculators and Reference Sheets

Science specialists from the New Hampshire and Rhode Island departments of education and the Vermont agency of education acknowledge that the use of calculators is a necessary and important skill. Calculators can save time and allow students to solve more sophisticated and intricate problems by reducing errors in calculations. For these reasons, it was decided that calculators should be permitted in all three sessions of the NECAP Science assessment. However, the state science specialists chose to prohibit scientific and graphing calculators in Session 3 because the inquiry task includes a graphing item.

A reference sheet is provided for the grade 8 and grade 11 tests. The grade 8 reference sheet includes solar system data, algebraic formulas used in science, the biological classification system, quantities with corresponding standard units of measure, a map of plate movements, and the periodic table of the elements. The grade 11 reference sheet includes genetic codes for amino acids, the electromagnetic spectrum, a map of plate movements, algebraic formulas used in science, and the periodic table of the elements.

3.2 TEST DEVELOPMENT PROCESS

3.2.1 Item Development

The item and inquiry task development process combined the expertise of the NECAP state science specialists, committees of NECAP educators, and Measured Progress test developers to help ensure items met the needs of the NECAP program. NECAP Science items are directly aligned to the assessment targets for each science domain. Each item addresses one assessment target. Measured Progress test developers worked with the state science specialists and teachers from the states to verify the alignment of items to the appropriate assessment target. All items used on the NECAP test were reviewed by a committee of NECAP teachers and by a NECAP bias and sensitivity committee.

The inquiry tasks were developed by Measured Progress test developers using *Guidelines for the Development of Science Inquiry Tasks*, a document created by the science specialists for each state agency/department of education. The document can be found in Appendix A of this report. The inquiry tasks were field-tested separately rather than as part of the embedded field test. Measured Progress test developers, the scoring content manager, and program managers observed field-testing of the inquiry tasks in schools located in Maine. The selected schools had varying demographics and population sizes. Each inquiry task was administered to approximately 200 students. Measured Progress test developers and program managers prepared a document titled *Inquiry Task Field Test Report* for the state science specialists to review. The state science specialists then approved the final form of each inquiry task. Due to space limitations, the *Inquiry Task Field Test Report* is not reproduced here. However, it can be obtained from any of the three NECAP states as a standalone document.

3.2.2 Item Reviews by Measured Progress

Measured Progress conducted two internal reviews of the multiple-choice and constructed-response items as well as a review of the inquiry tasks before presenting the items to the state science specialists. During these reviews, science test developers focused on three major areas.

- Item alignment to the assessment target: The reviewers considered whether the item measured the content as outlined in the assessment target and whether the content was grade appropriate. The reviewers also checked the DOK level of the item.
- Correctness of science content: The reviewers considered whether the information in the item was scientifically correct. For multiple-choice items, the keyed answer had to be the only correct answer. For constructed-response items, the scoring guide had to reflect correct science content and grade-level appropriate responses.
- Universal design: The reviewers considered item structure, clarity, possible ambiguity, and the appropriateness and relevance of graphics. For constructed-response items, the reviewers

considered whether the item adequately prompted an examinee to give a response similar to the one in the scoring guide.

3.2.3 Item Reviews by the States

The state science specialists reviewed the items. Measured Progress revised the items based on edits requested by the specialists.

Item review committees (IRC) comprising state teachers and curriculum supervisors were formed to conduct another review of the items. A list of the 2014– 15 NECAP IRC participants for science in grades 4, 8, and 11 and their affiliations is included in Appendix B. On August 7–8, 2014, the NECAP Science IRC was held at the Stoweflake Mountain Resort & Spa in Stowe, Vermont. Participants' primary role was to evaluate and provide feedback on potential field-test items. For each grade level, the committee members reviewed potential multiple-choice and constructed-response field-test items as well as potential inquiry tasks. During the meeting, committee members were asked to evaluate the items for the following criteria:

- Assessment target alignment:
  - Is the test item aligned to the identified assessment target?
- DOK:
  - Are the items coded to the appropriate DOK level?
- Scientific correctness:
  - Are the items and distractors correct with respect to content and grade-level appropriateness?
  - Are the scoring guides consistent with the item and do they provide grade-level appropriate responses?
- Universal design:
  - Is the item language clear and grade appropriate?
  - Is the item language accurate (syntax, grammar, conventions)?
  - Is there an appropriate use of simplified language (is language that interferes with the assessment target avoided)?
  - Are charts, tables, and diagrams easy to read and understandable?
  - Are charts, tables, and diagrams necessary to the item?
  - Are instructions easy to follow?
  - Is the item amenable to accommodations—read aloud, signed, or in Braille?

3.2.4 Bias and Sensitivity Review

Bias review is an essential component of the development process. During the bias review process, NECAP Science items were reviewed by a committee of general education teachers, English language learner (ELL) specialists, special education teachers, and other educators and members of major constituency groups who represent the interests of legally protected and/or educationally disadvantaged groups. A list of bias and sensitivity review committee participants and affiliations is included in Appendix B. Items were examined for issues that might offend or dismay students, teachers, or parents/guardians. Including such educators 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.

3.2.5 Reviewing and Refining

After the IRC and bias and sensitivity review committee meetings, Measured Progress test developers and the state science specialists met to review the committees' feedback. The specialists decided what edits should be made to the items.

3.2.6 Item Editing

Measured Progress editors then reviewed and edited the items to ensure uniform style (based on *The Chicago Manual of Style*, 16th edition) and adherence to sound testing principles. These principles included the stipulation that items were

- correct with regard to grammar, punctuation, usage, and spelling;
- written in a clear, concise style;
- written at a reading level that allows the student to demonstrate his or her knowledge of science, regardless of reading ability;
- written in a way that did not cue the correct answer (for multiple-choice options); and
- free of potentially sensitive content.

3.2.7 Item Selection and Operational Test Assembly

In preparation for the item selection meeting with the state science specialists, test developers and psychometricians at Measured Progress considered the following when selecting sets of items to propose for the common (including items for release) and the embedded field tests:

- **Content coverage/match to test design.** The test design stipulates a specific number of multiple-choice and constructed-response items from each content area. Item selection for the embedded field test was based on the number of items in the existing pool of items eligible for the common item set.

- **Item difficulty and complexity.** Item statistics drawn from the data analysis of previously tested items were used to ensure similar levels of difficulty and complexity from year to year as well as quality psychometric characteristics.
- **"Cueing" items.** Items were reviewed for any information that might "cue" or provide information that would help to answer another item.

At the item selection meeting, the state specialists reviewed the proposed sets of items and made the final selection of items for the common item set, including which items would be released after the test was administered. The state specialists also made the final selection of items for the embedded field test and approved the final wording of these items.

During assembly of the test forms, the following criteria were considered:

- **Option balance.** Items were balanced among the forms so that each form contained a fairly equal distribution of keys (correct answers).
- **Key patterns.** The sequence of keys was reviewed to ensure that key order appeared random.
- **Page fit.** Item placement was modified to ensure the best fit and arrangement of items on any given page.
- **Facing-page issues.** For multiple items associated with a single stimulus (inquiry task) and multiple-choice items with large graphics, consideration was given to whether those items needed to begin on a left- or right-hand page and to the nature and amount of material that needed to be placed on facing pages. These considerations serve to minimize the amount of page flipping required of students.
- **Relationship between forms.** Although equating and field-test items differ across forms, these items must take up the same number of pages in each form 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 taken into consideration, including such aspects as the amount of white space, the density of the text, and the number of graphics.

3.2.8 Operational Test Draft Review

After the forms were laid out as they would appear in the final test booklets, they were again thoroughly reviewed by Measured Progress editors and test developers to ensure that the items appeared exactly as the state science specialists had requested. Finally, all the forms were reviewed by the state science specialists for their final approval.

3.2.9 Alternative Presentations

Common items for grades 4, 8, and 11 were translated into Braille by a subcontractor that specializes in test materials for students who are blind or visually impaired. In addition, Form 1 for each grade was also adapted into a large-print version.

CHAPTER 4 TEST ADMINISTRATION

4.1 RESPONSIBILITY FOR ADMINISTRATION

The 2015 NECAP Science *Principal/Test Coordinator Manual* indicated that principals and/or their designated NECAP test coordinators were responsible for the proper administration of the NECAP Science test. The *Test Administrator Manual*, which contained explicit directions and read-aloud scripts, was used to ensure the uniformity of administration procedures from school to school.

4.2 ADMINISTRATION PROCEDURES

Principals and/or their schools' designated NECAP coordinators were instructed to read the *Principal/Test Coordinator Manual* before testing and familiarize themselves with the instructions provided in the *Test Administrator Manual*. The *Principal/Test Coordinator Manual* provided each school with checklists to help them prepare for testing. The checklists outlined tasks to be performed by school staff before, during, and after test administration. Besides these checklists, the *Principal/Test Coordinator Manual* described the testing material being sent to each school and how to inventory the material, track it during administration, and return it after testing was complete. The *Test Administrator Manual* included checklists for the administrators to ready themselves, their classrooms, and the students for the administration of the test. It also contained sections detailing the procedures to be followed for each test session and instructions for preparing the material before its return to Measured Progress.

4.3 PARTICIPATION REQUIREMENTS AND DOCUMENTATION

For New Hampshire, Rhode Island, and Vermont, the intent of No Child Left Behind (NCLB) legislation is for *all* students in grades 4, 8, and 11 to participate in science testing through standard administration, administration with accommodations, or alternate assessment. Furthermore, any student who is absent during any session of the NECAP Science test is expected to make up the missed sessions within the 18-day testing window.

Schools were required to return a student answer booklet for every enrolled student in the grade level. On those occasions when it was deemed impossible to test a particular student, school personnel were required to inform their state agency/department of education. The states included a grid on the student answer booklets listing the approved reasons why a student answer booklet could be returned blank for one or more sessions of the science test.

- **Student withdrew from school after May 4, 2015.** If a student withdrew after May 4, 2015, but before completing all of the test sessions, school personnel were instructed to code this reason on the student's answer booklet.
- **Student enrolled in school after May 4, 2015.** If a student enrolled after May 4, 2015, and was unable to complete all of the test sessions before the end of the testing administration window, school personnel were instructed to code this reason on the student's answer booklet.
- **State-approved special consideration.** Each state agency/department of education had a process for documenting and approving circumstances that made it impossible or not advisable for a student to participate in testing. Schools were required to obtain state approval before beginning testing.
- **Student was enrolled on May 4, 2015, and did not complete test for reasons other than those listed above.** If a student was not tested for a different reason, school personnel were instructed to code this reason on the student's answer booklet. These "other" categories were considered not state approved.

Appendix C lists the science participation rates of the three states combined.

4.3.1 Students With Disabilities

All students were expected to participate in the 2015 NECAP Science tests, unless they completed the state-specific alternate assessment in New Hampshire, Rhode Island, or Vermont during the 2014–15 school year.

Form 1 of the grades 4, 8, and 11 science tests was enlarged to 20-point font for students with visual impairments. At all three grades, only the common items were translated into Braille by American Printing House for the Blind, a subcontractor that specializes in test materials for students who are blind or have visual impairments.

4.3.2 Students With Limited English Proficiency

Students who were new to the United States after October 1, 2014, and were designated as limited English proficient (LEP) were required to take the 2015 NECAP Science test.

4.4 ADMINISTRATOR TRAINING

In addition to distributing the *Principal/Test Coordinator Manual* and *Test Administrator Manual*, the New Hampshire and Rhode Island departments of education and Vermont agency of education, along with Measured Progress, conducted test administration workshops or a webinar to inform school personnel about the NECAP Science test and to provide training on the policies and procedures regarding administration.

4.5 DOCUMENTATION OF ACCOMMODATIONS

Though every effort was made to provide a test that would be as accessible as possible, a need still remained to allow some students to take the test with accommodations. An operating principle employed during the development of the accommodations protocols and policy development was to allow only accommodations that would not change the construct of what was being measured by the item.

The *Principal/Test Coordinator Manual* and *Test Administrator Manual* provided directions for coding the information related to accommodations and modifications on page 2 of the student answer booklet. All accommodations used during any test session were required to be coded by authorized school personnel—not students—after testing was completed.

The *NECAP Accommodations Guide* also provides detailed information on planning and implementing accommodations. This guide can be found on each state's agency/department of education website. The states collectively made the decision that accommodations be made available to all students based on individual need, regardless of disability status. Decisions regarding accommodations were made by the students' educational teams on an individual basis and were consistent with those used during the students' regular classroom instruction. Making accommodations decisions on an entire group basis rather than on an individual basis was not permitted. If the decision made by a student's educational team required an accommodation not listed in the state-approved NECAP Table of Standard Accommodations, schools were instructed to contact their agency/department of education in advance of testing for specific instructions for coding the "Other Accommodations (O)" and/or "Modifications (M)" sections.

Table 4-1 shows the accommodations observed for the May 2015 NECAP Science administration. Appendix D shows the breakdown of students who tested with which accommodation. The accommodation codes are defined in the NECAP Table of Standard Accommodations, found in Appendix E.

Table 4-1. 2014–15 NECAP Science: Number of Students Tested With and Without Accommodations by Grade

Grade	*Number of Students Tested*	
	Without Accommodations	*With Accommodations*
4	24,253	5,420
8	26,333	3,852
11	26,015	2,251

4.6 TEST SECURITY

Maintaining test security is critical to the success of NECAP and the continued partnership among the three states. The *Principal/Test Coordinator Manual* and *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 test coordinator and principal immediately. The test

coordinator and/or principal were responsible for immediately reporting the concern to the district superintendent and the state director of testing at the agency/department of education. Test security was strongly emphasized in the test administration workshops and webinars conducted for all three states. The states required the principal of each school that participated in testing to log on to a secure website to complete the Principal's Certification of Proper Test Administration form for each grade level tested. The principal was required to provide the number of secure tests received from Measured Progress, the number of tests administered to students, and the number of secure test materials being returned to Measured Progress. The principal was then required to enter his or her name in the online form as an electronic signature. By signing the form, the principal was certifying that the tests were administered according to the procedures outlined in the *Principal/Test Coordinator Manual* and *Test Administrator Manual*, that he or she maintained the security of the test materials, that no secure material was duplicated or in any way retained in the school, and that all test materials had been accounted for and scheduled for return to Measured Progress.

4.7 Test and Administration Irregularities

There were no testing or administration irregularities reported to Measured Progress or documented by Measured Progress for the 2015 NECAP Science test.

4.8 Test Administration Window

The test administration window was May 4–28, 2015.

4.9 NECAP Service Center

To provide additional support to schools before, during, and after testing, Measured Progress established the NECAP Service Center. The additional support that the service center provided was an essential element to the successful administration of the three-state test program. Individuals in the field could call the centralized location using a toll-free number and ask questions or report any problems they were experiencing.

The service center was staffed based on call volume and was available from 8:00 a.m. to 4:00 p.m. beginning two weeks before the start of testing and ending two weeks after testing. The representatives were responsible for receiving, responding to, and tracking calls and then routing issues to the appropriate person(s) for resolution.

CHAPTER 5 SCORING

Upon receipt of used NECAP Science answer booklets following testing, Measured Progress scanned all student responses, along with student identification and demographic information. Imaged data for multiple-choice items were machine-scored. Images of constructed-response items were processed and organized by iScore, secure server-to-server electronic scoring software designed by Measured Progress, for hand-scoring.

Student responses that could not be physically scanned (e.g., answer documents damaged during shipping) were physically reviewed and scored on an individual basis by trained, qualified scorers. These scores were linked to the student's demographic data and merged with the student's scoring file by Measured Progress's data processing department.

5.1 MACHINE-SCORED ITEMS

Multiple-choice responses were compared to scoring keys using item analysis software. Correct answers were assigned a score of 1 point; incorrect answers were assigned a score of 0 points. Student responses with multiple marks or blank responses were also assigned 0 points.

The hardware elements of the scanners monitored themselves continuously for correct reads, and the software driving these scanners monitored the correct data reads. Standard checks included recognition of a sheet that did not belong, was upside down, or was backward; identification of missing critical data, including a student ID number or test form that was out of range or missing; and identification of page/document sequence errors. When a problem was detected, the scanner stopped and displayed an error message directing the operator to investigate and correct the situation.

5.2 PERSON-SCORED ITEMS

The images of student responses to constructed-response items were hand-scored through the iScore system. Using iScore minimized the need for scorers to physically handle actual answer booklets and related scoring materials. Student confidentiality was easily maintained, as all NECAP Science scoring was blind (i.e., district, school, and student names were not visible to scorers). The iScore system maintained the link between the student response images and their associated test booklet numbers.

Through iScore, qualified scorers accessed electronically scanned images of student responses at computer terminals. The scorers evaluated each response and recorded each student's score via keypad or mouse entry through the iScore system. When a scorer finished one response, the next response immediately appeared 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 item at a time; however, when necessary, imaged responses from a student's entire booklet were available for viewing, and the physical booklet was also available to the on-site Scoring Content Specialist.

The use of iScore also helped ensure that access to student response images was limited to only those who were scoring or who were working for Measured Progress in a scoring management capacity.

5.3 Inquiry Task Scoring

Of special interest during this cycle of scoring the 2014–15 NECAP Science test was implementing the scoring requirements associated with inquiry task items. These items were unique in that students conducted a single scientific experiment and then answered the eight assigned questions about that experiment. The questions were designed to stand alone, meaning that each one could be scored separately, instead of as part of a set of several combined questions. This maximized the number of scorers that could be assigned to score responses for each student.

5.4 Scoring Location and Staff

Scoring Location

The iScore database, its operation, and its administrative controls are all based in Dover, New Hampshire; in addition, all 2014–15 NECAP Science responses were scored in Longmont, Colorado (grade 8 and 11 inquiry task items), and Dover (all operational grades as well as grades 4 items).

The iScore system monitored accuracy, reliability, and consistency across the scoring site. Constant daily communication and coordination were accomplished through e-mail, telephone, and secure websites, to ensure that critical information and scoring modifications were shared and implemented throughout the scoring site.

Staff Positions

The following staff members were involved with scoring the 2014–15 NECAP Science responses:

- The NECAP Science scoring project manager, based in Dover, New Hampshire, oversaw the communication and coordination of scoring constructed-response items.
- The iScore operational manager and iScore administrators, based in Dover, New Hampshire, coordinated technical communication pertaining to the scoring of constructed-response items.
- A Scoring Content Specialist in science ensured the consistency of scoring across the scoring site for all grades tested. The Scoring Content Specialist, based in Longmont, Colorado, also provided read-behind activities for Scoring Supervisors.
- Numerous Scoring Supervisors, selected from a pool of experienced Scoring Team Leaders (STLs) for their ability to score accurately and their ability to instruct and train scorers,

participated in benchmarking activities for each grade. Scoring Supervisors provided read-behind activities for STLs.

- Numerous STLs, selected from a pool of skilled and experienced scorers, provided read-behind activities for the scorers at their scoring tables (2 to 12 scorers at each table).
- Scorers at the scoring sites scored the 2014–15 NECAP Science operational and field-test student responses.

5.4.1 Scorer Recruitment and Qualifications

For scoring of the 2014–15 NECAP Science test, Measured Progress actively sought a diverse scoring pool that was representative of the population of the three participating NECAP Science states. The broad range of scorers included scientists, business professionals, authors, teachers, graduate school students, and retired educators. Demographic information for scorers (e.g., gender, race, educational background) was electronically captured and reported.

Although a four-year college degree or higher was preferred for all scorers, scorers of the NECAP Science test responses of grades 4, 8, and 11 students were required to have successfully completed at least two years of college and to have a demonstrated knowledge of science. This permitted the recruitment of scorers who were currently enrolled in a college program, a sector of the population that had relatively recent exposure to classroom practices and current trends in their field of study. In all cases, potential scorers submitted documentation (e.g., resume and/or transcripts) of their qualifications.

Table 5-1 summarizes the qualifications of the 2014–15 NECAP Science scoring leadership (Scoring Supervisors and STLs) and scorers.

Table 5-1. 2014–15 NECAP Science: Qualifications of Scoring Leadership and Scorers

Scoring Responsibility	*Spring 2015 Administration Educational Credentials*				*Total*
	Doctorate	*Master's*	*Bachelor's*	*Other*	
Scoring leadership	5	14	13	2*	34
Scorers	21	68	100	14	203

* Indicates the 1 STL with an associate's degree and the 1 STL with at least 48 college credits.

Scorers were recruited through the services of one or more temporary employment agencies. All scorers signed a nondisclosure/confidentiality agreement.

5.4.2 Scorer Training

Scorer training began with an introduction of on-site scoring staff and an overview of the NECAP Science program's purpose and goals, including a discussion about the security, confidentiality, and proprietary nature of testing, scoring materials, and procedures.

Next, scorers thoroughly reviewed and discussed the scoring guide for the item to be scored. Each item-specific scoring guide included the item itself and score point descriptions.

Following review of the item-specific scoring guide for any constructed-response item, scorers began reviewing or scoring response sets organized for specific training purposes:

- Anchor set
- Practice set
- Qualifying set

5.4.2.1 Anchor Set

Scorers first reviewed an anchor set of exemplary responses, approved by the state science specialists, for the item to be scored. Responses in anchor sets were typical rather than unusual or uncommon; solid rather than controversial or borderline; and true, meaning that they had scores that could not be changed by anyone other than the NECAP Science client and Measured Progress test development staff.

For constructed-response items, each item-specific anchor set contained, for each respective score point, a client-approved sample response that was to be considered a midrange example of its respective score point. When necessary, a second sample response was included to illustrate a substantially alternate way to achieve that score point.

Responses were read aloud to the room of scorers and presented in descending score order. Trainers then announced the true score of each anchor response and facilitated a group discussion of the response in relation to the score point descriptions to allow scorers to internalize typical characteristics of each score point.

This anchor set served as a reference for scorers as they continued with calibration, scoring, and recalibration activities for that item.

5.4.2.2 Practice Set

Next, scorers practiced applying the scoring guide and anchors to responses in the practice set. The practice set included student responses designed to help establish the score point range and the range of responses within each score point. The practice set often represented unusual responses that were less clear or solid (e.g., were shorter than normal, employed atypical approaches, contained both very low and very high attributes, or included difficult handwriting). Responses in the practice set were presented in randomized score point order.

After scorers had independently read and scored a practice set response, trainers polled scorers or used online training system reports to record the initial range of scores. Then they led a group discussion of one or two responses, directing scorer attention to scoring issues that were particularly relevant to the specific scoring group, such as the line between two score points. Trainers modeled for scorers how to discuss scores by referring to the anchor set and scoring guides.

5.4.2.3 Qualifying Set

After the practice set had been completed for an item, scorers were required to measurably demonstrate their ability to accurately and reliably score all responses in the item's qualifying set, according to the appropriate anchor set in concert with its scoring rubric, by scoring the qualifying set. The qualifying set consisted of 10 responses selected from an array of responses that clearly illustrated the range of score points for that item. The set was chosen in accordance with the responses reviewed and approved by the state specialists.

To be eligible to score operational 2014–15 NECAP Science responses, scorers were required to demonstrate scoring accuracy rates of at least 80 percent exact agreement and at least 90 percent exact or adjacent agreement across all qualifying set responses. In other words, exact scores were required on at least eight of the qualifying set responses and either exact or adjacent scores were required on at least nine. Scorers were allowed one discrepant score as long as they had at least eight exact scores.

5.4.2.4 Retraining

Scorers who did not pass the first qualifying set were, at the discretion of scoring leadership, retrained as a group by reviewing their performance with scoring leadership and then scoring a second qualifying set of responses. If they achieved a minimum scoring accuracy rate of 80 percent exact and 90 percent exact or adjacent agreement on this second set, they were allowed to score operational responses.

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

5.4.3 Scoring Supervisor and STL Training

Scoring Supervisors and select STLs were trained in a separate session that occurred immediately prior to scorer training. In addition to discussing the items and their responses, Scoring Supervisor and STL training included emphasis on the states' rationale behind the score points. This rationale was discussed in greater detail with Scoring Supervisors and STLs than with regular scorers to better equip leadership to handle questions from the scorers.

5.4.4 Benchmarking Meetings

In preparation for implementing NECAP Science guidelines for the scoring of field-test responses, Measured Progress scoring staff prepared and facilitated benchmarking meetings held with the NECAP state science specialists. The purpose of the meetings was to establish item-specific guidelines for scoring each NECAP Science item for the current field-test scoring session and for future operational scoring sessions.

Prior to these meetings, scoring staff collected a set of several dozen student responses that Scoring Content Specialists identified as being illustrative midrange examples of their respective score points. The Scoring Content Specialists and science specialists worked collaboratively during benchmarking meetings to finalize an authoritative set of score point exemplars for each field-test item. As a matter of practice, each of these authoritative sets is included as part of the scoring training materials and used to train scorers each time that item is scored—both as a field-test item and as part of a future NECAP Science administration.

This repeated use of approved sets of midrange score point exemplars helps ensure that, each time a particular NECAP Science item is scored, scorers follow the guidelines established by the state science specialists.

5.5 METHODOLOGY FOR SCORING CONSTRUCTED-RESPONSE ITEMS

Constructed-response items were scored based on possible score points and scoring procedures, as shown in Table 5-2.

Table 5-2. 2014–15 NECAP Science: Possible Score Points for Open-Response Items

Item Type	*Possible*	
	Score Points	*Highest Score*
Constructed-response	0–4	4
Inquiry task—constructed-response	0–3	3
Inquiry task—short-answer	0–2	2
Nonscorable	0	0

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).[1]
- Response was written in a language other than English.
- Response was completely off task or off topic.
- Response included an insufficient amount of material to make scoring possible.
- Response was an exact copy of the assignment.
- Response was incomprehensible.
- Student made a statement refusing to write a response to the question.

[1] Unreadable and wrong-location responses were eventually resolved, whenever possible, by researching the actual answer document (electronic copy or hard copy, as needed) to identify the correct location or to more closely examine the response and then assign a score.

Scoring Procedures

Scoring procedures for constructed-response items included both single scoring and double scoring. Single-scored items were scored by one scorer. Double-scored items were scored independently by two scorers whose scores were tracked for agreement (known as interrater agreement). For further discussion of double scoring and interrater agreement, see Subsection 5.5.1.3.

Table 5-3 shows the method(s) by which common and equating constructed-response items for each operational test were scored.

Table 5-3. 2014–15 NECAP Science: Methods of Scoring Common and Equating Constructed-Response Items by Grade and Test

Grade	*Test/Field Test Name*	*Responses*	
		*Single Scored**	*Double Scored**
4	Science	100%	2% randomly
8	Science	100%	2% randomly
11	Science	100%	2% randomly
All	Unreadable responses	100%	100%
All	Blank responses	100%	100%

* Per grade and test/field test.

For each field-test item, 1,500 student responses were scored.

5.5.1 Monitoring of Scoring Quality Control and Consistency

Scorers were monitored for continued accuracy rates and scoring consistency throughout the scoring process, using the following methods and tools:

- Embedded committee-reviewed responses (CRRs)
- Read-behind procedures
- Double-blind scoring
- Recalibration sets
- Scoring reports

If scorers met or exceeded the expected accuracy rate, they continued scoring operational responses. Any scorer who fell below the expected accuracy rate for the particular item and monitoring method was retrained on that item and, upon approval by the Scoring Supervisor or Scoring Content Specialist as appropriate, was allowed to resume scoring.

It is important to note the difference between the accuracy rate each scorer must have achieved to qualify for scoring live responses and the accuracy rate each scorer must have maintained to continue scoring live responses. Specifically, the qualification accuracy rate was stricter than the live scoring accuracy rate. The reason for this difference is that an “exact score” in double-blind statistics requires that two scorers both

identify the same score for a response; an exact score during qualification requires that an individual scorer match the score predefined by leadership. Thus, the latter is dependent on matching an expert, not a peer.

During live scoring, scorer accuracy rates are monitored using an array of techniques, thereby providing a more complete picture of a scorer's performance than would be the case if relying on just one technique. These techniques are described in the next subsections.

5.5.1.1 Embedded CRRs

Previously scored CRRs were selected and loaded into iScore for blind distribution to scorers as a way to monitor accuracy. Embedded CRRs, either chosen before scoring had begun or selected by leadership during scoring, were inserted into the scoring queue so as to be indistinguishable from all other live student responses. Up to 10 embedded CRRs were distributed at random points throughout the first full day of scoring an item to ensure that scorers were sufficiently calibrated at the beginning of the scoring period. If any scorer fell below the required live scoring accuracy rate, he or she was retrained before being allowed by the Scoring Supervisor to continue. Once the scorer was allowed to resume scoring, leadership carefully monitored him or her by increasing the number of read-behinds.

5.5.1.2 Read-Behind Procedures

Read-behind scoring refers to the practice of scoring leadership, usually an STL, scoring a response after a scorer has already scored it.

Responses to be placed into the read-behind queue were randomly selected by scoring leadership; scorers were not made aware as to which of their responses would be reviewed by their STL. The iScore system allowed one, two, or three responses per scorer at a time to be placed into the read-behind queue.

The STL entered his or her score into iScore before being allowed to see the score assigned by the scorer for whom the read-behind was being performed. The STL then compared the two scores, and the ultimate reported score was determined as follows:

- If there was exact agreement between the scores, no action was taken; the regular scorer's score remained.
- If the scores were adjacent (i.e., the difference was not greater than 1), the STL's score became the score of record. If there was a significant number of adjacent scores for this scorer across items, an individual scoring consultation was held with the scorer, and the Scoring Supervisor determined whether or when the scorer could resume scoring.
- If there was a discrepant difference between the scores (greater than 1 point), the STL's score became the score of record. An individual consultation was held with the scorer, and the Scoring Supervisor determined whether or when the scorer could resume scoring.

These three scenarios are illustrated in Table 5-4.

Table 5-4. 2014–15 NECAP Science: Examples of Read-Behind Scoring Resolutions

Scorer	*Leadership Resolution*	*Final**
4	4	4
4	3	3
4	2	2

* Leadership score is score of record.

Scoring Team Leaders conducted read-behinds on all scorers assigned to their group, and continued to conduct these read-behinds throughout the course of scoring. STLs focused additional read-behinds on scorers who were in need of extra coaching and attention.

In addition to the daily read-behinds, scoring leadership could choose to read behind any scorer at any point during the scoring process and thereby take an immediate, real-time snapshot of a scorer's accuracy.

5.5.1.3 Double-Blind Scoring

Double-blind scoring refers to the practice of two scorers independently scoring a response, each without knowing the response had already been or soon would be scored by another scorer. Table 5-3 provides information about the proportion of responses that were double scored.

If there was a discrepancy (a difference greater than 1) between scores, the response was placed in an arbitration queue. Arbitration responses were reviewed by scoring leadership (STL or Scoring Supervisor) without any background knowledge of the scores assigned by the two previous scorers.

Scoring leadership consulted individually with any scorer whose scoring rates on the different monitoring methods fell below the required accuracy percentage, and the Scoring Supervisor determined whether or when the scorer could resume scoring. Once the scorer was allowed to resume scoring, leadership carefully monitored him or her by increasing the frequency of read-behinds.

5.5.1.4 Recalibration Sets

To determine whether scorers were still calibrated to the scoring standard, scorers were required to take an online recalibration set at the start and midpoint of the shift of their resumption of scoring.

Each recalibration set consisted of five responses representing the entire range of possible scores, including some with a score point of 0.

- Scorers who were discrepant on two of five responses of the first recalibration set, or exact on two or fewer, were not permitted to score on that item that day and were either assigned to a different item or dismissed for the day.
- Scorers who were discrepant on only one of five responses of the first recalibration set, and/or exact on three, were retrained by their STL by discussing the recalibration set responses in terms of the score point descriptions and the original anchor set. After this retraining, such scorers began scoring operational responses under the proviso that the

scorer's scores for that day and that item would be kept only if the scorer was exact on all five of five responses of the second recalibration set administered at the shift midpoint. The Scoring Supervisor determined whether or when these scorers had received enough retraining to resume scoring operational responses. Scoring leadership also carefully monitored the accuracy of such scorers by significantly increasing the number of their read-behinds.

- Scorers who were not discrepant on any response of the first recalibration set, and exact on at least four, were allowed to begin scoring operational responses immediately, under the proviso that this recalibration performance would be combined with that of the second recalibration set administered at the shift midpoint.

The results of both recalibration sets were combined with the expectation that scorers would have achieved an overall 80 percent-exact and 90 percent-adjacent standard for that item for that day.

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

Recalibration sets were employed for all constructed-response items. They were not used for 2-point short-answer items, for which read-behind and double-blind techniques are more informative and cost effective.

5.5.2 Scoring Reports

Measured Progress's electronic scoring software, iScore, generated multiple reports that were used by scoring leadership to measure and monitor scorers for scoring accuracy, consistency, and productivity.

Reports Generated During Scoring

Because the 2014–15 NECAP Science test administration was complex, computer-generated reports were necessary to ensure all of the following:

- Overall group-level accuracy, consistency, and reliability of scoring
- Immediate, real-time individual scorer data availability for early scorer intervention when necessary
- Scoring schedule maintenance

The following reports were produced by iScore:

- The **Read-Behind Summary** showed the total number of read-behind responses for each scorer and noted the numbers and percentages of scores that were exact, adjacent, and discrepant between that scorer and the STL or Scoring Supervisor. Scoring leadership could choose to generate this report by selecting options such as "Today," "Past Week," or "Cumulative" from a pull-down menu. The report could also be filtered to display 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 (i.e., sent back out to the floor to be rescored by other scorers). The benefit of this report is that it measures the degree to which

individual scorers agree with his or her Scoring Supervisor or STL on how to best score live responses.

- The **Double-Blind Summary** showed the total number of double-score responses scored by each scorer and noted the numbers and percentages of scores that were exact, adjacent, and discrepant between that scorer and the second scorer. This report was used in conjunction with other reports to determine whether a scorer's scores would be voided. The benefit of this report is that it reveals the degree to which scorers are in agreement with each other about how to best score live responses.
- The **Accuracy Summary** combined read-behind and double-score data, showing the total number of read-behind and double-score responses scored for each scorer and noting his or her accuracy percentages and score point distributions.
- The **Embedded CRR Summary** showed, for each scorer and for either a particular item or across all items, the total number of responses scored, the number of embedded CRRs scored, and the numbers and percentages of scores that were exact, adjacent, and discrepant between the scorer and the Scoring Content Specialist (by virtue of the Scoring Content Specialist's approval of the prescored embedded CRRs). This report was used in conjunction with other reports to determine whether a scorer's scores would be voided. The benefit of this report is that it measures the degree to which individual scorers agree with their Scoring Content Specialist on how to best score live responses—and since embedded responses are administered during the first hours of scoring, this report provides an early indication of agreement between scorers and their Scoring Content Specialist.
- The **Qualification Statistics Report** listed all scorers by name and ID number, identifying which qualifying set(s) they did and did not take and, for the ones they did take, whether they passed or failed. The total number of qualifications passed and failed was noted for each scorer, as was the total number of individuals passing or failing a particular qualifying set. The Scoring Supervisor could use this report to determine how the scorers within his or her specific scoring group performed on a specific qualifying set.
- The **Summary Report** showed the total number of student responses for an item and identified, for the time at which the report was generated, (1) the number of single and double scorings that had been performed and (2) the number of single and double scorings yet to be performed.

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 the quality of a test 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* (2004) include standards for identifying quality items. Items should assess only knowledge or skills that are identified as part of the domain being tested and should avoid assessing irrelevant factors. Items should also be unambiguous and free of grammatical errors, potentially insensitive content or language, and other confounding characteristics. In addition, items must not unfairly disadvantage students, in particular racial, ethnic, or gender groups.

Both qualitative and quantitative analyses are conducted to ensure that NECAP Science test 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 NECAP Science test in spring 2015. Note that the information presented in this chapter is based on the items common to all forms, since those are the items on which student scores are calculated. (Item analyses are also performed for field-test items, and the statistics are then used during the item review process and form assembly for future administrations.)

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); so, for these items, the difficulty index is simply the proportion of students who correctly answered the item. Polytomously scored items include short-answer, for which students can receive scores of 0, 1, or 2, and constructed-response, which are worth either 3 or 4 points total. 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 0 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 NECAP 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 0.0 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. 2014–15 NECAP Science: Summary of Item Difficulty and Discrimination Statistics by Grade

Grade	*Item Type*	*Number of Items*	p-*Value*		*Discrimination*	
			Mean	*Standard Deviation*	*Mean*	*Standard Deviation*
4	ALL	44	0.61	0.17	0.36	0.09
	CR	5	0.45	0.12	0.50	0.05
	MC	33	0.67	0.11	0.34	0.07
	SA	6	0.37	0.21	0.37	0.10
8	ALL	44	0.56	0.13	0.40	0.12
	CR	5	0.37	0.12	0.59	0.05
	MC	33	0.61	0.11	0.34	0.08
	SA	6	0.46	0.08	0.54	0.05
11	ALL	44	0.50	0.17	0.38	0.12
	CR	5	0.36	0.10	0.56	0.05
	MC	33	0.56	0.15	0.34	0.09
	SA	6	0.31	0.06	0.50	0.03

A comparison of indices across grade levels is complicated because these indices are population dependent. Direct comparisons would require that either the items or students were common across groups. Since that is not the case, it cannot be determined whether differences in performance across grade levels are because of differences in student abilities, differences in item difficulties, or both. With this caveat in mind, it appears generally that students in grades 8 and 11 found their items more difficult than students in grade 4.

Comparing the difficulty indices of multiple-choice items and open-response (short-answer or 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 open-response items. Similarly, discrimination indices for the open-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 F; 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 were a small number of items with low discrimination indices, but none were 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 NECAP Science test. Item-level score point distributions are provided for open-response items in Appendix G; for each item, the percentage of students who received each score point is presented.

6.2 DIFFERENTIAL ITEM FUNCTIONING (DIF)

Code of Fair Testing Practices in Education (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, NECAP Science items were evaluated in terms of DIF statistics.

For NECAP 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 open-response items. Dorans and Holland (1993) suggested that index values between -0.05 and 0.05 should be considered negligible. The preponderance of NECAP 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 to 0.10 range (i.e., "high" DIF) are more unusual and should be examined very carefully.[2]

For the 2014–15 NECAP Science tests, five subgroup comparisons were evaluated for DIF:

- Male versus female
- No disability versus disability
- Noneconomically disadvantaged versus economically disadvantaged
- Non-LEP versus LEP
- White versus Asian

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

6.3 DIMENSIONALITY ANALYSES

The NECAP 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,

[2] It should be pointed out here 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 Measured Progress content specialist. The content specialist consults with the agency/department of education to determine whether to include the flagged item in a future operational test administration.

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 2014-15 NECAP 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 and (b) the nature of the multidimensionality. Findings from dimensionality analyses performed on the 2014-15 NECAP science common items for grades 4, 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 covariance 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 constructed-response (CR) items, we would expect MC items to have positive conditional covariances with each other, CR items to have positive conditional covariances with each other, and MC items to have negative conditional covariances with CR items.

DIMTEST is a hypothesis-testing procedure for detecting violations of local independence. For the exploratory analyses conducted for the NECAP 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 NECAP 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).

DIMTEST and DETECT were applied to the spring 2014-15 NECAP Science assessments for grades 4, 8, and 11. The data for each grade were split into a training sample and a cross-validation sample. Each grade had at least 28,000 student examinees. Because DIMTEST was limited to using 24,000 students, the training and cross-validation samples for the DIMTEST analyses used 12,000 each, randomly sampled from the total sample. DETECT, on the other hand, had an upper limit of 50,000 students, so every training sample and cross-validation sample used with DETECT had at least 14,000 students. DIMTEST was then applied to each of the science grades. 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 the NECAP science tests, DIMTEST would be sensitive to even quite small violations of unidimensionality, and the null hypothesis was rejected at a significance level of 0.01 for every dataset. 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 6-2 below displays the multidimensional effect size estimates from DETECT.

As shown in Table 6-2, all of the DETECT values indicate multidimensionality is either very weak (DETECT values less than 0.20) or weak (DETECT values between 0.20 and 0.30) for every test analyzed. Also shown in Table 6-2 are the DETECT values from last year's dimensionality analysis. This year's results are seen to be similar to last year's in that both sets of results indicated weak or very weak multidimensionality for every test.

Table 6-2. 2014–15 NECAP Science: Multidimensionality Effect Sizes

Grade	*Multidimensionality Effect Size*	
	2014–15	*2013–14*
4	0.14	0.19
8	0.23	0.19
11	0.21	0.22

We also investigated how DETECT divided the tests into clusters to see if there were any discernable patterns with respect to the item types (multiple choice as compared to constructed response). In all grades, the multiple-choice items and constructed-response items tended to cluster separately from each other. This separation has occurred in all seven years of NECAP science dimensionality analyses. In 21 analyses (7 years x 3 grade levels), only once has more than 25% of the CR points occurred in an MC-dominated cluster; and never has more than 10% of the MC points occurred in a CR-dominated cluster. Despite this evidence of multidimensionality between the MC and CR items, the multidimensional effect sizes are weak, so no changes in test design, scoring or administration are warranted.

CHAPTER 7 SCALING AND EQUATING

This chapter describes the procedures used to calibrate, equate, and scale the NECAP Science assessment. During the course of these psychometric analyses, a number of quality-control procedures and checks on the processes were implemented. These procedures included evaluations 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, examination of 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, examination of α-plots and b-plots for reasonableness); and evaluation of the scaling results (e.g., parallel processing by the Measured Progress Psychometrics and Research and Data and Reporting Services Departments, comparing look-up tables to the previous year's). An equating report, which provided complete documentation of the quality-control procedures and results, was submitted to the agency/departments of education for their 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. The number of items identified for evaluation was very typical across the grades. Descriptions of the evaluations and results are included in the Item Response Theory (IRT) Results and Equating Results sections below.

Table 7-1. 2014–15 NECAP Science: Items That Required Intervention During IRT Calibration and Equating

Grade	*Item Number*	*Reasons*	*Action*
4	174255	c-parameter	set c = 0
	263765	a-parameter	a set to initial
	60348	c-parameter	set c = 0
	87135	c-parameter	set c = 0
8	144554	c-parameter	set c = 0
	144589	c-parameter	set c = 0
	175376	c-parameter	set c = 0
	219445	c-parameter	set c = 0
	256610	c-parameter	set c = 0
	258033	c-parameter	set c = 0
	50012	b/b analysis	removed from equating
11	176774	delta analysis	removed from equating
	177090	c-parameter	set c = 0
	242618	c-parameter	set c = 0
	46140	c-parameter	set c = 0
	47917	c-parameter	set c = 0
	48214	c-parameter	set c = 0
	60181	c-parameter	set c = 0
	89407	b/b analysis	removed from equating

7.1 ITEM RESPONSE THEORY (IRT)

All NECAP Science items were calibrated using IRT. The IRT uses mathematical models to define a relationship between an unobserved measure of student performance, usually referred to as theta (¸), 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 ¸). Another way to think of ¸ is as a mathematical representation of the latent trait of interest. Several common IRT models are used to specify the relationship between ¸ and p (Hambleton & Swaminathan, 1985; Hambleton & van der Linden, 1997). The process of determining the specific mathematical relationship between ¸ 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 ¸ and p. Once the item parameters are known, an estimate of ¸ 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 2014–15 NECAP Science assessment, the three-parameter logistic (3PL) model was used for dichotomous (multiple-choice) items and the graded-response model (GRM) was used for polytomous (open-response) 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, denoted as m 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 m categories can be characterized by k item category threshold curves (ICTCs) of the two-parameter logistic form:

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

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

$$P_{ik}\left(k\middle|\theta_j\right) = \frac{\exp\left[D\alpha_i\left(\theta_j - b_i + d_k\right)\right]}{1+\exp\left[D\alpha_i\left(\theta_j - b_i + d_k\right)\right]} - \frac{\exp\left[D\alpha_i\left(\theta_j - b_i + d_{k+1}\right)\right]}{1+\exp\left[D\alpha_i\left(\theta_j - b_i + d_{k+1}\right)\right]}.$$

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:

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

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

7.2 IRT Results

The tables in Appendix I give the IRT item parameters of all common items on the 2014–15 NECAP Science tests by grade. In addition, Appendix J shows graphs of the TCCs and TIFs, which are defined below.

The 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

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

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 to 4), and
$E\left(X\middle|\theta_j\right)$ 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 above lists items that were flagged based on the quality-control checks implemented during the calibration process. (Note that some items were flagged as a result of the evaluations of the equating items; those results are described below.) In most cases, items flagged during this step were identified because the pseudo-guessing parameter (*c*-parameter) was poorly estimated. Difficulty in estimating the *c*-parameter is not at all unusual and is well documented in the psychometric literature (see, e.g., Nering & Ostini, 2010), especially when the item's discrimination is below 0.50. In all cases, fixing the *c*-parameter resulted in reasonable and stable item parameter estimates and improved model fit.

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

Table 7-2. 2014–15 NECAP Science: Number of Newton Cycles Required for Convergence

Subject	*Grade*	*Cycles*
Science	4	45
	8	73
	11	64

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 given an unfair advantage or disadvantage because the test form they took is easier or harder than those taken by other students.

The 2014–15 administration of NECAP 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 2014–15 NECAP Science tests are not equivalent to the groups who took them in the reference years. IRT is particularly useful for equating scenarios that involve nonequivalent groups (Allen & Yen, 1979). Equating for NECAP 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. Subsets of the equating items are distributed across forms.

Item parameter estimates for 2014–15 were placed on the 2013–14 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 2013–14 and 2014–15 NECAP Science tests should have the same item parameters. After the item parameters for each 2014–15 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 2014–-15 TCC for the equating items was as close as possible to that of 2013–-14.

7.4 EQUATING RESULTS

Prior to calculating the Stocking and Lord (1983) transformation constants, a variety of evaluations of the equating items were conducted. Items that were flagged as a result of these evaluations are listed in Table 7-1 at the beginning of this chapter. These items were scrutinized and a decision was made as to whether to include the item as an equating item or to discard it. The procedures used to evaluate the equating items are described below.

Appendix K presents the results from the delta analysis, applied to both the dichotomous and polytomous equating items. This procedure was one of several used to evaluate the adequacy of the equating items; the discard status presented in the appendix tables indicates whether or not the item was flagged as potentially inappropriate for use in equating based on the delta analysis alone. The discard status presented in the plots, however, indicates whether an item was flagged by any procedures used to evaluate the equating items. Also presented in Appendix K 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 well below the criterion value for excluding an item as an equating item, 0.80 in absolute value.

In addition to the delta and rescore analyses, evaluations based on analyzing a-plots and b-plots, which show IRT parameters for 2014–15 plotted against the values for 2013–14, were conducted and the

results are presented in Appendix L. Any items that appeared as outliers in the plots were evaluated in terms of suitability for use as equating items.

Finally, α-plots and *b*-plots, which show IRT parameters for 2014–15 plotted against the values for 2013–14, are presented in Appendix L. Any items that appeared as outliers in the plots were evaluated in terms of suitability for use as equating items.

Once all flagged items had been evaluated and appropriate action taken, the Stocking and Lord (1983) method of equating was used to place the item parameters onto the previous year's scale, as described earlier. The Stocking and Lord transformation constants are presented in Table 7-3.

Table 7-3. 2014–15 NECAP Science: Stocking and Lord Transformation Constants

Content Area	*Grade*	α-*Slope*	b-*Intercept*
Science	4	1.052	-0.137
	8	0.982	-0.012
	11	1.050	0.202

The next administration of NECAP Science (2015–16) will be scaled to the 2014–15 administration using the same equating method described above.

7.5 ACHIEVEMENT STANDARDS

Cutpoints for NECAP Science to establish the four achievement levels (Substantially Below Proficient, Partially Proficient, Proficient, and Proficient With Distinction) were set at a standard setting meeting held in August 2008. Details of the standard setting procedures can be found in the standard setting report created at that time, as well as in the 2007–08 technical report. The cuts on the theta scale that were established via standard setting and used for reporting in 2014–15 are presented in Table 7-4. Also shown in the table are the cutpoints on the reporting score scale (described below). These cuts will remain fixed throughout the assessment program unless standards are reset for any reason.

Table 7-4. 2014–15 NECAP Science: Cut Scores on the Theta Metric and Reporting Scale by Grade

Grade	Theta			Scaled Score				
	Cut 1	Cut 2	Cut 3	Minimum	Cut 1	Cut 2	Cut 3	Maximum
4	-1.222	0.048	2.371	400	427	440	463	480
8	-0.612	0.751	2.578	800	829	840	855	880
11	-0.432	0.788	2.193	1100	1130	1140	1152	1180

Table M-1 in Appendix M shows achievement level distributions by grade. Results are shown for each of the past three years.

7.6 REPORTED SCALED SCORES

Because the ¸ scale used in IRT calibrations is not readily understood by most stakeholders, reporting scales were developed for NECAP Science. The reporting scales are simple linear transformations of the underlying ¸ scale. The reporting scales are developed such that they range from *x*00 through *x*80 (where *x* is grade level). In other words, grade 4 scaled scores range from 400 through 480, grade 8 from 800 through 880, and grade 11 from 1100 through 1180. The lowest scaled score in the Proficient range is fixed at *x*40 for each grade level. For example, to be classified in the Proficient achievement level or above, a minimum scaled score of 440 was required at grade 4, 840 at grade 8, and 1,140 at grade 11.

By providing information that is more specific about the position of a student's results, scaled scores supplement achievement level scores. School- and district-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 2014–15 NECAP Science tests were translated to scaled scores using a data analysis process called scaling. Scaling simply converts from one scale to another. In the same way that a given temperature can be expressed on either Fahrenheit or Celsius scales, or the same distance can be expressed in either miles or kilometers, student scores on the 2014–15 NECAP 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 NECAP Science are reported instead of raw scores. Scaled scores make consistent the reporting of results. To illustrate, standard setting typically results in different raw cut scores across grades. The raw cut score between Partially Proficient and Proficient could be, say, 38 in grade 4 and 42 in grade 8, yet both of these raw scores would be transformed to scaled scores of *x*40 (i.e., 440 and 840). 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 the fact that they are linear transformations of ¸. Since the ¸ scale is used for equating, scaled scores are comparable from one year to the next. Raw scores are not.

The scaled scores are obtained by a simple translation of ability estimates ($\hat{\theta}$) using the linear relationship between threshold values on the ¸ metric and their equivalent values on the scaled score metric.

Students' ability estimates are based on their raw scores and are found by mapping through the TCC. Scaled scores are calculated using the linear equation

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

where
m is the slope and
b is the intercept.

A separate linear transformation is used for each grade level. For NECAP Science, the transformation function is determined by fixing the Partially Proficient/Proficient cut score and the bottom of the scale—that is, the *x*40 and the *x*00 values (e.g., 440 and 400 for grade 4). The *x*00 location on the ¸ scale is beyond (i.e., below) the scaling of all items. To determine this location, a chance score (approximately equal to a student's expected performance by guessing) is mapped to a value of -4.0 on the ¸ scale. A raw score of 0 is also assigned a scaled score of *x*00. The maximum possible raw score is assigned a scaled score of *x*80 (e.g., 480 in the case of grade 4). Because only two points within the ¸ scaled score space are fixed, the scaled score cuts between Substantially Below Proficient and Partially Proficient and between Proficient and Proficient With Distinction are free to vary across grades.

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

Table 7-5. 2014–15 NECAP Science: Scaled Score Slope and Intercept by Subject and Grade

Content Area	*Grade*	*m-Slope*	*b-Intercept*
Science	4	9.881	439.524
	8	8.420	833.678
	11	8.354	1133.414

Appendix N contains raw score to scaled score look-up tables for the 2014–15 NECAP Science tests. These are the actual tables used to determine student scaled scores, error bands, and achievement levels.

Appendix O contains scaled score distribution graphs for each grade. These distributions were calculated using the sparse data matrix files that were used in the IRT calibrations. As the curves move to the right, they represent an increase in performance.

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 forms 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), which 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 2014–15 NECAP Science test:

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

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 Error of Measurement

Table 8-1 presents descriptive statistics, Cronbach's α coefficient, and raw score standard errors of measurement (SEM) for each grade. (Statistics are based on common items only.)

Table 8-1. 2014–15 NECAP Science: Raw Score Descriptive Statistics, Cronbach's Alpha, and SEM by Grade

Grade	*Number of Students*	*Raw Score*			*Alpha*	*SEM*
		Maximum	*Mean*	*Standard Deviation*		
4	29,673	63	35.03	10.32	0.88	3.59
8	30,185	63	32.21	11.62	0.91	3.58
11	28,266	63	28.67	11.21	0.90	3.63

Because 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 2014–15 Subgroup Reliability

The reliability coefficients discussed in the previous section were based on the overall population of students who took the 2014–15 NECAP Science test. Appendix P presents reliabilities for various subgroups of interest. Subgroup Cronbach's α's were calculated using the formula defined earlier 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 grades 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 also on the statistical distribution of the studied subgroup. For example, it can be readily seen in Appendix P that subgroup sample sizes may vary considerably, which results in natural variation in reliability coefficients. Or 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 Reporting Subcategory Reliability

Of even more interest are reliabilities for the reporting subcategories within NECAP Science content areas, described in Chapter 3. 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, approximately to the degree one would expect based on classical test theory. Qualitative differences between grades once again preclude valid inferences about the quality of the full test based on statistical comparisons among subtests.

8.4 Interrater Consistency

Chapter 5 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 2 percent 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 NECAP Science test. 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. The table shows the number of score categories, number of included scores, percent exact agreement, percent adjacent agreement, correlation between the first two sets of scores, and percent of responses that required a third score. This same information is provided at the item level in Appendix Q.

Table 8-2. 2014–15 NECAP 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		
4	3	3,717	82.59	16.63	0.82	0.70
	4	1,238	70.27	27.95	0.74	1.78
	5	1,955	69.62	27.06	0.87	3.17
8	3	3,709	81.02	17.55	0.78	1.46
	4	1,216	74.26	23.85	0.85	1.48
	5	1,784	63.23	33.46	0.74	3.03
11	3	3,318	71.94	26.37	0.63	1.60
	4	1,084	80.26	18.36	0.90	1.29
	5	1,660	73.07	24.82	0.85	1.99

8.5 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 NECAP Science, students are classified into one of four achievement levels: Substantially Below Proficient, Partially Proficient, Proficient, or Proficient with Distinction. This section of the report explains the methodologies used to assess the reliability of classification decisions.

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 consistency of classification decisions based on a single administration of a test. The Livingston and Lewis (1995) technique was used for the 2014–15 NECAP Science assessment because it is easily adaptable to all types of testing formats, including mixed-format tests.

The accuracy and consistency estimates reported in Appendix R 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 2014–15 NECAP Science test, 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 are other consistency estimates.

8.5.1 Accuracy and Consistency Results

The accuracy and consistency analyses described above are provided in Table R-1 of Appendix R. The table includes overall accuracy and consistency indices, including kappa. Accuracy and consistency values conditional upon achievement level are also given. For these calculations, the denominator is the proportion of students associated with a given achievement level. For example, the conditional accuracy value is 0.86 for Substantially Below Proficient for grade 8. This figure indicates that among the students whose true scores placed them in this classification, 86 percent would be expected to be in this classification when categorized according to their observed scores. Similarly, a consistency value of 0.80 indicates that 80 percent of students with observed scores in the Substantially Below Proficient 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 level thresholds. For example, in testing done for No Child Left Behind accountability purposes, the primary concern is distinguishing between students who are proficient and those who are not yet proficient. In this case, the accuracy of the Partially Proficient/Proficient threshold is of greatest interest. For 2014–15 NECAP Science, Table R-2 in Appendix R provides accuracy and consistency estimates at each cutpoint, as well as false positive and false negative decision rates. (A false positive is the proportion of students whose observed scores were above the cut and whose true scores were below the cut. A false negative is the proportion of students whose observed scores were below the cut and whose true scores were above the cut.)

The above indices are derived from Livingston and Lewis's (1995) method of estimating the accuracy and consistency of classifications. It should be noted that Livingston and Lewis discuss two versions of the accuracy and consistency tables. A standard version performs calculations for forms parallel to the form taken. An "adjusted" version adjusts the results of one form to match the observed score distribution obtained in the data. The tables use the standard version for two reasons: (1) this "unadjusted" version can be considered a smoothing of the data, thereby decreasing the variability of the results; and (2) for results dealing with the consistency of two parallel forms, the unadjusted tables are symmetrical, indicating that the two parallel forms have the same statistical properties. This second reason is consistent with the notion of forms that are parallel; that is, it is more intuitive and interpretable for two parallel forms to have the same statistical distribution.

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 R should be interpreted with caution. In addition, it is important to remember that it is inappropriate to compare DAC statistics between grades.

CHAPTER 9 SCORE REPORTING

The data used for the NECAP Science reports are the results of the spring 2015 administration of the NECAP Science test. NECAP Science tests are based on the NECAP Science assessment targets, which cover the grade spans K–4, 5–8, and 9–11. For example, the grade 8 NECAP Science test is based on the assessment targets of grades 5 through 8. Because the assessment targets cover grade spans, the state agency/departments of education determined that assessing science in the spring—as opposed to the fall, when mathematics, reading, and writing are assessed—would allow students and schools adequate time to cover all assessment targets through the curriculum and would also avoid a testing overload in the fall. All students who participated in NECAP Science were represented in testing year reports, because the students took the test in the school where they completed their learning of the assessment targets for their particular grade span.

9.1 PRIMARY REPORTS

Measured Progress created four primary reports for the 2014–15 NECAP Science test:

- Student Report
- Interactive Reporting
- School-, District-, and State Grade-Level Results Reports
- District and State Summary Report

With the exception of the Student Report, all reports were available for schools and districts to view or download on a password-secure website hosted by Measured Progress. Student-level data files were also available for districts to download. Each of these reports is described in the following subsections. Sample reports are provided in Appendix S.

9.2 STUDENT REPORT

The NECAP Science Student Report is a single-page, two-sided report printed on 8.5" by 11" paper. The front side of the report includes informational text about the design and uses of the assessment. It also describes the three corresponding sections of the reverse side of the report as well as the achievement levels. The reverse side provides a complete picture of an individual student’s performance on the NECAP Science test, divided into three sections. The first section provides the student’s overall performance for science. In addition to giving the student’s achievement level, it presents the scaled score numerically and in a graphic

that places the score, including its standard error of measurement (SEM), within the full range of possible scaled scores demarcated into the four achievement levels.

The second section of the report displays the student's achievement level in science relative to the percentage of students at each achievement level across the school, district, and state.

The third section shows the student's performance compared to school, district, and statewide performances in each of the four tested science domains: Physical Science, Earth Space Science, Life Science, and Scientific Inquiry.

Student performance is reported in the context of possible points: average points earned for the school, district, and state; and average points earned by students who are minimally proficient on the test (scaled score of 440, 840, or 1,140). The average points earned is reported as a range, because it is the average of all students who are minimally proficient, plus or minus one standard deviation.

To provide a more complete picture of the inquiry task portion of the science test (Session 3), each report includes a description of the inquiry task that was administered to all students at that grade. The grade 4 inquiry task always contains a hands-on experiment; the grade 8 inquiry task sometimes contains a hands-on experiment and sometimes contains a paper-and-pencil data analysis; and the grade 11 inquiry task always contains a paper-and-pencil data analysis.

The NECAP Student Report is confidential and should be kept secure within the school and district. The Family Educational Rights and Privacy Act (FERPA) requires that access to individual student results be restricted to the student, the student's parents/guardians, and authorized school personnel.

9.3 INTERACTIVE REPORTING

Four interactive reports were available: Item Analysis Roster, Achievement Level Summary, Released Items Summary Data, and Longitudinal Data. Each of these interactive reports is described in the following sections. Sample interactive reports are provided in Appendix T. To access these four interactive reports, the user clicked the interactive tab on the home page of the system and selected the report desired from the dropdown menu. Next, the user applied basic filtering options, such as the name of the district or school and the grade level/content area test, to open the specific report. At this point, the user had the option of printing the report for the entire grade level or applying advanced filtering options to select a subgroup of students to analyze. Advanced filtering options include gender, ethnicity, limited English proficient (LEP), individualized education program (IEP), and socioeconomic status (SES). All interactive reports, with the exception of the Longitudinal Data Report, allowed the user to provide a custom title for the report.

9.3.1 Item Analysis Roster Report

The NECAP Science Item Analysis Roster Report provides a roster of all students in a school and provides performance on the common items that are released to the public. The student names and

identification numbers are listed as row headers down the left side of the report. The items are listed as column headers in the same order they appeared in the released item document.

For each item, the following are shown:

- The Depth of Knowledge (DOK) code
- The item type
- The correct response key for multiple-choice items
- The possible points
- The content standard

For each student, multiple-choice items are marked either with a plus sign (+), indicating that the student chose the correct multiple-choice response, or a letter (from A to D), indicating the incorrect response chosen by the student. For short-answer and constructed-response items, the number of points earned is shown. All responses to released items are shown in the report, regardless of the student's participation status. The columns on the right side of the report show the Total Test results, broken into several categories. Subcategory Points Earned columns show points earned by the student in each content area subcategory relative to total possible points. A Total Points Earned column is a summary of all points earned and total possible points in the content area. The last two columns show the student's scaled score and achievement level. Students reported as Not Tested are given a code in the achievement level column to indicate the reason the student was not tested. It is important to note that not all items used to compute student scores are included in this report; only released items are included. At the bottom of the report, the average percentage correct for each multiple-choice item and average scores for the short-answer and constructed-response items are shown for the school, district, and state. When the user applies advanced filtering criteria, the School and District Percent Correct/Average Score rows at the bottom of the report are blanked out and only the Group row and the State row for the group selected will contain data. This report can be saved, printed, or exported as a PDF.

The Item Analysis Roster is confidential and should be kept secure within the school and district. FERPA requires that access to individual student results be restricted to the student, the student's parents/guardians, and authorized school personnel.

9.3.2 Achievement Level Summary

The Achievement Level Summary provides a visual display of the percentages of students in each achievement level for a selected grade. The four achievement levels are represented by various colors in a pie chart. A separate table is also included below the chart that shows the number and percentage of students in each achievement level. This report can be saved, printed, or exported as a PDF file.

9.3.3 Released Items Summary Data Report

The Released Items Summary Data Report is a school-level report that provides a summary of student responses to the released items for a selected grade/content area. The report is divided into three sections by item type (Multiple-Choice, Constructed-Response, and Inquiry Task). For multiple-choice items, the total number/percent of students who answered the item correctly and the number of students who chose each incorrect option or provided an invalid response are reported. An invalid response on a multiple-choice item is defined as "the item was left blank" or "the student selected more than one option for the item." For constructed-response and inquiry task items, point value and average score for the item are reported. Users are also able to view the actual released items within this report. If a user clicks on a particular magnifying glass icon next to a released item number, a pop-up box will open displaying the released item.

9.3.4 Longitudinal Data Report

The Longitudinal Data Report is a confidential student-level report that provides individual student performance data for multiple test administrations. The state-assigned student identification number is used to link students across test administrations. Student performance on future test administrations will be included on this report over time. This report can be saved, printed, or exported as a PDF file. It should be noted that New Hampshire provides an alternate student identification number unique to the test administration year; therefore, no students are included in the New Hampshire Longitudinal Data Report.

9.4 School, District, and State Grade-Level Results Reports

The NECAP School Results Report and the NECAP District Results Report consist of three parts: the grade-level summary report (page 2), the content area results (page 3), and the disaggregated content area results (page 4).

The grade-level summary report provides a summary of participation in the NECAP and a summary of NECAP results. The participation section on the top half of the page shows the number and percentage of students who were enrolled on or after May 4, 2015. The total number of students enrolled is defined as the number of students tested plus the number of students not tested.

Data are provided for the following groups of students who are considered tested in NECAP:

- **Students Tested:** This category provides the total number of students tested.
- **Students Tested With an Approved Accommodation:** Students in this category tested with an accommodation and did not have their test invalidated.
- **Current LEP Students:** Students in this category are currently receiving LEP services.

- **Current LEP Students Tested With an Approved Accommodation:** Students in this category are currently receiving LEP services, tested with an accommodation, and did not have their test invalidated.
- **IEP Students:** Students in this category have an IEP.
- **IEP Students Tested With an Approved Accommodation:** Students in this category have an IEP, tested with an accommodation, and did not have their test invalidated.

Because students who were not tested did not participate, average school scores were not affected by nontested students. These students were included in the calculation of the percentage of students participating, but not in the calculation of scores. For students who participated in some but not all sessions of the NECAP Science test, overall raw and scaled scores were reported. These reporting decisions were made to support the requirement that all students participate in the NECAP testing program.

Data are provided for the following groups of students, who may not have completed the entire NECAP Science test:

- **Alternate Assessment:** Students in this category completed an alternate assessment for the 2014–15 school year.
- **Withdrew After May 4:** Students withdrawing from a school after May 4, 2015, may have taken some sessions of the NECAP Science test prior to their withdrawal from the school.
- **Enrolled After May 4:** Students enrolling in a school after May 4, 2015, may not have had adequate time to participate fully in all sessions of the NECAP Science test.
- **Special Consideration:** Schools received state approval for special consideration for an exemption on all or part of the NECAP Science test for any student whose circumstances were not described by the previous categories but for whom the school determined that taking the NECAP Science test would not be possible.
- **Other:** Occasionally, students did not complete the NECAP Science test for reasons other than those listed. These "other" categories were considered not state approved.

The results section, on the bottom half of the page, shows the number and percentage of students performing at each achievement level in science across the school, district, and state. In addition, a mean scaled score is provided across school, district, and state levels. For the district version of this report, the school information is blank.

The content area results page provides information on performance in the four tested science domains (Physical Science, Earth Space Science, Life Science, and Scientific Inquiry). The purpose of this section is to help schools determine the extent to which their curricula are effective in helping students achieve the particular standards and benchmarks contained in the NECAP Science assessment targets. Information about the content area for school, district, and state includes

- the total number of students enrolled, not tested for a state-approved reason, not tested for another reason, and tested;
- the total number and percentage of students at each achievement level (based on the number in the Tested column); and
- the mean scaled score.

Information about each science domain includes the following:

- **The total possible points for that domain.** To provide as much information as possible for each domain, the total number of points includes both the common items used to calculate scores and additional items in each category.
- **A graphic display of the percentage of total possible points for the school, state, and district.** In this graphic display, symbols represent school, district, and state performance. In addition, a line symbolizes the SEM. This statistic indicates how much a student's score could vary if the student were examined repeatedly with the same test (assuming that no learning were to occur between test administrations).

The disaggregated content area results pages present the relationship between performance and student reporting variables in science across school, district, and state levels. The report shows the number of students categorized as enrolled, not tested for a state-approved reason, not tested for another reason, and tested. The report also provides the number and percentage of students within each of the four achievement levels and the mean scaled score by each reporting category.

The list of student reporting categories is as follows:

- All Students
- Gender
- Primary Race/Ethnicity
- LEP Status
- IEP
- SES
- Migrant
- Title I
- 504 Plan

The data for achievement levels and mean scaled score are based on the number shown in the Tested column. Reporting categories data were provided by records linked to the student labels. Because performance is being reported by categories that can contain relatively low numbers of students, school personnel are advised, under FERPA guidelines, to treat these pages confidentially.

It should be noted that for New Hampshire and Vermont, no data were reported for the 504 plan. In addition, for Vermont, no data were reported for Title I.

9.5 District and State Summary Reports

The NECAP District Summary Report provides details on student performance for all grade levels of NECAP Science tested in the district. The purpose of the report is to help districts determine the extent to which their schools and students achieve the particular standards and benchmarks contained in the NECAP Science assessment targets. The NECAP District Summary Report contains no individual school data. The information provided includes

- the total number of students enrolled, not tested for a state-approved reason, not tested for another reason, and tested;
- the total number and percentage of students at each achievement level (based on the number in the Tested column); and
- the mean scaled score.

9.6 Decision Rules

To ensure that the reported results for the 2014–15 NECAP Science test are accurate relative to collected data and other pertinent information, a document that delineates analysis and reporting rules was created. These decision rules were observed in the analyses of test data and in reporting the test results. Moreover, these rules served as the main reference for quality-assurance checks.

The decision rules document used for reporting results of the May 2015 administration of the NECAP Science test is found in Appendix U.

The first set of rules pertains to general issues in reporting scores. Each issue is described, and pertinent variables are identified. The actual rules applied are described by the way they impact analyses and aggregations and their specific impact on each of the reports. The general rules are further grouped into issues pertaining to test items, school type, student exclusions, and number of students for aggregations.

The second set of rules pertains to reporting student participation. These rules describe which students were counted and reported for each subgroup in the student participation report.

9.7 Quality Assurance

Quality-assurance measures are embedded throughout the entire process of analysis and reporting. The data processor, data analyst, and psychometrician assigned to work on the NECAP Science assessment implement quality-control checks of their respective computer programs and intermediate products. Moreover, when data are handed off to different functions within the Psychometrics and Research and Data

and Reporting Services Departments, the sending function verifies that the data are accurate before handoff. 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 for science 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 achievement levels assigned are compared across all students for 100 percent agreement. Different exclusions assigned to students that determine whether each student receives scaled scores and/or is included in different levels of aggregation are also parallel processed. Using the decision rules document, two data analysts independently write a computer program that assigns students' exclusions. For each grade, the exclusions assigned by each data analyst are compared across all students. Only when 100 percent agreement is achieved can the rest of the data analysis be completed.

The third aspect of quality control involves the procedures implemented by the quality-assurance group to check the veracity and accuracy of reported data. Using a sample of schools and districts, the quality-assurance group verifies that reported information is correct. The step is conducted in two parts: (1) Verify that the computed information was obtained correctly through appropriate application of different decision rules, and (2) verify that the correct data points populate each cell in the NECAP Science reports. The selection of sample schools and districts for this purpose is very specific and can affect the success of the quality-control efforts. Two sets of samples are selected, though they may not be mutually exclusive.

The first set includes those that satisfy the following criteria:

- One-school district
- Two-school district
- Multischool district

The second set of samples includes districts or schools that have unique reporting situations, as indicated by decision rules. This set is necessary to check that each rule is applied correctly. The second set includes the following criteria:

- Private school
- Small school that receives no school report
- Small district that receives no district report
- District that receives a report but all schools are too small to receive a school report
- School with excluded (not tested) students
- School with homeschooled students

The quality-assurance group uses a checklist to implement its procedures. After the checklist is completed, sample reports are circulated for psychometric checks and program management review. The appropriate sample reports are then presented to the client for review and signoff.

CHAPTER 10 VALIDITY

Because interpretations of test scores, and not a test itself, are evaluated for validity, the purpose of *the 2014–15 NECAP Science Technical Report* is to describe several technical aspects of the NECAP 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 speaks to different aspects of validity but are not 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 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 3 and 4. Item alignment with NECAP 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 NECAP Science questions are aligned by educators from the member states to specific NECAP Science content standards 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, and all test coordinators and administrators are required to familiarize themselves with and adhere to all of the procedures outlined in the NECAP *Principal/Test Coordinator Manual* and *Test Administrator Manual*.

The scoring information in Chapter 5 describes the steps taken to train and monitor hand-scorers, as well as quality-control procedures related to scanning and machine scoring. However, studies of student response processes would also be helpful and might include an investigation of students' cognitive methods using think-aloud protocols.

Evidence based on internal structure is presented in great detail in the discussions of item analyses, reliability, and scaling and equating in Chapters 6 through 8. 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 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 reporting information in Chapter 9, as well as in the *Guide to Using the 2015 NECAP Science Reports*, which is a separate document. Each of these chapters speaks to the efforts undertaken to promote 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.

To further support the validation of the assessment program, additional studies might be considered to provide evidence regarding the relationship of NECAP Science results to other variables, including the extent to which scores from NECAP Science converge with other measures of similar constructs and the extent to which they diverge from measures of different constructs. Relationships among measures of the same or similar constructs can sharpen the meaning of scores and appropriate interpretations by refining the definition of the construct.

10.1 QUESTIONNAIRE DATA

External validity of the NECAP 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 NECAP Science test. Note that no inferential statistics are included in the results presented below; however, because the numbers of students are quite large, differences in average scores may be statistically significant.

10.1.1 Difficulty of Assessment

Examinees in all three grades were asked how difficult the science test was. Figures 10-1 through 10-3 show that students in grades 8 and 11 who thought the test was easier than their regular science schoolwork did better overall than those who thought it was more difficult; the trend was especially dramatic for grade 11. The pattern of responses for grade 4 was inconsistent.

Question: How difficult was this science test?

Figure 10-1. 2014–15 NECAP Science: Questionnaire Responses Difficulty—Grade 4

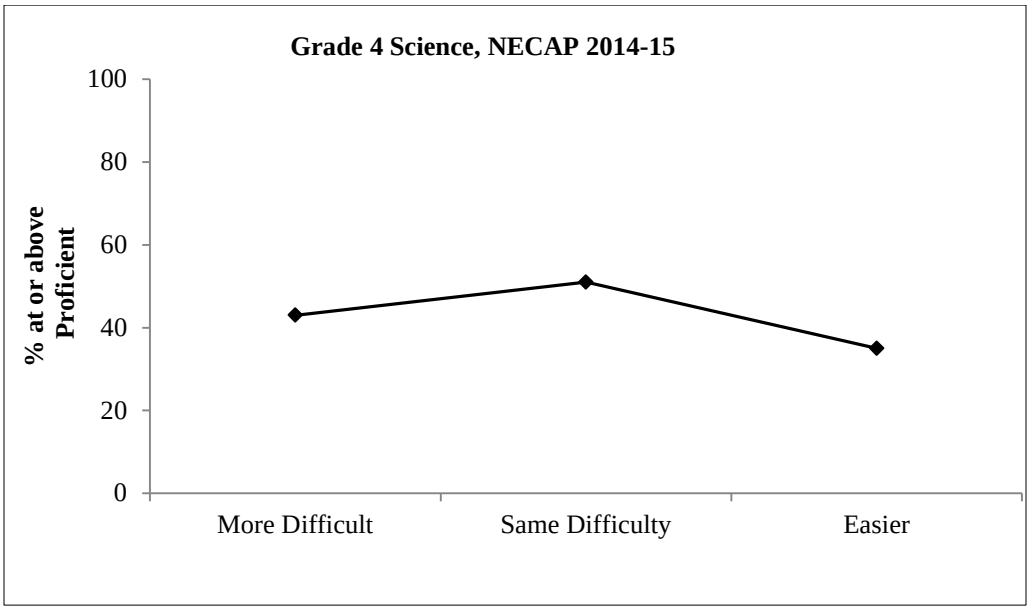


Figure 10-2. 2014–15 NECAP Science: Questionnaire Responses Difficulty—Grade 8

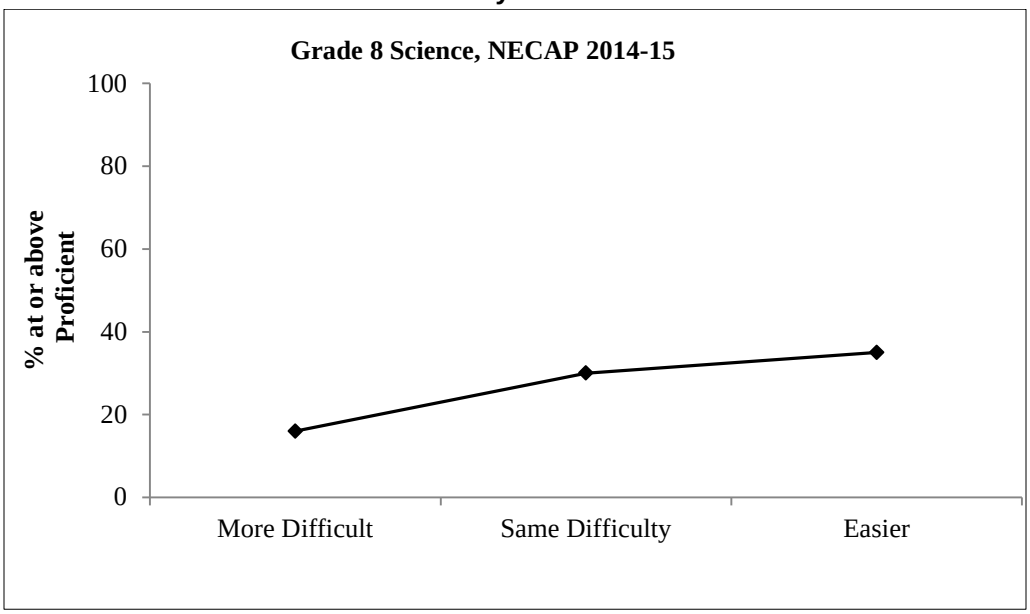
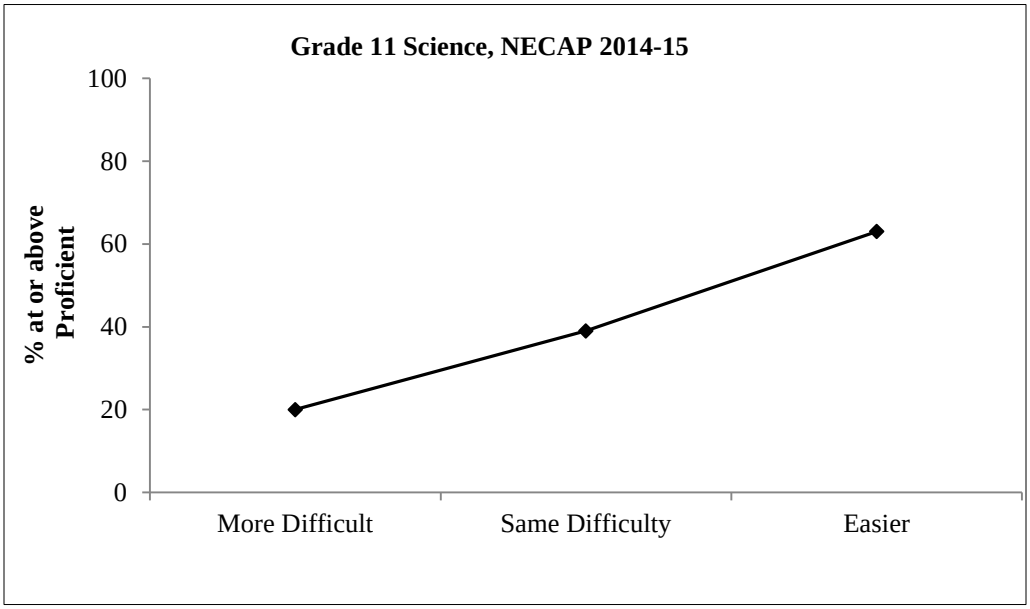


Figure 10-3. 2014–15 NECAP Science: Questionnaire Responses Difficulty—High School



10.1.2 Content

Examinees in all three grades were asked how often they do science experiments or inquiry tasks in their class. Figures 10-4 through 10-6 indicate a slight positive relationship between frequency of performing experiments or inquiry tasks and NECAP Science scores (i.e., higher scores are associated with greater frequency) for grade 8 and a relatively flat relationship for students who perform an inquiry task at least yearly in grades 4 and 11. In all cases, by excluding the "Weekly" response, the relationship is positive although the differences are slight, especially in grade 11.

Question: How often do you do science experiments or inquiry tasks in your class?

Figure 10-4. 2014–15 NECAP Science: Questionnaire Responses Content—Grade 4

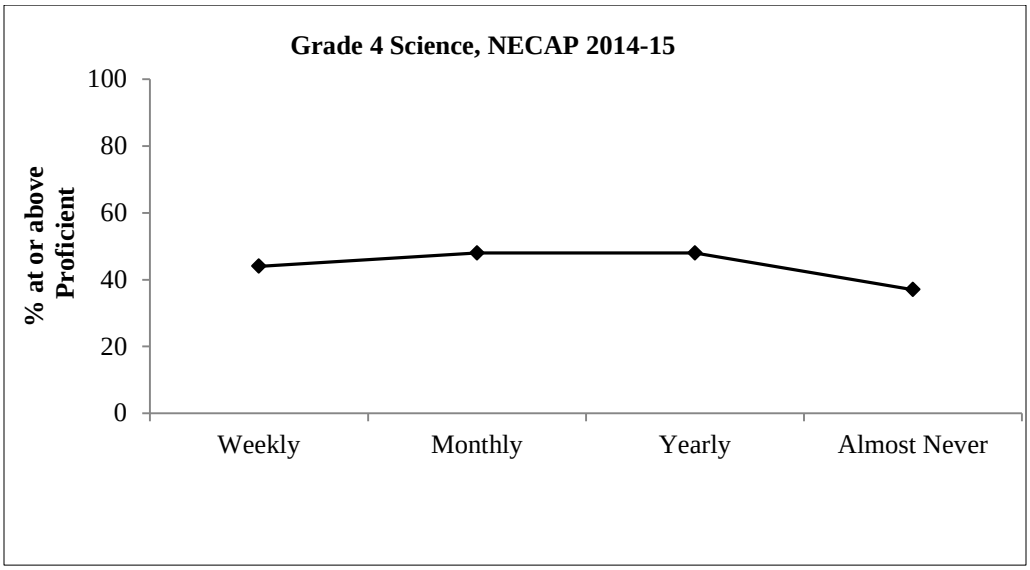


Figure 10-5. 2014–15 NECAP Science: Questionnaire Responses Content—Grade 8

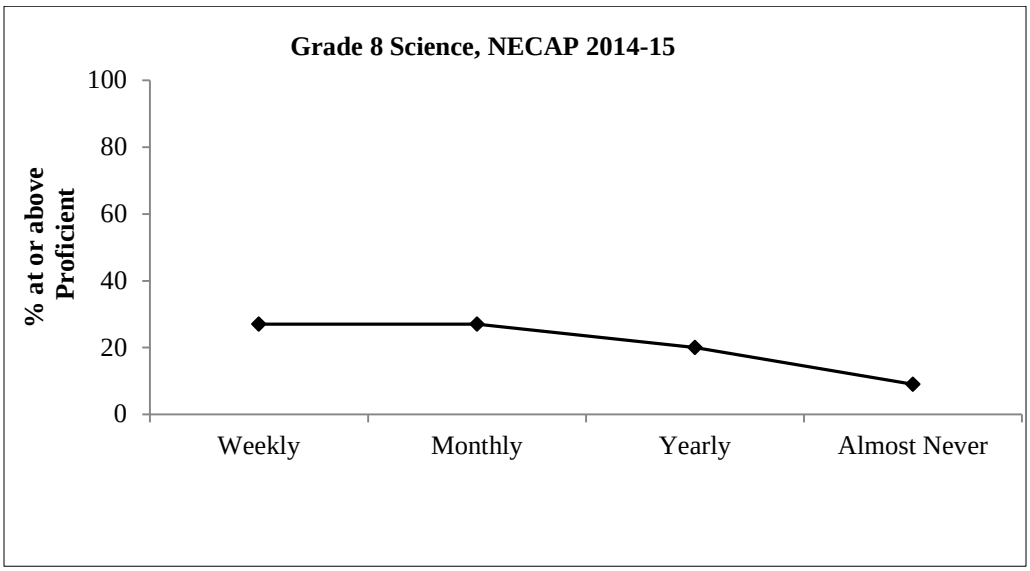
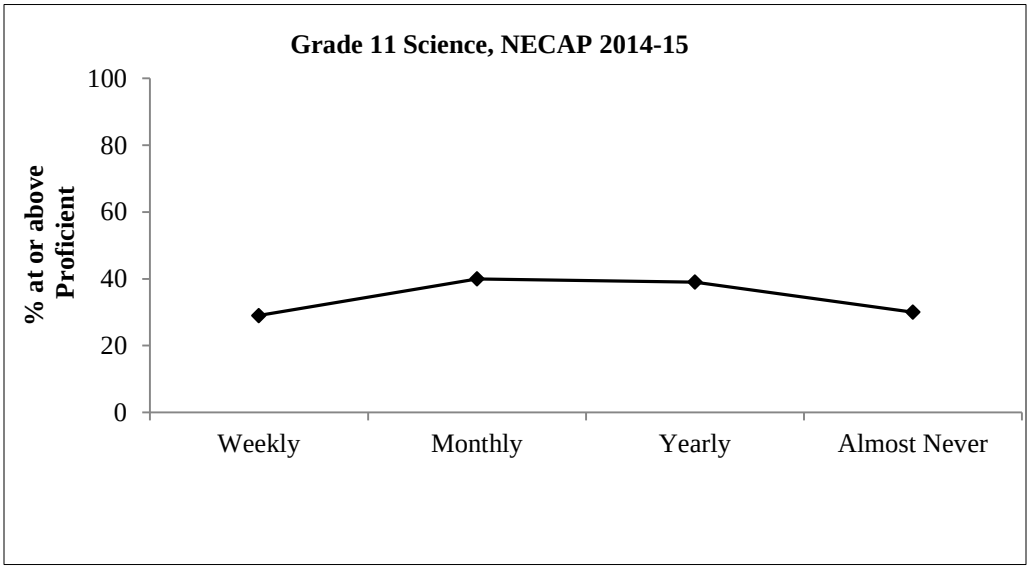


Figure 10-6. 2014–15 NECAP Science: Questionnaire Responses Content—High School



10.1.3 Homework

Examinees in all three grades were asked how often they have science homework. Figures 10-7 through 10-9 indicate a strong positive relationship between frequency of homework and NECAP Science scores for grades 11. The results for grade 4, on the other hand, indicate a negative relationship between frequency of homework and NECAP Science scores. The results for Grade 8 show a fairly flat relationship between frequency of homework and NECAP Science scores.

Question: How often do you have science homework?

Figure 10-7. 2014–15 NECAP Science: Questionnaire Responses Homework—Grade 4

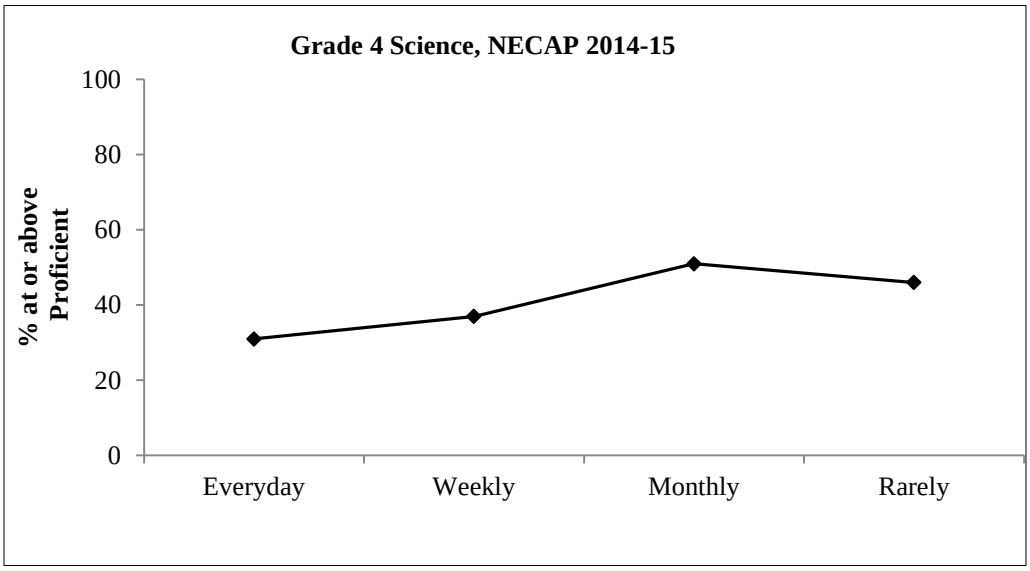


Figure 10-8. 2014–15 NECAP Science: Questionnaire Responses Homework—Grade 8

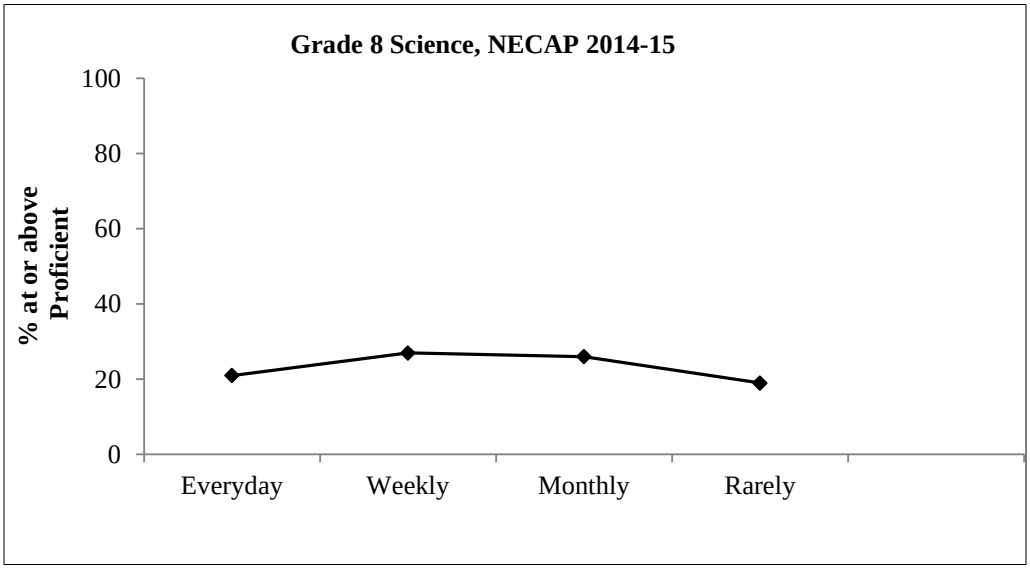
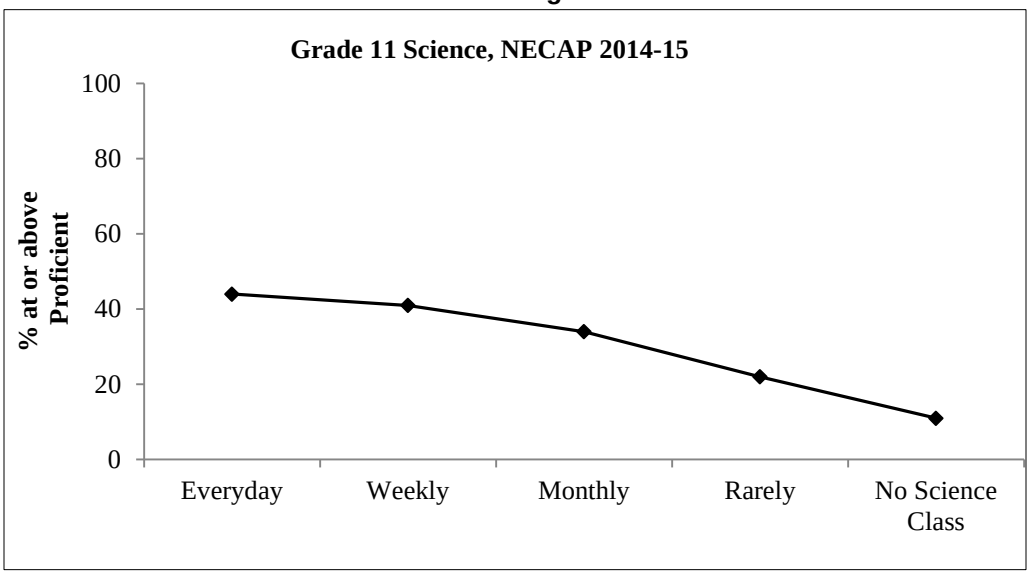


Figure 10-9. 2014–15 NECAP Science: Questionnaire Responses Homework—High School



10.1.4 Performance in Science Class

Students in grades 8 and 11 were asked what their most recent science grade was. Figures 10-10 and 10-11 indicate that, for grades 8 and 11, there was a strong positive relationship between the most recent science grade and NECAP Science scores.

Question: What was your science grade on your most recent report card?

Figure 10-10. 2014–15 NECAP Science: Questionnaire Responses Grade in Science—Grade 8

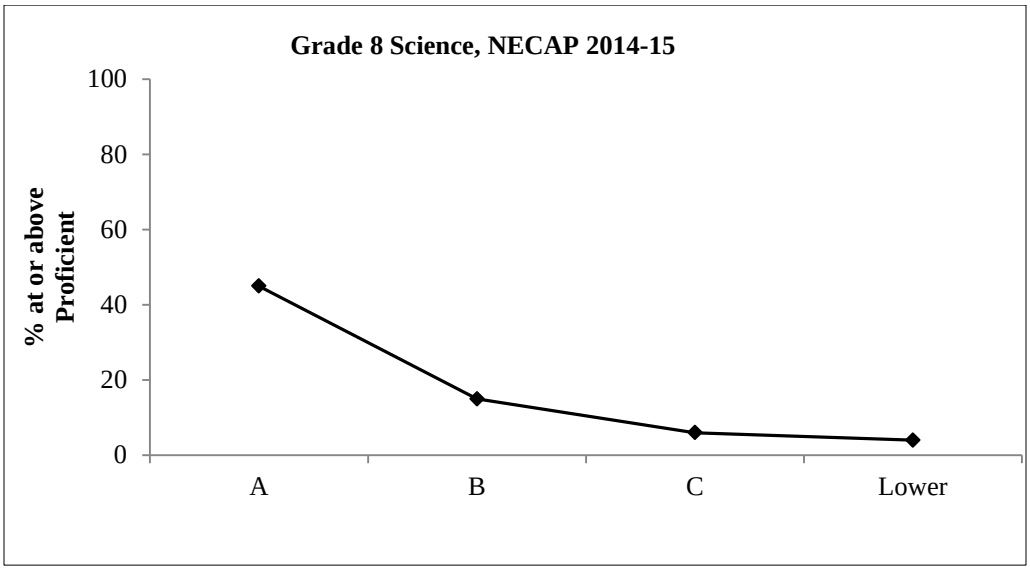
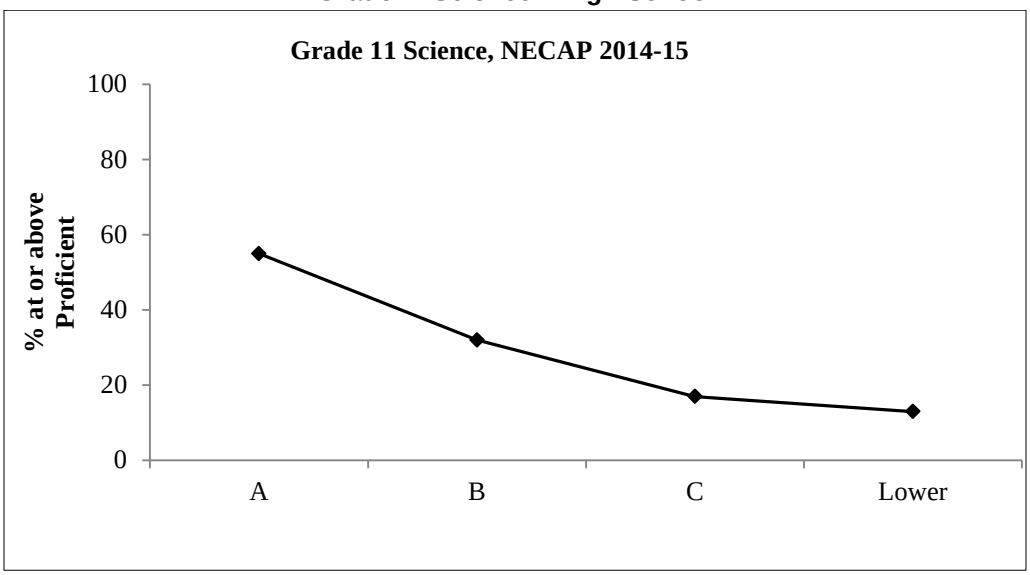


Figure 10-11. 2014–15 NECAP Science: Questionnaire Responses Grade in Science—High School



The evidence presented in this report supports inferences made about student achievement on the content represented in the NECAP Science assessment targets. As such, the evidence provided also supports the use of NECAP Science results 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, DC: American Educational Research Association.

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

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

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: John Wiley & Sons.

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

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

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

Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education*. Washington, DC: 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 mental test scores.* Reading, MA: Addison-Wesley.

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

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

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). Washington, DC: American Council on Education/Macmillan.

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.

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: 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—GUIDELINES FOR THE DEVELOPMENT OF SCIENCE INQUIRY TASKS

NECAP
Science Assessment

Guidelines for the Development of Science Inquiry Tasks

Created in Partnership by the New Hampshire, Rhode Island, and Vermont Departments of Education

February 2008

Introduction: Inquiry in the NECAP Science Assessment

Defining the NECAP Science Assessment Inquiry Task

Focus – The Science Inquiry Task at every grade level should be rich and engaging. The task may be an experimental question or observational question – it is the quality of the task that is most important. Regardless of the type of task, all Four Broad Areas of Inquiry as defined in the *NECAP Schema for Assessing Scientific Inquiry*, (see column headings in the table on page 4), will be assessed. The task should flow from beginning to end in a purposeful way that allows students to make connections, express their ideas, and provide evidence of scientific thinking.

Design – Inquiry Tasks should be rooted in one or more NECAP Science Assessment Targets (one of which should have INQ code) and over time should address a variety of content domains. For every task at grades four and eight there must be scoreable components from each of the Four Broad Areas of Inquiry. At grade 11, while the focus of the task may be on constructs in the Area of Developing and Evaluating Explanations (column 4), scoreable items from each of the other three Broad Areas of Inquiry should also be included.

Task development will be guided by *Guidelines for the Development of Science Inquiry Tasks* (*GDIT*). For each item within a Science Inquiry Task, the developer must identify the Depth of Knowledge (DOK), the Inquiry Construct number, score points, and key elements (scoring notes). Over time, all Inquiry Constructs should be addressed at each grade level. See the Appendix for additional information about the Inquiry Task development process.

Goal – Science Inquiry Tasks will engage students in a range of Depth of Knowledge experiences up to and including strategic thinking (DOK 3). Individual tasks may look different, but each should focus on providing insight into how students engage in scientific thinking. The goal is to encourage the meaningful inclusion of inquiry in classrooms at all levels.

Applying the Guidelines of the Science NECAP Assessment Task in the Classroom

Background – The first version of *Guidelines for Development of Science Inquiry Tasks* was originally created by the Science Specialists from the New Hampshire, Rhode Island, and Vermont Departments of Education to facilitate and refine the development of Inquiry Tasks for the NECAP Science Assessment. It became clear that such a tool would be useful to teachers and local science specialists to guide them in the development of similar tasks for classroom use at all levels. The State Science Specialists have collaborated on this version of *GDIT* to help educators understand and employ the constructs of the Four Broad Areas of Inquiry as they design or evaluate inquiry tasks for classroom instruction and assessment.

Focus - Classroom inquiry tasks should be relevant, engaging and meaningful learning experiences for students. The classroom inquiry tasks included on the state Department of Education website are examples of the kinds of tasks found in the NECAP Science Assessment. In the classroom any inquiry activity should provide regular opportunities for students to experience the science process as defined in the *NECAP Schema for Assessing Scientific Inquiry* (see page 4). Analysis of student performance on classroom inquiry tasks can inform instruction by providing data on student proficiencies within the constructs across the Four Broad Areas of Inquiry. Classroom inquiry tasks might be used as a component of local assessment or as a classroom summative assessment for a specific unit.

Design - While there are many ways to design inquiry experiences and an assessment for the classroom, *GDIT* provides a framework for the development of rich performance assessments that are aligned with this component of the NECAP Science Assessment. *GDIT* offers the necessary details for teachers to develop classroom inquiry tasks that are similar in structure to the NECAP Science Inquiry Tasks. Each classroom inquiry task will include elements from each of the Four Broad Areas of Inquiry, and address specific constructs within each Broad Area. Classroom inquiry tasks can span a class period, a few days or the length of a unit. Classroom inquiry tasks related to units of study provide opportunities for students to become familiar with the format of the NECAP Science Inquiry Tasks and will help to prepare them for the state assessment

Goals - The main goals of *Guidelines for Development of Science Inquiry Tasks* are to help educators:

- encourage the inclusion of engaging and relevant inquiry experiences in classrooms that contribute to increasing the science literacy of the citizens of New Hampshire, Rhode Island and Vermont;

- develop, evaluate and implement rich science tasks that allow students to gain skills across the Four Broad Areas of Inquiry;

- understand the process and parameters used in the development of Inquiry Tasks for the NECAP Science Assessment;

- provide opportunities for students to become familiar with the format and requirements of the NECAP Science Inquiry Tasks.

NECAP Science Inquiry Constructs for all Grade Levels

NECAP Science Schema for Assessing Scientific Inquiry (with DOK levels for constructs)				
Broad Areas of Inquiry to be Assessed	**Formulating Questions & Hypothesizing**	**Planning and Critiquing of Investigations**	**Conducting Investigations**	**Developing and Evaluating Explanations**
Constructs for each Broad Area of Inquiry (including intended DOK Ceiling Levels, based on Webb Depth of Knowledge Levels for Science – see also Section II) ***Inquiry Constructs answer the question: What is it about the broad area of Inquiry that we want students to know and be able to do?***	**1.** Analyze information from observations, research, or experimental data for the purpose of formulating a question, hypothesis, or prediction: **(DOK 3)** **1a.** Appropriate for answering with scientific investigation **1b.** For answering using scientific knowledge **2.** Construct coherent argument in support of a question, hypothesis, prediction **(DOK 2 or 3 depending on complexity of argument)** **3.** Make and describe observations in order to ask questions, hypothesize, make predictions related to topic **(DOK 2)**	**4.** Identify information/evidence that needs to be collected in order to answer the question, hypothesis, prediction **(DOK 2 – routine; DOK 3 non-routine/ more than one dependant variable)** **5.** Develop an organized and logical approach to investigating the question, including controlling variables **(DOK 2 – routine; DOK 3 non-routine)** **6.** Provide reasoning for appropriateness of materials, tools, procedures, and scale used in the investigation **(DOK 2)**	**7.** Follow procedures for collecting and recording qualitative or quantitative data, using equipment or measurement devices accurately **(DOK 1 – use tools; routine procedure;** DOK 2 – follow multi-step procedures; make observations) **8.** Use accepted methods for organizing, representing, and manipulating data **(DOK 2 – compare data; display data)** **9.** Collect sufficient data to study question, hypothesis, or relationships **(DOK 2 – part of following procedures)** **10.** Summarize results based on data **(DOK 2)**	**11.** Analyze data, including determining if data are relevant, artifact, irrelevant, or anomalous **(DOK 2 – specify relationships between facts; ordering, classifying data)** **12.** Use evidence to support and justify interpretations and conclusions or explain how the evidence refutes the hypothesis **(DOK 3)** **13.** Communicate how scientific knowledge applies to explain results, propose further investigations, or construct and analyze alternative explanations **(DOK 3)**

NECAP Science Assessment Inquiry Task Flow

Administration of each Science Inquiry Performance Task (Grades 4 and 8) should follow the sequence below:

Prior to start of Session 3:

- Set up materials
- Group students

Standard Flow of NECAP Science Inquiry Performance Tasks: (Grades 4 and 8)

1. Directions read aloud by Test Administrator (basic info)
2. Scenario read aloud by Test Administrator (context)
3. Description of the materials and/or model explained by Test Administrator. Students make a prediction individually
4. Students conduct investigation with partner
5. Students clean up kits/experiment with partner
6. Students return to desks with their own Task Booklet to work individually
7. Test Administrator distributes Student Answer Booklets to students
8. Students copy data from Task Booklet to Student Answer Booklet (non-scored)
9. Students answer eight (8) scored questions in Student Answer Booklet

   A. For analyzing the prediction, there will be Yes/No check boxes with space for the narrative below.

   B. At grades 4 and 8, the question where students must graph data will have a hard-coded grid (1/2- inch squares) in the answer box with lines for *x* and *y* axis labels as well as a title. At grade 11, use 1/4- inch squares.

   .

Standard Flow of NECAP Science Inquiry Data Analysis Tasks: (Grades 8 and 11)

1. Test Administrator distributes Student Answer Booklets to students
2. Directions read aloud by Test Administrator (basic info)
3. Scenario read aloud by Test Administrator (task context)
4. Students answer questions related to the scenario and complete data analysis in the Student Answer Booklet.
5. Items will require high school students to consider the Inquiry Constructs in relation to a selected data set.
6. Upon completion of the task students sit quietly and read until dismissal.

Broad Area 1: Formulating Questions and Hypothesizing

Grade 4

Standard: *Task must provide students a scenario that describes objects, organisms, or events within the environment. The scenario must include information relevant to grade 4 students and sufficient for them to construct questions and/or predictions based upon observations, past experiences, and scientific knowledge.*

Note: *bullets addressing constructs are not all inclusive.*

Inquiry Construct:	Items addressing this construct require students to:
1. Analyze information from observations, research, or experimental data for the purpose of formulating a question, hypothesis, or prediction: **1a.** Appropriate for answering with scientific investigation **1b.** For answering using scientific knowledge **DOK 3**	• analyze scientific data and use that information to generate a testable question or a prediction that includes a cause and effect relationship; • generate a question or prediction which is reasonable in terms of available evidence; • support a question or prediction with an explanation. **Note:** Addressing this construct may appear at the beginning of the task, the end, or both.
2. Construct coherent argument in support of a question, hypothesis, prediction **DOK 2 or 3** depending on complexity of argument	• identify evidence that supports or does not support a question or prediction.
3. Make and describe observations in order to ask questions, hypothesize, make predictions related to topic **DOK 2**	• connect observations to a question or prediction. **Note:** Items may refer to an existing, new, or student-generated question or prediction.

Broad Area 2: Planning and Critiquing of Investigations

Grade 4

Standard: *Task requires students to plan or analyze a simple experiment based upon questions or predictions derived from the scenario. The experiment and related items should emphasize fairness in its design.*

Note: The words "procedure" and "plan" are synonymous.

Inquiry Construct:	Items addressing this construct require students to:
4. Identify information and/or evidence that needs to be collected in order to answer the question, hypothesis, prediction **DOK 2** (routine) **DOK 3** (non-routine or more than one dependant variable)	• identify the types of evidence that should be gathered to answer the question; • design an appropriate format, such as data tables or charts, for recording data. **Note:** These items could appear at the end of the task.
5. Develop an organized and logical approach to investigating the question, including controlling variables **DOK 2** (routine) **DOK 3** (non-routine)	• develop a procedure to gather sufficient evidence (including multiple trials) to answer the question or test the prediction; • develop a procedure that lists steps logically and sequentially; • develop a procedure that changes one variable at a time. **Note:** These items could appear at the end of the task. Use of the term "variable" should not appear in the item stem.
6. Provide reasoning for appropriateness of materials, tools, procedures, and scale used in the investigation **DOK 2**	• explain why the materials, tools, or procedure for the task are or are not appropriate for the investigation.

Broad Area 3: Conducting Investigations

Grade 4

Standard: *The procedure requires the student to demonstrate simple skills (observing, measuring, basic skills involving fine motor movement). The investigation requires the student to use simple scientific equipment (rulers, scales, thermometers) to extend their senses. The procedure provides the student with an opportunity to collect sufficient data to investigate the question, prediction, or relationships. Student is required to organize and represent qualitative or quantitative data using blank graph/chart templates. Student is required to summarize data.*

Note: Metric measurements are used for Grade 4, except for those pertaining to weather.

Note: Multiple trials mean repeating the experiment to collect multiple sets of data.

Inquiry Construct	Items addressing this construct require students to:
7. Follow procedures for collecting and recording qualitative or quantitative data, using equipment or measurement devices accurately **DOK 1:** use tools; routine procedure; **DOK 2:** follow multi-step procedures; make observations	• record precise data and observations that are consistent with the procedure of the investigation; • include appropriate units of all measurements; • use appropriate measurement tools correctly to collect data; • record and label relevant details within a scientific drawing or diagram.
8. Use accepted methods for organizing, representing, and manipulating data **DOK 2:** compare data; display data	• represent data accurately in a graph/table/chart; • include titles , labels, keys or symbols as needed; • select a scale appropriate for the range of data to be plotted; • use common terminology to label representations; • identify relationships among variables based upon evidence.
9. Collect sufficient data to study question, hypothesis, or relationships **DOK 2** part of following procedures	• show understanding of the value of multiple trials; • relate data to original question and prediction; • determine if the quantity of data is sufficient to answer the question or support or refute the prediction.
10. Summarize results based on data **DOK 2**	• consider all data when developing an explanation and/or conclusion; • identify patterns and trends in data.

Broad Area 4: Developing and Evaluating Explanations

Grade 4

Standard: *Task must provide the opportunity for students to use data to construct an explanation based on their science knowledge and evidence from experimentation or investigation.*

Inquiry Construct	Items addressing this construct require students to:
11. Analyze data, including determining if data are relevant, artifact, irrelevant, or anomalous **DOK 2 -** specify relationships between facts; ordering, classifying data	• identify data relevant to the task or question ; • identify factors that may affect experimental results (e.g. variables, experimental error, environmental conditions); • classify data into meaningful categories.
12. Use evidence to support and justify interpretations and conclusions or explain how the evidence refutes the hypothesis **DOK 3**	• identify data that seem inconsistent ; • use evidence to support or refute a prediction; • use evidence to justify an interpretation of data or trends; • identify and explain differences or similarities between prediction and experimental data; • provide a reasonable explanation that accurately reflects data; • use mathematical reasoning to determine or support conclusions.
13. Communicate how scientific knowledge applies to explain results, propose further investigations, or construct and analyze alternative explanations **DOK 3**	• explain how experimental results compare to accepted scientific understanding; • suggest ways to modify the procedure in order to collect sufficient data; • identify additional data that would strengthen an investigation; • connect the investigation or model to a real world example; • propose new questions, predictions, next steps or technology for further investigations; • design an investigation to further test a prediction.

Broad Area 1: Formulating Questions and Hypothesizing

Grade 8

Standard: *Task must provide students a scenario that describes objects, organisms, or events to which the student will respond. The task will provide the student with the opportunity to develop their own testable questions or predictions based upon their experimental data, observations, and scientific knowledge. The task could include opportunities for the student to refine and refocus questions or hypotheses related to the scenario using their scientific knowledge and information*

Inquiry Construct	Items addressing this construct require students to:
1. Analyze information from observations, research, or experimental data for the purpose of formulating a question, hypothesis, or prediction: **(DOK 3)** **1a.** Appropriate for answering with scientific investigation **1b.** For answering using scientific knowledge	• analyze scientific data and use that information to generate a testable question or a prediction that includes a cause and effect relationship; • generate a question or a prediction which is reasonable in terms of available evidence; • support their question or prediction with a <u>scientific</u> explanation; • <u>refine or refocus a question or hypothesis using experimental data, research, or scientific knowledge.</u> **Note:** Addressing this construct may appear at the beginning of the task, the end, or both.
2. Construct coherent argument in support of a question, hypothesis, prediction **DOK 2 or 3** depending on complexity of argument	• identify evidence that supports or does not support a question, <u>hypothesis</u> or prediction; • <u>explain the cause and effect relationship within the hypothesis or prediction;</u> • <u>use a logical argument to explain how the hypothesis or prediction is connected to a scientific concept, or observation.</u>
3. Make and describe observations in order to ask questions, hypothesize, make predictions related to topic **DOK 2**	• connect observations to a question or prediction. **Note:** Items may refer to an existing, new, or student-generated question or prediction.

Broad Area 2: Planning and Critiquing of Investigations

Grade 8

Standard: *The task will require students to plan or analyze an experiment or investigation based upon questions, hypothesis, or predictions derived from the scenario. An experiment must provide students with the opportunity to identify and control variables. The task will provide opportunities for students to think critically about experiments and investigations and may ask students to propose alternatives.* **Note:** Scale refers to proportionality between the model and what it represents or the frequency with which data are collected.	
Inquiry Construct	***Items addressing this construct require students to:***
4. Identify information/evidence that needs to be collected in order to answer the question, hypothesis, prediction **DOK 2:** routine; **DOK 3:** non-routine/ more than one dependant variable	• identify the types of evidence that should be gathered to answer the question, or support or refute the prediction ; • identify the variables that may affect the outcome of the experiment or investigation; • design an appropriate format for recording data; • evaluate multiple data sets to determine which data are relevant to the question, hypothesis or prediction. **Note:** These items could appear at the end of the task
5. Develop an organized and logical approach to investigating the question, including controlling variables **DOK 2**: routine (replicates existing procedure); **DOK 3:** non-routine (extends, refines, or improves existing procedure)	• develop a procedure to gather sufficient evidence (including multiple trials) to answer the question, or test the hypothesis, or prediction; • develop a procedure that lists steps sequentially and logically; • explain which variable will be manipulated or changed (independent) and which variable will be affected by those changes (dependent); • identify variables that will be kept constant throughout the investigation; • use scientific terminology that supports the identified procedures; • evaluate the organization and logical approach of a given procedure including variables, controls, materials, and tools; • evaluate investigation design, including opportunities to collect appropriate and sufficient data. **Note:** These items could appear at the beginning or the end of the task.
6. Provide reasoning for appropriateness of materials, tools, procedures, and scale used in the investigation **DOK 2**	• explain why the materials, tools, procedure, or scale for a task are appropriate or are inappropriate for the investigation. • evaluate the investigation for the safe and ethical considerations of the materials, tools, and procedures.

Broad Area 3: Conducting Investigations

Grade 8

Standard: *The procedure requires the student to demonstrate skills (observing, measuring, basic skills involving fine motor movement) and mathematical understanding. The materials involved in the investigation are authentic to the task required. The procedure provides the student with an opportunity to collect sufficient data to investigate the question, prediction/hypothesis, or relationships. Student is required to organize and represent qualitative or quantitative data. Student is required to summarize data to form a logical argument.* **Note:** Metric units are used for all Grade 8 measurements. **Note:** Multiple trials means repeating the experiment to collect multiple sets of data.	
Inquiry Construct	**Items addressing this construct require students to:**
7. Follow procedures for collecting and recording qualitative or quantitative data, using equipment or measurement devices accurately **DOK 1:** use tools; routine procedure; **DOK 2:** follow multi-step procedures; make observations	• record precise data and observations that are consistent with the procedure of the investigation; • include appropriate units of all measurements; • use appropriate measurement tools correctly to collect data; • record and label relevant details within a scientific drawing.
8. Use accepted methods for organizing, representing, and manipulating data **DOK 2:** compare data; display data	• represent data accurately in an appropriate graph/table/chart; • include titles, labels, keys or symbols as needed; • select a scale appropriate for the range of data to be plotted; • use scientific terminology to label representations; • identify relationships among variables based upon evidence. **Note:** The standard practice of graphing in science is to represent the independent on the *x*-axis and the dependent variable on the *y*- axis.
9. Collect sufficient data to study question, hypothesis, or relationships **DOK 2**: part of following procedures	• show understanding of the value of multiple trials; • relate data to original question, hypothesis or prediction; • determine if the quantity of data is sufficient to answer the question or support or refute the hypothesis or prediction.
10. Summarize results based on data **DOK 2**	• consider all data when developing an explanation/conclusion; • identify patterns and trends in data.

Broad Area 4: Developing and Evaluating Explanations

Grade 8

Standard *Task must provide the opportunity for students to use data to construct an explanation based on their science knowledge and evidence from experimentation or investigation. The task requires students to use qualitative and quantitative data to communicate conclusions and support/refute prediction/hypothesis.*	
Inquiry Construct	**Items addressing this construct require students to:**
11. Analyze data, including determining if data are relevant, artifact, irrelevant, or anomalous **DOK 2:** specify relationships between facts; ordering, classifying data	• identify data relevant to the task or question; • identify factors that may affect experimental results (e.g. variables, experimental error, environmental conditions); • classify data into meaningful categories; • compare experimental data to accepted scientific data provided as part of the task; • use mathematical and statistical techniques to analyze data; • provide a reasonable explanation that accurately reflects data; • use content understanding to question data that might seem inaccurate; • evaluate the significance of experimental data.
12. Use evidence to support and justify interpretations and conclusions or explain how the evidence refutes the hypothesis **DOK 3**	• identify and explain data, interpretations or conclusions that seem inaccurate; • use evidence to support or refute question or hypothesis; • use evidence to justify an interpretation of data or trends; • identify and explain differences or similarities between predictions and experimental data; • provide a reasonable explanation that accurately reflects data; • use mathematical computations to determine or support conclusions.
13. Communicate how scientific knowledge applies to explain results, propose further investigations, or construct and analyze alternative explanations **DOK 3**	• explain how experimental results compare to accepted scientific understanding; • recommend changes to procedures to produce data that would provide sufficient data and more accurate analysis; • identify and justify additional data that would strengthen an investigation; • connect the investigation or model to an authentic situation; • propose and evaluate new questions, predictions, next steps or technology for further investigations or alternative explanations; • account for limitations and/or sources of error within the experimental design; • apply experimental results to a new problem or situation.

Broad Area 1: Formulating Questions and Hypothesizing

Grade 11

***Standard**: Task must provide students a scenario with information and detail sufficient for the student to create a testable prediction or hypothesis. Students will draw upon their science knowledge base to advance a prediction or hypothesis using appropriate procedures and controls; this may include an experimental design.*

Inquiry Construct	Items addressing this construct require students to:
1. Analyze information from observations, research, or experimental data for the purpose of formulating a question, hypothesis, or prediction. **1a.** Appropriate for answering with scientific investigation **1b.** For answering using scientific knowledge **DOK 3**	• analyze scientific data and use that information to generate a testable question, hypothesis, or prediction that includes a cause and effect relationship; • generate a question, hypothesis or a prediction which is reasonable in terms of available evidence; • show connections between hypothesis or prediction and scientific knowledge, observations, or research; • support their question, hypothesis, or prediction with a scientific explanation; • refine or refocus a question or hypothesis using experimental data, research, or scientific knowledge. **Note:** Addressing this construct may appear at the beginning of the task, the end, or both.
2. Construct coherent argument in support of a question, hypothesis, prediction. **DOK 2 or 3:** depends on complexity of argument	• identify evidence that supports or does not support a question, hypothesis or prediction • explain the cause and effect relationship within the hypothesis or prediction; • use a logical argument to support the hypothesis or prediction using scientific concepts, principles, or observations.
3. Make and describe observations in order to ask questions, hypothesize, make predictions related to topic. **DOK 2**	• connect observations and data to a question, hypothesis, or prediction. **Note:** Items may refer to an existing, new, or student-generated question, hypothesis, or prediction.

Broad Area 2: Planning and Critiquing of Investigations

Grade 11

Standard: *The task will require students to plan or analyze an experiment or investigation based upon questions, hypothesis, or predictions derived from the scenario. An experiment must provide students with the opportunity to identify and control variables. The task will provide opportunities for students to think critically and construct an argument about experiments and investigations and may ask students to propose alternatives. Task will require the student to identify and justify the appropriate use of tools, equipment, materials, and procedures involved in the experiment.* **Note:** Scale refers to proportionality between the model and what it represents or the frequency with which data are collected.	
Inquiry Construct	**Items addressing this construct require students to:**
4. Identify information/evidence that needs to be collected in order to answer the question, hypothesis, prediction **DOK 2:** routine; **DOK 3:** non-routine; more than one dependent variable	• identify the types of evidence that should be gathered to answer the question, or support or refute the hypothesis or prediction; • identify the variables that may affect the outcome of the experiment or investigation; • design an appropriate format for recording data and include relevant technology; • evaluate multiple data sets to determine which data are relevant to the question, hypothesis or prediction. **Note:** These items could appear at the end of the task.
5. Develop an organized and logical approach to investigating the question, including controlling variables **DOK 2**: routine (replicates existing procedure); **DOK 3:** non-routine (extends, refines, or improves existing procedure)	• develop a procedure to gather sufficient evidence (including multiple trials) to answer the question, or test the hypothesis, or prediction; • develop a procedure that lists steps sequentially and logically and incorporates the use of appropriate technology; • explain which variable will be manipulated or changed (independent) and which variable will be affected by those changes (dependent); • identify variables that will be kept constant throughout the investigation; • distinguish between the control group and the experimental group in an investigation; • use scientific terminology that supports the identified procedures; • evaluate the organization and logical approach of a given procedure including variables, controls, materials, and tools. • evaluate investigation design, including opportunities to collect appropriate and sufficient data. **Note:** These items could appear at the beginning or the end of the task.

6. Provide reasoning for appropriateness of materials, tools, procedures, and scale used in the investigation **DOK 2**	• explain why the materials, tools, procedure, or scale for a task are appropriate or inappropriate for the investigation. • evaluate the investigation for the safe and ethical considerations of the materials, tools, and procedures.

Broad Area 3: Conducting Investigations

Grade 11

Standard: *The procedure requires the student to collect data through observation, inference, and prior scientific knowledge. Mathematics is required for the student to determine and report data. The task scenario is authentic to the realm of the student. The task requires the student to collect sufficient data to investigate the question, prediction/hypothesis, or relationships. Student is required to organize and represent qualitative or quantitative data. Student is required to summarize data to form a logical argument.* **Note:** Metric units are used for all Grade 11 measurements **Note:** Multiple trials mean repeating the experiment to collect multiple sets of data.	
Inquiry Construct	**Items addressing this construct require students to:**
7. Follow procedures for collecting and recording qualitative or quantitative data, using equipment or measurement devices accurately **DOK 1:** use tools; routine procedure; **DOK 2**: follow multi-step procedures; make observations	• record precise data and observations that are consistent with the procedure of the investigation; • include appropriate units of all measurements; • use appropriate measurement tools correctly to collect data; record and label relevant details within a scientific drawing.
8. Use accepted methods for organizing, representing, and manipulating data **DOK 2** : compare data; display data	• represent data accurately in an appropriate graph/table/chart; • include titles, labels, keys or symbols as needed; • select a scale appropriate for the range of data to be plotted; • use scientific terminology to label representations; • identify relationships among variables based upon evidence. **Note:** The standard practice of graphing in science is to represent the independent on the *x*-axis and the dependent variable on the *y*- axis.
9. Collect sufficient data to study question, hypothesis, or relationships **DOK 2** : part of following procedures	• show understanding of the value of multiple trials • relate data to original question, hypothesis or prediction; • determine if the quantity of data is sufficient to answer the question or support or refute the hypothesis or prediction.
10. Summarize results based on data **DOK 2**	• consider all data when developing an explanation/conclusion; • identify patterns and trends in data.

Broad Area 4: Developing and Evaluating Explanations

Grade 11

Standard: *Task must provide the opportunity for students to use data to construct an explanation based on their science knowledge and evidence from experiment or investigation. The task requires students to use qualitative and quantitative data to communicate conclusions and support/refute prediction/hypothesis. The task provides students the opportunity to recognize and analyze alternative methods and models to evaluate other plausible explanations.* ***Note:*** *The complexity of the scenario and associated data sets distinguishes this task from an 8th Grade task.*	
Inquiry Construct	**Items addressing this construct require students to:**
11. Analyze data, including determining if data are relevant, artifact, irrelevant, or anomalous **DOK 2**: specify relationships between facts; ordering, classifying data	• identify data relevant to the task or question; • identify factors that may affect experimental results (e.g. variables, experimental error, environmental conditions); • analyze data and sort into meaningful categories; • compare experimental data to accepted scientific data provided as part of the task; • use mathematical and statistical techniques to analyze data; • provide a reasonable explanation that accurately reflects data; • use content understanding to question data that might seem inaccurate • evaluate the significance of experimental data.
12. Use evidence to support and justify interpretations and conclusions or explain how the evidence refutes the hypothesis **DOK 3**	• identify and explain data, interpretations or conclusions that seem inaccurate; • use evidence to support or refute question or hypothesis; • use evidence to justify an interpretation of data or trend; • identify and explain differences or similarities between hypothesis and predictions and experimental data; • use evidence to justify a conclusion or explanation based on experimental data; • use mathematical computations to determine or support conclusions; • evaluate potential bias in the interpretation of evidence.

13. Communicate how scientific knowledge applies to explain results, propose further investigations, or construct and analyze alternative explanations **DOK 3**	• explain how experimental results compare to accepted scientific understanding; • recommend changes to procedures to produce data that would provide sufficient data and more accurate analysis; • identify and justify additional data that would strengthen an investigation; • connect the investigation or model to an authentic situation; • propose and evaluate new questions, predictions, next steps or technology for further investigations or alternative explanations; • account for limitations and/or sources of error within the experimental design; • apply experimental results to a new problem or situation; • consider the impact (safety, ethical, social, civic, economic, environmental) of additional investigations.

APPENDIX

NECAP Science
Inquiry Task Development Process

Initial Steps for the Development of an Inquiry Task

1. Identify the NECAP Assessment **TARGET** to be addressed within the major idea for the task.

2. Refer to the ***Guidelines for the Development of Science Inquiry Tasks*** (***GDIT***). Brainstorm constructs that would be addressed under each broad area within the major idea for the task.

Formulating Questior and Hypothesizing	**Planning and Critiquing of Investigations**	**Conducting Investigations**	**Developing and Evaluating Explanation**

3. Develop a draft **SCENARIO** aligned to the major idea of the task that could generate testable questions.*

4. Identify an authentic **Data Set** (Grades 8 & 11) that applies to the **TARGET** and relates to the **SCENARIO** *

OR

Provide opportunity for **Collection of Data** (Grade 4 & 8) that applies to the **TARGET** and relates to the **SCENARIO** *

* **Note**: *The previous steps are interdependent. The construction of the draft SCENARIO and the identification of a data set, will inform one another. Either may necessitate modifications for alignment, as the task items are being developed.*

Components of the Final Inquiry Task

Each **Inquiry Task** must include:

- A cohesive series of scoreable items, totaling 16-18 points, that assess student understanding in each of the four broad areas of inquiry, as described in the ***GDIT***.
- Scoreable items that have sufficient complexity for students to demonstrate scientific thinking related to inquiry.
- An identified DOK level for each scoreable item.
- A scoring rubric for each scoreable item.

APPENDIX B—NECAP SCIENCE COMMITTEE MEMBERS

Table B-1. 2014–15 NECAP Science: Item Review Committee Participants
August 7 and 8, 2014

State	*Name*	*School/Association Affiliation*	*Position*
New Hampshire	Mary Fougere	Gilmanton	Science teacher – Grades 7, 8
	Ann West	Bow Elementary	Teacher
	Marisa Bozek	Epping MS	Teacher
	Cheryl Patty	Westmoreland School	Science Teacher
	Leslie McRobie	Idlehurst Elementary	Teacher
	Kristen Gauthier	Ross A Lurgio	Science Teacher
	Kelly Marcotte	Richards School	Teacher – Grade 4
	Dennis Vienneau	Moultonborough Academy	Science Department Chair
	Judy Sidileau	Alvirne High School	Chem Teacher
	Mike Sedlisky	Alvirne High School	Chem Teacher
	Erik Anderson	Sage School	Teacher
	Rhetta Sabean	Alvirne High School	Science Teacher
Rhode Island	Melynda Antunes	Fort Barton School	Teacher
	Immacolata Beauchemin	Chariho Middle	Teacher
	Carey Bledsoe	The Learning Community	Teacher
	Kathryn Conley	Paul Cuffee School	Teacher
	Marie Contildes	N.A. Ferri Middle	Teacher
	Lisa DelBove	Tiogue School	Teacher
	Karen Finlan	North Kingstown High	Science Teacher
	Adam Flynn	Davies Vocational	Science Teacher
	Sophia Lambertsen	Blackstone Valley Prep	Teacher
	Frank Lenox	Greenwich High	Science Teacher
	Jennirfer Polacek	Globe Park Elementary	Teacher
	Maryjane Utley	Westerly High	Science Teacher
Vermont	Lynn Murphy	Waits River Valley	Science Teacher
	Jennifer Burdick	Burlington School District	Teacher
	Tamara Cooley	Barre Town Elementary	Teacher - Grade 4
	Dana Dezotell	Missisquoi Valley UHS	Science Teacher
	Jenny Hewitt	Pomfret School	Self Contained Teacher
	Nathan Reutter	Mt. Anthony High School	Science Teacher
	Charlotte Sherman	Washington West SU	Retired
	Lacey Smith	Whiting Elementary	Teacher
	Elizabeth Tarno	Warren Elementary	Teacher

Table B-2. 2014–15 NECAP Science: Bias and Sensitivity Committee Participants August 7, 2014

State	*Name*	*School/Association Affiliation*	*Position*
New Hampshire	Carol Lemay	Amherst Street School	Teacher
	Jill Glassmith	Lebanon Middle School	Self-Contained Teacher
	Erika Langlais	Gilmanton Schools	Spec Ed Teacher
	Jessica Fraser	Lebanon Middle School	Math Teacher
Rhode Island	Kathleen Beebe	Portsmouth High School	Chemistry teacher
	Angela Goulart	Calcutt Middle School	Science Teacher
	Shelagh Goulis	Northern Lincoln Elem	Teacher
	Joanne Lynch	Greystone Elementary	Teacher
	Kristen Skwirz	Pathways	Special Educator
	Melissa Valetta	Gladestone Elementary	Teacher
	Robert Wall	Pawtrucket School	Special Educator
Vermont	Aranka Gyuk	Integrated Arts Academy	ELL teacher
	Brenda Seitz	INSPIRE for Autism INC.	Director
	Susanna Mason	Addison Northwest SU	ELL Teacher
	Laura Nugent	Burlington School District	Special Educator

APPENDIX C—PARTICIPATION RATES

Table C-1. 2014–15 NECAP Science: Participation Rates

Description	*Tested*	
	Number	*Percent*
All Students	88,124	100.00
Male	45,050	51.12
Female	43,020	48.82
Gender Not Reported	54	0.06
Hispanic or Latino	9,018	10.23
American Indian or Alaskan Native	382	0.43
Asian	2,642	3.00
Black or African American	3,565	4.05
Native Hawaiian or Pacific Islander	79	0.09
White (non-Hispanic)	69,825	79.23
Two or More Races (non-Hispanic)	2,517	2.86
No Primary Race/Ethnicity Reported	96	0.11
Currently receiving LEP services	2,871	3.26
Former LEP student – monitoring year 1	894	1.01
Former LEP student – monitoring year 2	402	0.46
LEP: All Other Students	83,957	95.27
Students with an IEP	12,409	14.08
IEP: All Other Students	75,715	85.92
Economically Disadvantaged Students	31,113	35.31
SES: All Other Students	57,011	64.69
Migrant Students	32	0.04
Migrant: All Other Students	88,092	99.96
Students receiving Title 1 Services	14,868	16.87
Title 1: All Other Students	73,256	83.13
Plan 504	733	0.83
Plan 504: All Other Students	87,391	99.17

APPENDIX D—ACCOMMODATION FREQUENCIES

Table D-1. 2014–15 NECAP Science: Accommodation Frequencies

Accommodation	*Grade*		
	4	*8*	*11*
SCIAccomT1	2,862	1,629	776
SCIAccomT2	79	57	50
SCIAccomT3	2,016	1,139	430
SCIAccomT4	57	38	39
SCIAccomS1	3,088	1,804	1,370
SCIAccomS2	72	34	44
SCIAccomP1	680	309	181
SCIAccomP2	2,893	2,284	1,286
SCIAccomP3	2,391	926	339
SCIAccomP4	335	348	104
SCIAccomP5	802	371	120
SCIAccomP6	38	20	14
SCIAccomP7	1,214	390	226
SCIAccomP8	35	21	15
SCIAccomP9	2	3	0
SCIAccomP10	38	41	67
SCIAccomP11	48	20	4
SCIAccomR1	1,106	258	59
SCIAccomR2	17	53	5
SCIAccomR3	106	60	18
SCIAccomR4	38	86	41
SCIAccomR5	106	33	24
SCIAccomR6	5	7	4
SCIAccomR7	29	23	80
SCIAccomO1	2	0	0
SCIAccomM1	0	4	10
SCIAccomM3	0	0	0

APPENDIX E—NECAP TABLE OF STANDARD ACCOMMODATIONS

NECAP Table of Standard Accommodations

Revised August 2009

Any 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 accommodations must be made on an individual student basis – not for a large group, entire class, or grade level. 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 *NECAP Accommodations Guide*.

T. Timing		
Code	Tests were administered	Details on Delivery of Accommodations
T1	with time to complete a session extended beyond the scheduled administration time within the same day.	NECAP tests are not designed to be timed or speeded tests. 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.
T2	so that only a portion of the test session was administered on a particular day.	In rare and severe cases, the extended time accommodation (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 only has access to the items administered on that day (see the *NECAP Accommodations Guide* for details).
T3	with short, supervised breaks.	Multiple or frequent breaks may be required by a student whose attention span, distractibility, or physical condition, requires shorter working periods.
T4	at the time of day or day of week that takes into account the student's medical needs or learning style.	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 accommodation may not be used specifically to change the order of administration of test sessions. This accommodation 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.

S. Setting		
Code	Tests were administered	Details on Delivery of Accommodations
S1	in a separate location within the school by trained school personnel.	A student or students may be tested individually or in small groups in an alternative site within the school to reduce distractions for themselves or others, or to increase physical access to special equipment.
S2	in an out-of-school setting by trained 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.

P. Presentation		
Code	Tests were administered	Details on Delivery of Accommodations
P1	individually.	Individual or small group testing may be used to minimize distractions for a student or students whose test is administered out of the classroom or so that others will not be distracted by other accommodations being used (e.g., dictation)
P2	in a small group.	
P3	with test and directions read aloud in English or signed to the student. **(NOT allowed for the Reading test.)**	A reader may be used for a student whose inability to read would hinder performance on the Mathematics, Science, or Writing test. Words must be read as written. Guidelines for reading mathematical symbols must be followed. No translations (with the exception of signed language) or explanations are allowed. Trained personnel may use sign language to administer the test.
P4	with only test directions read aloud or signed to the student.	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 NECAP test occur at the beginning of the test session and are already read aloud by the test administrator. 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 routinely presents information in a foreign language, directions may not be translated.
P5	with administrator verification of student understanding following the reading of test directions.	After test directions have been read, the test administrator 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.
P6	using alternative or assistive technology that is part of the student's communication system.	The test may be presented through his/her regular communication system to a student who uses alternative or assistive technology on a daily basis. Technology may not be used to "read" the Reading test to the student.
P7	by trained school personnel known to the student other than the student's classroom teacher.	A student may be more comfortable with a test administrator who works with the student on a regular basis, but is not the student's regular teacher for the general curriculum or other staff assigned as test administrator. All test administrators must be trained school personnel familiar with test administration and accommodations procedures and guidelines.
P8	using a large-print version of assessment.	Both large-print and Braille versions of the assessment require special preparation and processing and must be pre-ordered. Directions for ordering these materials are included in communications sent to school principals prior to the test.
P9	using Braille version of assessment.	
P10	using a word-to-word translation dictionary for ELL students. **(NOT allowed for the Reading test.)**	A student with limited English proficiency may have a word-to-word dictionary available for individual use as needed. A word-to-word dictionary is one that does not include any definitions. Information on acceptable dictionaries is provided on the departments' websites.
P11	using 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, or acetate shields; or auditory devices such as special acoustics, amplification, noise buffers, whisper phones, or calming music.

R. Response		
Code	Tests were administered	Details on Delivery of Accommodations
R1	with a student dictating responses to school personnel. (**NOT allowed for the Writing test**. See O2 – using a scribe for the Writing test.)	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 test materials.
R2	with a student dictating responses using alternative or assistive technology/devices that are part of the student's communication system. (**NOT allowed for the Writing test**. See O2 – using a scribe for the Writing test.)	Technology is used to permit a student to respond to the test. When using a computer, word processing device, or other assistive technology, spell and grammar checks must be turned off. Policies regarding recorded answers must be followed prior to returning test materials.
R3	with a student using approved tools or devices 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.
R4	with a student writing responses using separate paper, a word processer, computer, brailler, or similar device.	A student may use technological or other tools (e.g., large-spaced paper) to write responses to constructed-response, short-answer, and extended 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, spell and grammar checks must be turned off, as well as access to the Web. This accommodation is intended for unique individual needs, not an entire class. Policies regarding recorded answers must be followed prior to returning test materials.
R5	with a student indicating responses to multiple-choice items to school personnel.	A student unable to write or otherwise unable to fill-in answers to multiple-choice questions may indicate a response to trained school personnel. The school personnel records the student's response in the student answer booklet.
R6	with a student responding with the use of visual aids.	Visual aids include any optical or non-optical 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 student with severe visual impairment or blindness on the Mathematics and Science tests. Note that the use of this accommodation still requires student responses to be recorded in a student answer booklet.
R7	with a student with limited English proficiency responding with use of a word-to-word dictionary. **(NOT allowed for the Reading test.)**	A student with limited English proficiency may have a word-to-word dictionary available for individual use as needed when responding. A word-to-word dictionary is one that does not include any definitions. Information on acceptable dictionaries is provided on each Department's website.

O. Other These accommodations require DOE approval.		
Code	Tests were administered	Details on Delivery of Accommodations
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. Non-approved accommodations used during test administration will be coded as an M3 modification.
O2	with a scribe used on the Writing test.	The use of a scribe for students dictating a response to the Writing test may only be used under limited circumstances and must be approved by the DOE. When approved as an accommodation, the scribe must follow established guidelines and procedures.

M. Modifications All modifications result in impacted items being scored as incorrect.		
Code	Tests were administered	Details on Delivery of Accommodations
M1	using a calculator and/or manipulatives on Session 1 of the Mathematics test or using a scientific or graphing calculator on Session 3 of the Science test	Inappropriate use of a calculator or other tools will result in impacted items being scored as incorrect.
M2	with the test administrator reading the Reading test.	The read aloud accommodation (P3) is not allowed for the Reading test. If it is used, all reading items in the sessions that are read aloud will be scored as incorrect.
M3	using an accommodation on this list not approved for a particular test or an accommodation not included on this list without prior approval of the DOE.	Inappropriate use of an accommodation included on this list or use of another accommodation without prior approval of the DOE will result in impacted items being scored as incorrect.

N. NimbleTools® 2011 The NECAP Science test was administered using NimbleTools online accommodations.	
Code	
N01	Allow breaks
N02	Read Aloud Text
N03	Magnifier
N04	Custom Masking or Answer Masking
N05	Color Overlay
N06	Reverse Contrast
N07	Font & Background Color Choice
N08	Auditory Calming
N09	Read Aloud Text and Graphics
N10	Microscope
N11	Magnifying Glass

Note: English Language Learners may qualify for any of the accommodations listed as appropriate and determined by a team. Refer to the *NECAP Accommodations Guide* for additional information.

APPENDIX F—ITEM-LEVEL CLASSICAL STATISTICS

Table F-1. 2014–15 NECAP Science: Item-Level Classical Test Theory Statistics—Grade 4

Item Number	Item Type	Difficulty	Discrimination	Percent Omitted
46263	MC	0.64	0.26	0
47508	MC	0.74	0.22	1
47591	MC	0.82	0.28	1
47662	MC	0.78	0.31	0
48046	MC	0.64	0.42	0
49872	MC	0.85	0.39	1
50385	CR	0.62	0.47	1
59429	MC	0.77	0.41	1
59915	MC	0.76	0.44	1
60292	MC	0.72	0.43	1
60342	MC	0.73	0.33	0
60348	MC	0.58	0.23	0
76739	MC	0.53	0.24	1
87201	MC	0.77	0.35	0
88128	MC	0.69	0.32	3
91741	MC	0.76	0.43	1
91965	MC	0.75	0.46	1
99017	CR	0.45	0.54	2
141411	CR	0.49	0.58	1
174255	MC	0.54	0.25	0
174283	MC	0.77	0.41	1
174373	MC	0.47	0.25	0
174377	MC	0.67	0.40	0
218371	MC	0.82	0.35	1
218403	MC	0.53	0.35	1
218439	MC	0.63	0.22	1
241826	MC	0.70	0.31	0
242407	MC	0.67	0.30	0
242427	MC	0.41	0.32	1
256584	MC	0.71	0.39	1
256592	MC	0.74	0.44	1
256625	MC	0.59	0.30	1
256651	MC	0.70	0.41	0
256679	MC	0.65	0.43	0
256714	MC	0.55	0.33	1
256767	MC	0.56	0.27	0
263753	SA	0.40	0.45	1
263754	SA	0.76	0.32	1
263757	CR	0.41	0.46	1
263761	SA	0.35	0.41	1
263763	SA	0.23	0.37	1
263765	SA	0.21	0.19	3
263766	SA	0.25	0.45	3
263767	CR	0.30	0.47	4

Table F-2. 2014–15 NECAP Science: Item-Level Classical Test Theory Statistics—Grade 8

Item Number	Item Type	Difficulty	Discrimination	Percent Omitted
48266	MC	0.69	0.37	1
48269	MC	0.46	0.33	1
50005	MC	0.73	0.39	1
50018	MC	0.79	0.37	1
50021	MC	0.61	0.24	1
50140	MC	0.67	0.37	1
58385	MC	0.77	0.49	1
58390	MC	0.58	0.32	1
60005	CR	0.26	0.54	3
60015	MC	0.46	0.20	1
60052	MC	0.71	0.36	1
60141	MC	0.67	0.35	1
60202	MC	0.50	0.38	1
82020	MC	0.54	0.29	1
82184	CR	0.34	0.65	4
87080	MC	0.43	0.33	1
91788	MC	0.66	0.42	1
144224	MC	0.60	0.37	1
144380	MC	0.57	0.37	1
144495	MC	0.70	0.51	1
144554	MC	0.76	0.35	1
174851	MC	0.62	0.36	1
175376	MC	0.68	0.24	1
219230	CR	0.27	0.53	2
219438	MC	0.59	0.30	1
219445	MC	0.50	0.30	1
220336	MC	0.46	0.44	1
241664	MC	0.53	0.35	1
242835	MC	0.53	0.35	1
256578	MC	0.71	0.42	1
256610	MC	0.77	0.30	1
256830	MC	0.63	0.45	1

continued

Item Number	Item Type	Difficulty	Discrimination	Percent Omitted
256882	MC	0.64	0.17	1
256933	MC	0.45	0.27	1
257855	MC	0.50	0.24	1
258033	MC	0.69	0.33	1
263771	SA	0.49	0.48	1
263773	CR	0.53	0.61	2

Item Number	Item Type	Difficulty	Discrimination	Percent Omitted
263780	SA	0.54	0.56	2
263781	SA	0.52	0.52	4
263782	CR	0.47	0.63	4
263784	SA	0.33	0.62	5
263785	SA	0.44	0.55	5
263786	SA	0.44	0.48	4

Table F-3. 2014–15 NECAP Science: Item-Level Classical Test Theory Statistics—Grade 11

Item Number	Item Type	Difficulty	Discrimination	Percent Omitted
46140	MC	0.74	0.27	1
47813	MC	0.39	0.29	1
47935	MC	0.38	0.30	1
48081	MC	0.75	0.40	1
48214	MC	0.58	0.21	1
49913	MC	0.87	0.39	1
49934	MC	0.80	0.39	1
59605	MC	0.81	0.38	1
59655	MC	0.28	0.21	1
60181	MC	0.64	0.22	1
60693	MC	0.47	0.28	2
61836	MC	0.65	0.31	1
89437	MC	0.77	0.52	1
90297	MC	0.70	0.49	1
91920	MC	0.64	0.48	2
146763	MC	0.57	0.34	2
146877	MC	0.41	0.19	1
146909	MC	0.56	0.32	1
146919	MC	0.42	0.20	1
146928	MC	0.55	0.30	1
146935	MC	0.66	0.49	1
147114	MC	0.45	0.31	1
169680	MC	0.67	0.42	1
176772	CR	0.45	0.58	3

Item Number	Item Type	Difficulty	Discrimination	Percent Omitted
176788	MC	0.55	0.37	1
177090	MC	0.53	0.27	1
177122	MC	0.48	0.32	1
241745	CR	0.25	0.56	7
241753	MC	0.48	0.40	1
242603	MC	0.50	0.35	1
242618	MC	0.53	0.27	1
242637	CR	0.37	0.59	5
256834	MC	0.38	0.24	1
256853	MC	0.30	0.32	1
256864	MC	0.48	0.35	1
256867	MC	0.55	0.47	1
262508	SA	0.40	0.49	3
262515	SA	0.34	0.51	5
262516	SA	0.34	0.55	5
262517	CR	0.45	0.47	7
262523	SA	0.31	0.52	7
262524	SA	0.23	0.47	7
262527	SA	0.25	0.47	9
262529	CR	0.26	0.59	7

APPENDIX G—ITEM-LEVEL SCORE POINT DISTRIBUTIONS

Table G-1. 2014–15 NECAP Science: Item-Level Score Distributions for Constructed Response Items by Subject and Grade

Grade	*Item Number*	*Total Possible Points*	*Percent of Students at Score Points*				
			0	*1*	*2*	*3*	*4*
4	50385	4	4.19	10.65	28.55	41.36	14.05
	99017	4	26.86	10.72	29.21	14.68	16.76
	141411	4	13.86	19.43	29.16	28.90	7.94
	263753	2	37.06	43.68	17.77		
	263754	2	4.70	37.39	57.33		
	263757	3	23.30	35.67	31.61	8.51	
	263761	2	40.22	46.89	11.50		
	263763	2	61.13	28.12	9.39		
	263765	2	57.23	37.32	2.58		
	263766	2	48.00	48.44	0.54		
	263767	3	25.75	54.50	13.56	2.62	
8	60005	4	19.17	56.37	17.73	3.40	0.62
	82184	4	24.79	29.61	24.19	13.71	4.07
	219230	4	26.84	42.81	20.01	6.56	1.60
	263771	2	20.49	57.74	20.35		
	263773	3	18.68	17.69	43.33	17.88	
	263780	2	25.13	38.13	34.76		
	263781	2	29.66	28.59	37.65		
	263782	3	26.48	15.15	35.16	18.88	
	263784	2	44.44	35.28	15.75		
	263785	2	29.32	44.59	21.27		
	263786	2	13.80	74.58	7.13		
11	176772	4	4.29	19.04	59.32	11.78	2.20
	241745	4	37.70	23.78	20.12	8.43	3.01
	242637	4	19.42	32.21	22.43	14.77	6.44
	262508	2	31.56	52.15	13.64		
	262515	2	34.38	53.64	6.98		
	262516	2	34.81	53.01	7.28		
	262517	3	36.57	6.33	23.39	27.04	
	262523	2	45.51	33.48	14.02		
	262524	2	58.82	22.18	11.93		
	262527	2	53.01	27.46	10.82		
	262529	3	38.88	32.83	18.15	3.29	

APPENDIX H—DIFFERENTIAL ITEM FUNCTIONING RESULTS

Table H-1. 2014–15 NECAP 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
4	Male	Female	CR	5	0	0	0	0	0	0
			MC	33	3	2	1	0	0	0
			SA	6	0	0	0	0	0	0
	No Disability	Disability	CR	5	0	0	0	0	0	0
			MC	33	5	5	0	0	0	0
			SA	6	0	0	0	0	0	0
	Non-EconDis	EconDis	CR	5	0	0	0	0	0	0
			MC	33	0	0	0	0	0	0
			SA	6	0	0	0	0	0	0
	Non-LEP	LEP	CR	5	0	0	0	0	0	0
			MC	33	8	8	0	0	0	0
			SA	6	0	0	0	0	0	0
	White	Asian	CR	5	0	0	0	0	0	0
			MC	33	4	4	0	0	0	0
			SA	6	0	0	0	0	0	0
8	Male	Female	CR	5	0	0	0	0	0	0
			MC	33	10	9	1	2	2	0
			SA	6	3	0	3	0	0	0
	No Disability	Disability	CR	5	2	2	0	0	0	0
			MC	33	2	0	2	0	0	0
			SA	6	4	4	0	0	0	0
	Non-EconDis	EconDis	CR	5	0	0	0	0	0	0
			MC	33	0	0	0	0	0	0
			SA	6	1	1	0	0	0	0
	Non-LEP	LEP	CR	5	0	0	0	0	0	0
			MC	33	10	10	0	1	1	0
			SA	6	1	1	0	0	0	0
	White	Asian	CR	5	0	0	0	0	0	0

continued

Grade	Group		Item Type	Number of Items	Number "Low"			Number "High"		
	Reference	Focal			Total	Favoring		Total	Favoring	
						Reference	Focal		Reference	Focal
8	White	Asian	MC	33	2	1	1	0	0	0
			SA	6	0	0	0	0	0	0
11	Male	Female	CR	5	2	0	2	0	0	0
			MC	33	8	7	1	1	1	0
			SA	6	0	0	0	0	0	0
	No Disability	Disability	CR	5	1	1	0	0	0	0
			MC	33	4	4	0	0	0	0
			SA	6	0	0	0	0	0	0
	Non-EconDis	EconDis	CR	5	0	0	0	0	0	0
			MC	33	0	0	0	0	0	0
			SA	6	0	0	0	0	0	0
	Non-LEP	LEP	CR	5	1	1	0	1	1	0
			MC	33	6	4	2	4	2	2
			SA	6	1	0	1	0	0	0
	White	Asian	CR	5	0	0	0	0	0	0
			MC	33	6	4	2	0	0	0
			SA	6	0	0	0	0	0	0

APPENDIX I—ITEM RESPONSE THEORY CALIBRATION RESULTS

Table I-1. 2014–15 NECAP Science: IRT Parameters for Dichotomous Items—Grade 4

Item Number	Parameters and Measures of Standard Error					
	a	*SE (a)*	*b*	*SE (b)*	*c*	*SE (c)*
87201	0.59071	0.01556	-1.44302	0.07374	0.10674	0.02926
174255	0.31605	0.00725	-0.44380	0.02339	0.00000	0.00000
47591	0.47468	0.01233	-2.10734	0.09697	0.08293	0.03460
256679	0.91532	0.02157	-0.32119	0.02682	0.20630	0.01126
60348	0.30709	0.00725	-0.82867	0.02759	0.00000	0.00000
256651	0.75769	0.01896	-0.71109	0.04137	0.17442	0.01681
241826	0.46033	0.01325	-1.16734	0.08533	0.07705	0.02743
48046	0.80203	0.01946	-0.38522	0.03190	0.16827	0.01296
60342	0.70624	0.02318	-0.52011	0.05705	0.34156	0.01792
47662	0.50445	0.01392	-1.64828	0.09287	0.09516	0.03349
60292	0.82021	0.01906	-0.78889	0.03640	0.16115	0.01586
242427	0.73766	0.02244	0.70918	0.02355	0.14838	0.00824
59429	0.71554	0.01540	-1.32985	0.04504	0.07354	0.02030
91965	0.93258	0.01970	-0.93129	0.03047	0.14315	0.01478
91741	0.78267	0.01778	-1.13304	0.04268	0.12274	0.01949
218439	0.31035	0.01205	-0.94017	0.14586	0.07962	0.03299
256767	0.46640	0.02038	0.08676	0.08010	0.18474	0.02213
46263	0.36033	0.01167	-0.93042	0.10275	0.06487	0.02664
174377	0.71101	0.01807	-0.60212	0.04180	0.14772	0.01648
174373	0.45519	0.02141	0.64086	0.06289	0.15614	0.01806
242407	0.56237	0.02080	-0.42395	0.07498	0.25715	0.02215
256714	0.54446	0.01741	-0.09691	0.05205	0.11290	0.01721
218371	0.61294	0.01321	-1.85301	0.05772	0.06140	0.02447
47508	0.33083	0.01381	-1.77903	0.21384	0.13477	0.05178
218403	0.68034	0.02027	0.18484	0.03383	0.16943	0.01196
49872	0.82088	0.02055	-1.58274	0.05738	0.19167	0.02764
256584	0.77969	0.02103	-0.58899	0.04230	0.24600	0.01608
59915	0.82413	0.01759	-1.10166	0.03669	0.10382	0.01734
256625	0.45896	0.01738	-0.28733	0.08289	0.12259	0.02429
256592	0.84050	0.01937	-0.89827	0.03719	0.16619	0.01665
76739	0.38080	0.01938	0.17939	0.10962	0.13623	0.02750
174283	0.73932	0.01746	-1.19117	0.04849	0.12443	0.02152
88128	0.46944	0.01266	-1.12860	0.07477	0.06575	0.02439

Table I-2. 2014–15 NECAP Science: IRT Parameters for Polytomous Items—Grade 4

Item Number	Parameters and Measures of Standard Error													
	a	*SE (a)*	*b*	*SE (b)*	*D0*	*SE (D0)*	*D1*	*SE (D1)*	*D2*	*SE (D2)*	*D3*	*SE (D3)*	*D4*	*SE (D4)*
263753	0.61083	0.00468	0.45718	0.01114	1.16640	0.01273	-1.16640	0.01544	0	0				
263754	0.39195	0.00358	-2.69439	0.01939	2.08980	0.03933	-2.08980	0.01849	0	0				
263761	0.55404	0.00422	0.90146	0.01254	1.46573	0.01365	-1.46573	0.02002	0	0				
263763	0.56424	0.00528	1.56881	0.01470	1.04226	0.01358	-1.04226	0.02153	0	0				
263765	0.28214	0	4.23296	0.02778	3.44229	0.02540	-3.44229	0.07659	0	0				
263766	0.71065	0.00751	2.34493	0.01093	2.41815	0.01098	-2.41815	0.06583	0	0				

continued

Item Number	Parameters and Measures of Standard Error													
	a	*SE (a)*	*b*	*SE (b)*	*D0*	*SE (D0)*	*D1*	*SE (D1)*	*D2*	*SE (D2)*	*D3*	*SE (D3)*	*D4*	*SE (D4)*
263757	0.58210	0.00367	0.50655	0.01101	2.01924	0.01481	0.13225	0.01296	-2.15149	0.02173	0	0		
263767	0.66403	0.00433	1.37328	0.01043	2.48713	0.01263	-0.28594	0.01488	-2.20119	0.03261	0	0		
141411	0.80483	0.00470	-0.00482	0.00785	1.86443	0.01352	0.75974	0.01021	-0.44987	0.01000	-2.17430	0.01668	0	0
50385	0.57453	0.00330	-0.98119	0.01104	2.55199	0.02711	1.11997	0.01697	-0.63315	0.01296	-3.03880	0.01794	0	0
99017	0.75626	0.00500	0.13525	0.00846	1.19635	0.01132	0.69800	0.01047	-0.54584	0.01080	-1.34851	0.01311	0	0

Table I-3. 2014–15 NECAP Science: IRT Parameters for Dichotomous Items—Grade 8

Item Number	Parameters and Measures of Standard Error					
	a	*SE (a)*	*b*	*SE (b)*	*c*	*SE (c)*
256882	0.24533	0.01284	-0.94042	0.25927	0.10860	0.04671
174851	0.53902	0.01345	-0.50702	0.04892	0.04888	0.01743
241664	0.79337	0.02457	0.42726	0.02849	0.22800	0.01043
256933	0.99150	0.03580	1.05552	0.02048	0.29022	0.00640
50018	0.70669	0.01874	-1.17742	0.06453	0.14968	0.02876
144380	0.75915	0.02202	0.16209	0.03236	0.19370	0.01225
91788	1.02752	0.02571	-0.09815	0.02514	0.27492	0.01059
220336	0.97344	0.02268	0.44013	0.01664	0.12811	0.00703
50005	0.71367	0.01863	-0.79256	0.05181	0.14254	0.02200
50021	0.35666	0.01756	-0.43504	0.16065	0.12357	0.03931
50140	0.63633	0.01853	-0.53463	0.05687	0.14132	0.02161
257855	0.37920	0.01946	0.38424	0.10456	0.10534	0.02745
256830	0.83426	0.01898	-0.31847	0.02913	0.11889	0.01294
144554	0.56407	0.00968	-1.45633	0.02385	0.00000	0.00000
60015	0.80347	0.03900	1.34403	0.02949	0.33061	0.00753
60202	0.87938	0.02452	0.47115	0.02233	0.19862	0.00872
60052	0.60470	0.01573	-0.90723	0.06056	0.08553	0.02419
242835	0.73923	0.02342	0.40228	0.03140	0.20727	0.01139

Item Number	Parameters and Measures of Standard Error					
	a	*SE (a)*	*b*	*SE (b)*	*c*	*SE (c)*
144495	1.22504	0.02498	-0.41741	0.01949	0.19416	0.01021
175376	0.34367	0.00795	-1.35454	0.03489	0.00000	0.00000
60141	0.68755	0.02149	-0.29573	0.05171	0.22972	0.01847
256610	0.47737	0.00922	-1.68718	0.03124	0.00000	0.00000
219438	0.47684	0.01896	-0.14693	0.08476	0.12984	0.02576
256578	0.75557	0.01740	-0.76616	0.04076	0.09667	0.01845
48269	0.78559	0.02567	0.72294	0.02497	0.20551	0.00903
258033	0.49867	0.00879	-1.10663	0.02182	0.00000	0.00000
48266	0.65039	0.01832	-0.65202	0.05705	0.13772	0.02239
87080	0.72713	0.02367	0.77229	0.02522	0.15581	0.00925
144224	0.76907	0.02278	0.09269	0.03463	0.22583	0.01285
82020	0.43611	0.01732	0.00495	0.08291	0.08966	0.02440
219445	0.41440	0.00802	0.01109	0.01720	0.00000	0.00000
58385	1.15625	0.02384	-0.74890	0.02442	0.17951	0.01338
58390	0.53680	0.01896	-0.08059	0.06368	0.12844	0.02106

Table I-4. 2014–15 NECAP Science: IRT Parameters for Polytomous Items—Grade 8

Item Number	Parameters and Measures of Standard Error													
	a	*SE (a)*	*b*	*SE (b)*	*D0*	*SE (D0)*	*D1*	*SE (D1)*	*D2*	*SE (D2)*	*D3*	*SE (D3)*	*D4*	*SE (D4)*
263771	0.71189	0.00500	0.03830	0.00975	1.37106	0.01276	-1.37106	0.01283	0	0				
263780	0.86726	0.00679	-0.15788	0.00786	0.75966	0.01016	-0.75966	0.00942	0	0				
263781	0.77571	0.00659	-0.09169	0.00874	0.59158	0.01050	-0.59158	0.01013	0	0				
263784	1.20713	0.00974	0.63286	0.00632	0.65067	0.00715	-0.65067	0.00906	0	0				
263785	0.86863	0.00655	0.28727	0.00793	0.90485	0.00962	-0.90485	0.01065	0	0				
263786	0.85759	0.00612	0.41666	0.00976	1.82348	0.01177	-1.82348	0.01633	0	0				
263773	0.93968	0.00613	-0.08214	0.00701	1.10597	0.01032	0.31577	0.00870	-1.42174	0.01063	0	0		
263782	1.03485	0.00713	0.13720	0.00645	0.82903	0.00857	0.25494	0.00789	-1.08397	0.00963	0	0		
219230	0.79490	0.00490	1.44484	0.00825	2.34501	0.01074	0.56680	0.01043	-0.76774	0.01620	-2.14408	0.03451	0	0
60005	0.88370	0.00552	1.65178	0.00788	2.83199	0.01078	0.50510	0.01038	-0.99215	0.02016	-2.34494	0.04947	0	0
82184	1.16237	0.00716	0.75023	0.00567	1.48926	0.00806	0.45747	0.00723	-0.44795	0.00882	-1.49878	0.01572	0	0

Table I-5. 2014–15 NECAP Science: IRT Parameters for Dichotomous Items—Grade 11

Item Number	Parameters and Measures of Standard Error					
	a	*SE (a)*	*b*	*SE (b)*	*c*	*SE (c)*
49934	0.70574	0.01488	-1.19076	0.04670	0.06255	0.02095
47935	1.13994	0.03617	1.33717	0.01633	0.21759	0.00511
242603	0.86452	0.02594	0.79964	0.02320	0.22989	0.00835
169680	0.95305	0.02417	0.04576	0.02774	0.26430	0.01102
49913	0.83950	0.01459	-1.58457	0.02905	0.02806	0.01232
90297	1.06360	0.02258	-0.23469	0.02295	0.18090	0.01072
47813	1.46890	0.04589	1.32080	0.01373	0.25467	0.00444
89437	1.16131	0.02149	-0.67041	0.02076	0.09284	0.01104
60181	0.29810	0.00744	-1.01042	0.03695	0.00000	0.00000
146877	0.58966	0.03549	1.87308	0.04223	0.26150	0.00988
48081	0.75614	0.01966	-0.60910	0.04826	0.18778	0.01954
146928	0.51664	0.02048	0.43918	0.06136	0.16563	0.01852
176788	0.79160	0.02319	0.52450	0.02855	0.21815	0.01031

Item Number	Parameters and Measures of Standard Error					
	a	*SE (a)*	*b*	*SE (b)*	*c*	*SE (c)*
146919	1.00011	0.04671	1.70757	0.02564	0.31526	0.00561
61836	0.45390	0.01514	-0.49121	0.08956	0.08905	0.02727
241753	1.13959	0.02896	0.79035	0.01582	0.21647	0.00628
146909	0.59650	0.02131	0.46022	0.04737	0.19345	0.01517
46140	0.40838	0.00835	-1.44856	0.03442	0.00000	0.00000
256834	0.67600	0.03101	1.61678	0.03042	0.21212	0.00835
146935	0.97129	0.01979	-0.20358	0.02272	0.11893	0.01040
59605	0.69874	0.01369	-1.31898	0.04213	0.04727	0.01852
59655	1.08156	0.04416	1.85739	0.02343	0.18344	0.00434
48214	0.27166	0.00723	-0.51090	0.03188	0.00000	0.00000
177122	0.73054	0.02524	0.93263	0.02839	0.21901	0.00961
147114	0.56693	0.02160	0.93930	0.03809	0.14106	0.01261
256864	0.57334	0.01802	0.55489	0.03793	0.08797	0.01306

continued

Item Number	Parameters and Measures of Standard Error					
	a	*SE (a)*	*b*	*SE (b)*	*c*	*SE (c)*
242618	0.35196	0.00757	0.00514	0.02117	0.00000	0.00000
256853	1.76710	0.04849	1.36046	0.01065	0.16489	0.00343
177090	0.35507	0.00759	-0.02149	0.02111	0.00000	0.00000
256867	0.91876	0.01979	0.23588	0.02001	0.11206	0.00848
146763	0.57063	0.01974	0.28719	0.05233	0.15991	0.01692

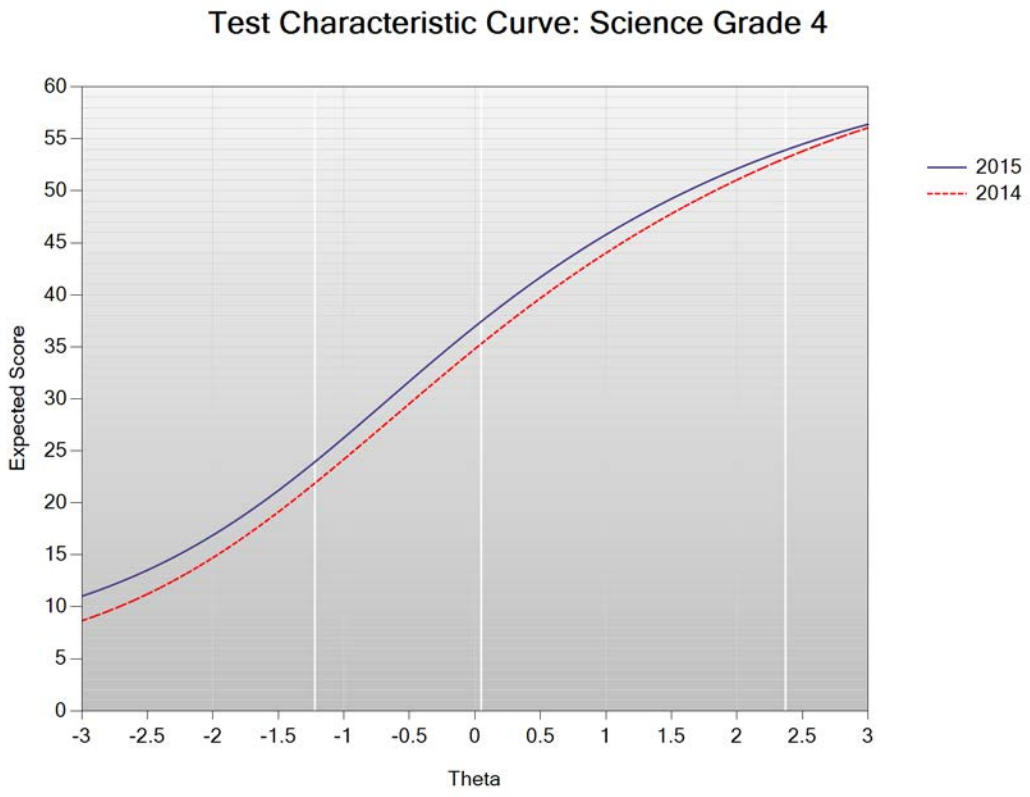
Item Number	Parameters and Measures of Standard Error					
	a	*SE (a)*	*b*	*SE (b)*	*c*	*SE (c)*
91920	1.28930	0.02911	0.19149	0.01730	0.26344	0.00784
60693	0.42913	0.01855	0.71970	0.06656	0.08983	0.01936

Table I-6. 2014–15 NECAP Science: IRT Parameters for Polytomous Items—Grade 11

Item Number	Parameters and Measures of Standard Error													
	a	*SE (a)*	*b*	*SE (b)*	*D0*	*SE (D0)*	*D1*	*SE (D1)*	*D2*	*SE (D2)*	*D3*	*SE (D3)*	*D4*	*SE (D4)*
262508	0.71814	0.00538	0.80435	0.01008	1.29120	0.01175	-1.29120	0.01515	0	0				
262515	0.78475	0.00610	1.20543	0.00972	1.42469	0.01067	-1.42469	0.01834	0	0				
262516	0.89772	0.00704	1.12243	0.00867	1.28781	0.00962	-1.28781	0.01605	0	0				
262523	0.83040	0.00710	1.12649	0.00930	0.79174	0.00986	-0.79174	0.01324	0	0				
262524	0.79050	0.00777	1.49646	0.01134	0.61748	0.01060	-0.61748	0.01457	0	0				
262527	0.74576	0.00691	1.49320	0.01126	0.78482	0.01092	-0.78482	0.01590	0	0				
262517	0.62729	0.00548	0.48876	0.01102	0.59265	0.01247	0.28615	0.01229	-0.87880	0.01357	0	0		
262529	0.96939	0.00707	1.43815	0.00762	1.37074	0.00882	0.08341	0.00999	-1.45415	0.02108	0	0		
176772	0.93516	0.00573	0.56757	0.00777	2.60073	0.01595	1.28462	0.01016	-1.22479	0.01191	-2.66056	0.02614	0	0
241745	0.88525	0.00613	1.52152	0.00803	1.51335	0.00944	0.58845	0.00967	-0.50699	0.01333	-1.59481	0.02389	0	0
242637	0.88864	0.00553	0.88518	0.00743	1.75307	0.01088	0.41277	0.00922	-0.53488	0.01071	-1.63095	0.01688	0	0

APPENDIX J—TEST CHARACTERISTIC CURVE AND TEST INFORMATION FUNCTION CHARTS

Figure J-1. 2014–15 NECAP Science: Grade 4 Charts
Top: Test Characteristic Curve Bottom: Test Information Function



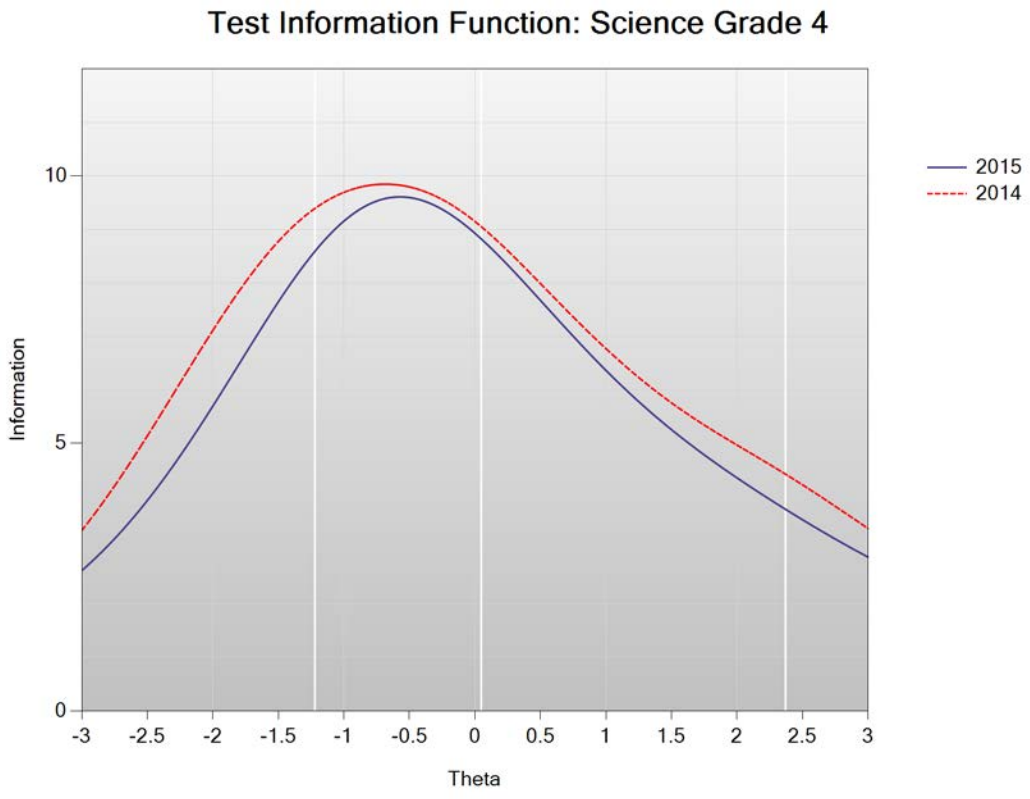
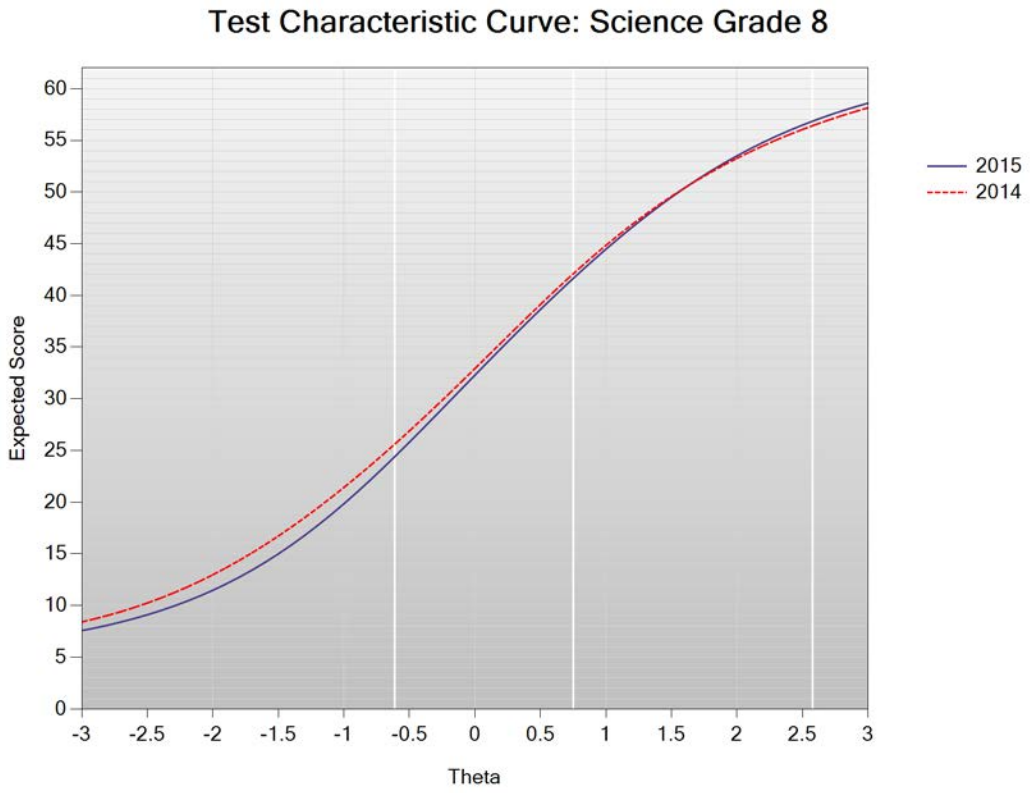


Figure J-2. 2014–15 NECAP Science: Grade 8 Charts
Top: Test Characteristic Curve Bottom: Test Information Function



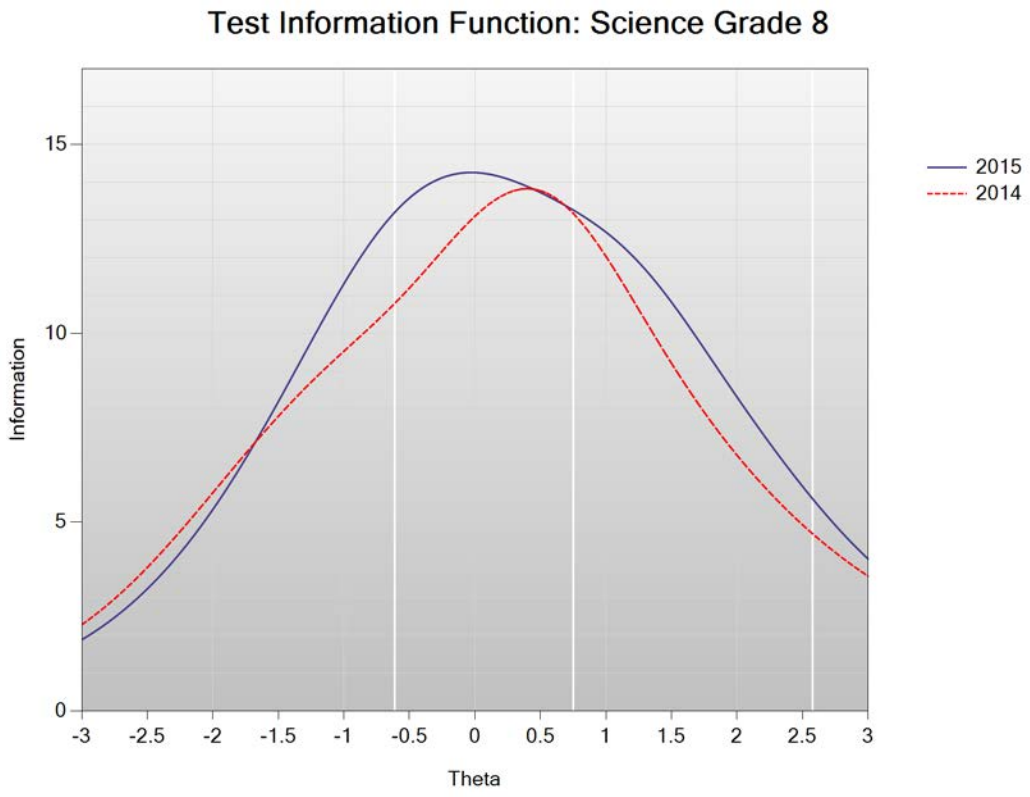
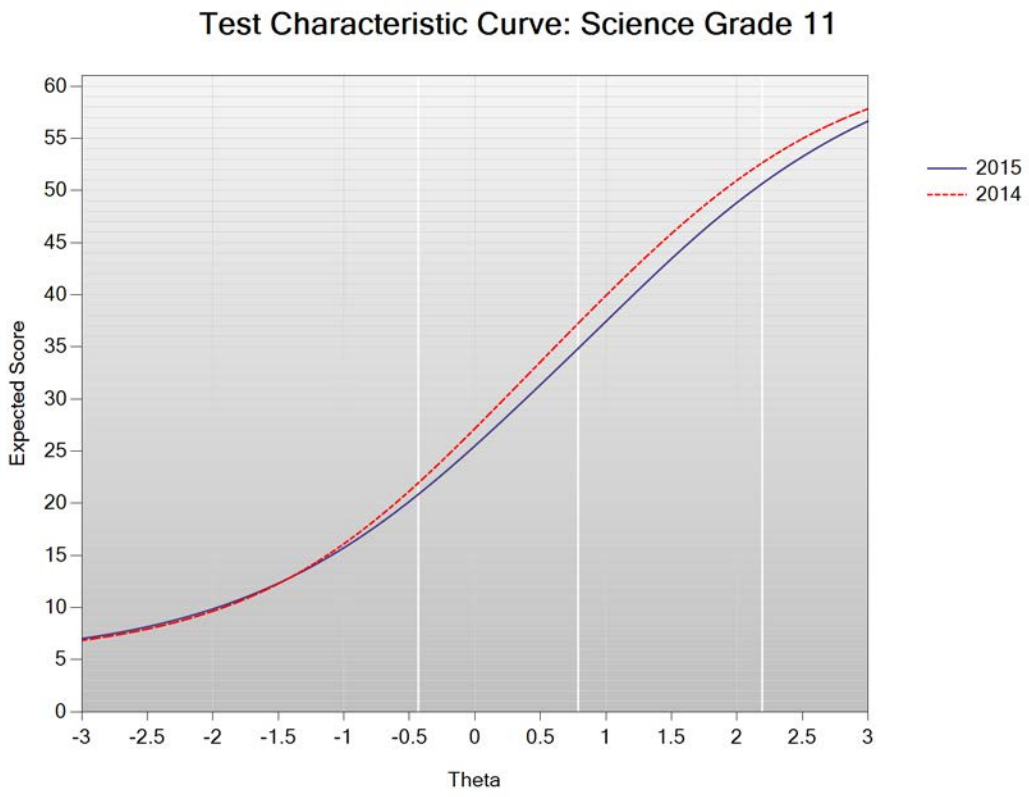
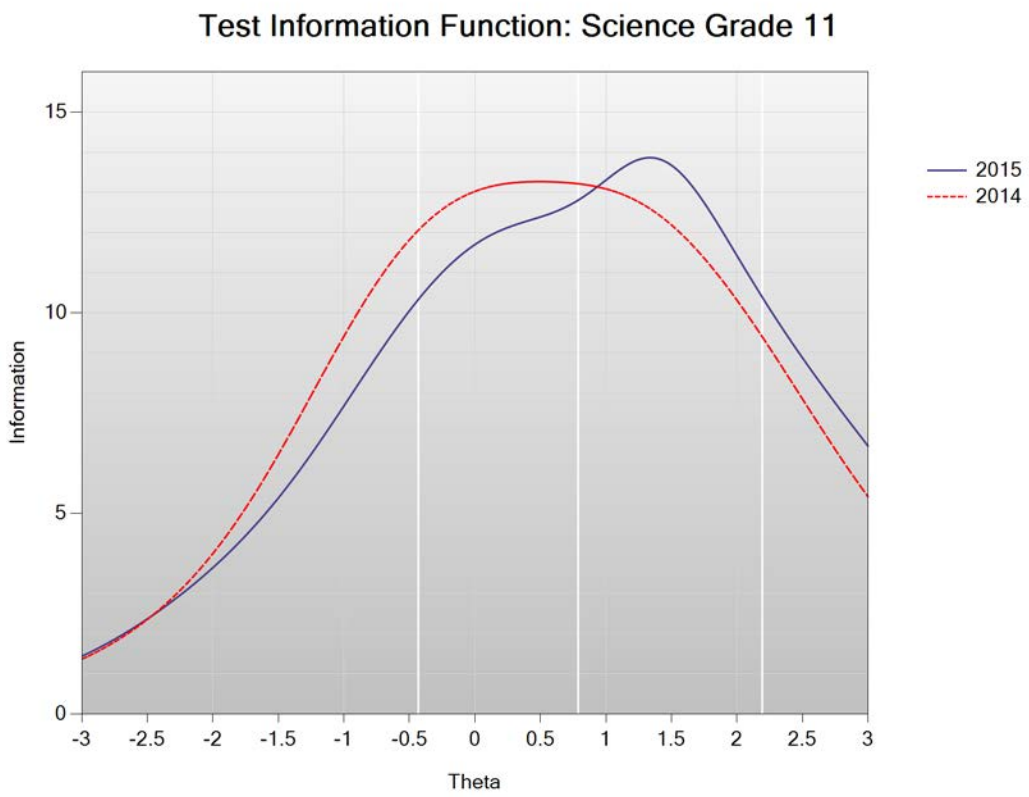


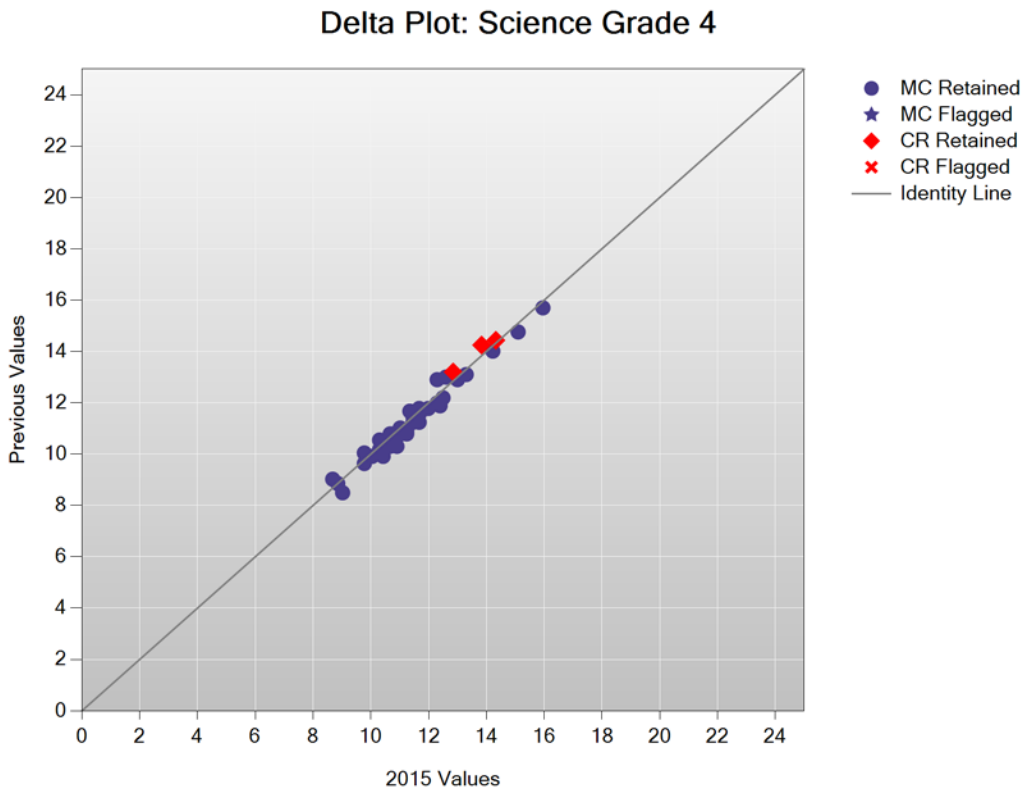
Figure J-3. 2014–15 NECAP Science: Grade 11 Charts
Top: Test Characteristic Curve Bottom: Test Information Function





APPENDIX K—DELTA ANALYSES AND RESCORE ANALYSES

Figure K-1. 2014–15 NECAP Science: Delta Plots
Top: Grade 4 Bottom: Grade 8



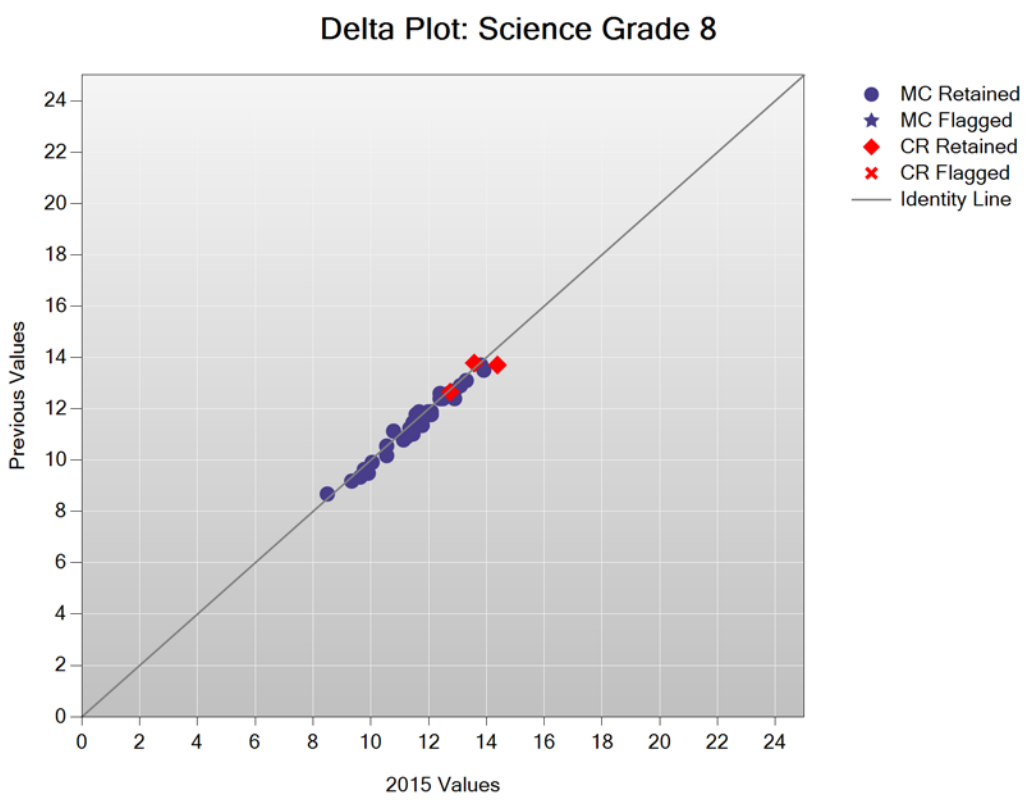


Figure K-2. 2014–15 NECAP Science: Delta Plot—Grade 11

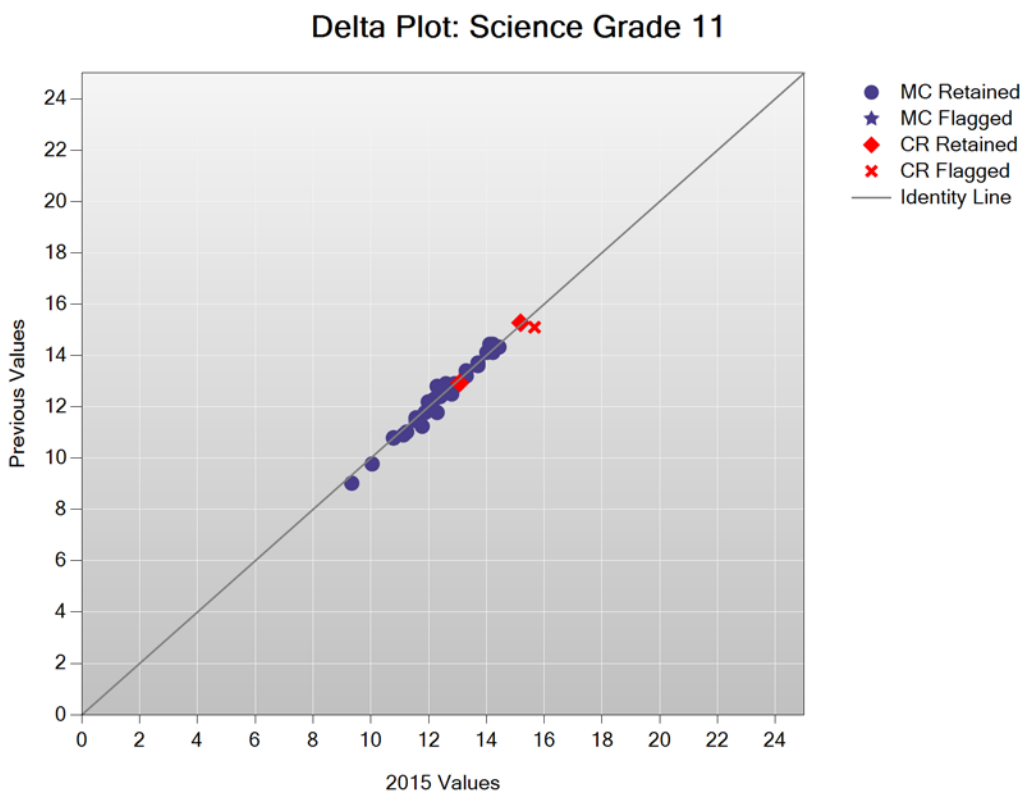


Table K-1. 2014–15 NECAP Science: Delta Analyses Results—Grade 4

Item Number	*Difficulty*		*Delta*		*Discard*	*Standard Deviation*
	Old	*New*	*Old*	*New*		
141192	0.51000	0.50000	12.89972	13.00000	False	-1.25602
142220	0.25000	0.23000	15.69796	15.95539	False	-0.02914
142224	0.67000	0.63000	11.24035	11.67259	False	0.55235
144866	0.62000	0.60000	11.77808	11.98661	False	-0.72634
144871	0.51000	0.57000	12.89972	12.29450	False	2.59725
144881	0.33000	0.30000	14.75965	15.09760	False	0.35406
144907	0.48000	0.51500	13.20061	12.84957	False	1.05005
174214	0.78000	0.77000	9.91123	10.04461	False	-1.36774
217967	0.40000	0.38000	14.01339	14.22192	False	-0.49507
256682	0.87000	0.84000	8.49444	9.02217	False	0.83783
46310	0.84000	0.86000	9.02217	8.67872	False	1.43703
46402	0.71000	0.67000	10.78646	11.24035	False	0.63450
46416	0.63000	0.66000	11.67259	11.35015	False	1.03751
47361	0.77000	0.79000	10.04461	9.77432	False	0.89494
47418	0.60000	0.57000	11.98661	12.29450	False	-0.11213
47448	0.75000	0.70000	10.30204	10.90240	False	1.45803
47493	0.71000	0.72000	10.78646	10.66863	False	-0.09125
47531	0.37750	0.41750	14.24821	13.83317	False	1.32336

continued

Item Number	Difficulty		Delta		Discard	Standard Deviation
	Old	New	Old	New		
47551	0.71000	0.71000	10.78646	10.78646	False	-0.79405
47624	0.58000	0.55000	12.19243	12.49735	False	-0.10851
47760	0.62000	0.63000	11.77808	11.67259	False	-0.26743
47992	0.76000	0.75000	10.17479	10.30204	False	-1.37706
49850	0.73000	0.75000	10.54875	10.30204	False	0.70207
49866	0.85000	0.85000	8.85427	8.85427	False	-0.59414
49894	0.65000	0.64000	11.45872	11.56616	False	-1.36234
49897	0.65000	0.65000	11.45872	11.45872	False	-0.86361
59267	0.67000	0.65000	11.24035	11.45872	False	-0.72331
59434	0.75000	0.72000	10.30204	10.66863	False	0.06371
59640	0.36000	0.37000	14.43384	14.32741	False	-0.53665
61931	0.78000	0.74000	9.91123	10.42662	False	0.91080
87135	0.69000	0.69000	11.01660	11.01660	False	-0.81786
87181	0.70000	0.67000	10.90240	11.24035	False	-0.04503
88061	0.66000	0.64000	11.35015	11.56616	False	-0.72599
88083	0.50000	0.54000	13.00000	12.59827	False	1.37314
88090	0.80000	0.79000	9.63352	9.77432	False	-1.35224
91504	0.61000	0.56000	11.88272	12.39612	False	1.10290
91509	0.72000	0.70000	10.66863	10.90240	False	-0.69065
91514	0.64000	0.63000	11.56616	11.67259	False	-1.35734
91971	0.49000	0.47000	13.10028	13.30108	False	-0.63565

Table K-2. 2014–15 NECAP Science: Rescore Analyses Results—Grade 4

Item Number	Number of Score Categories	Average Score		Standard Deviation		Effect Size	Discard
		Old	New	Old	New		
144907	4	2.04401	2.16137	1.31843	1.27893	0.08901	False
59640	4	1.43137	1.48039	0.96636	0.88925	0.05073	False
47531	4	1.48780	1.65366	1.09057	1.19986	0.15208	False

Table K-3. 2014–15 NECAP Science: Delta Analyses Results—Grade 8

Item Number	Difficulty		Delta		Discard	Standard Deviation
	Old	New	Old	New		
144589	0.66000	0.62000	11.35015	11.77808	False	0.70154
174954	0.81000	0.78000	9.48841	9.91123	False	0.89239
241628	0.86000	0.87000	8.67872	8.49444	False	0.80180
241639	0.62000	0.59000	11.77808	12.08982	False	-0.16849
256841	0.43000	0.36500	13.70550	14.38050	False	2.15378
260771	0.82000	0.80000	9.33854	9.63352	False	0.01076
46016	0.71000	0.68000	10.78646	11.12920	False	0.17058
46026	0.70000	0.67000	10.90240	11.24035	False	0.12270
46039	0.42250	0.44250	13.78201	13.57854	False	1.55875
46041	0.56000	0.51000	12.39612	12.89972	False	1.10676
46070	0.45000	0.41000	13.50265	13.91018	False	0.29566

continued

Item Number	Difficulty		Delta		Discard	Standard Deviation
	Old	New	Old	New		
46072	0.69000	0.65000	11.01660	11.45872	False	0.84208
46074	0.62000	0.59000	11.77808	12.08982	False	-0.16849
46082	0.61000	0.59000	11.88272	12.08982	False	-0.91789
46089	0.73000	0.73000	10.54875	10.54875	False	-0.26759
48184	0.76000	0.73000	10.17479	10.54875	False	0.46485
48228	0.67000	0.66000	11.24035	11.35015	False	-0.95623
48245	0.78000	0.77000	9.91123	10.04461	False	-1.19652
48267	0.62000	0.64000	11.77808	11.56616	False	1.37396
48268	0.61000	0.60000	11.88272	11.98661	False	-0.83633
48297	0.61000	0.63000	11.88272	11.67259	False	1.37422
48355	0.66000	0.65000	11.35015	11.45872	False	-0.93420
48421	0.54000	0.53000	12.59827	12.69892	False	-0.72638
48445	0.53000	0.51000	12.69892	12.89972	False	-1.06165
48449	0.53500	0.52500	12.64862	12.74917	False	-0.71952
48456	0.43000	0.42000	13.70550	13.80757	False	-0.60146
48563	0.80000	0.79000	9.63352	9.77432	False	-1.11048
49990	0.56000	0.56000	12.39612	12.39612	False	-0.04246
50012	0.68000	0.71000	11.12920	10.78646	False	2.21586
50026	0.64000	0.62000	11.56616	11.77808	False	-0.84542
50120	0.49000	0.47000	13.10028	13.30108	False	-1.11056
50511	0.83000	0.82000	9.18334	9.33854	False	-0.95425
76623	0.51000	0.49000	12.89972	13.10028	False	-1.08790
76626	0.63000	0.62000	11.67259	11.77808	False	-0.87322
90304	0.54000	0.56000	12.59827	12.39612	False	1.40513
91663	0.56000	0.55000	12.39612	12.49735	False	-0.75507
91775	0.65000	0.65000	11.45872	11.45872	False	-0.15669

Table K-4. 2014–15 NECAP Science: Rescore Analyses Results—Grade 8

Item Number	Number of Score Categories	Average Score		Standard Deviation		Effect Size	Discard
		Old	New	Old	New		
256841	4	1.83374	1.73350	1.27446	1.23863	-0.07866	False
46039	4	1.86765	2.03431	1.09371	1.08770	0.15239	False
48449	4	2.03415	2.23415	0.97087	1.05766	0.20600	False

Table K-5. 2014–15 NECAP Science: Delta Analyses Results—Grade 11

Item Number	Difficulty		Delta		Discard	Standard Deviation
	Old	New	Old	New		
135357	0.55000	0.52000	12.49735	12.79939	False	0.66269
146762	0.36000	0.39000	14.43384	14.11728	False	0.94957
146884	0.44000	0.43000	13.60388	13.70550	False	-0.47369
146889	0.28500	0.29250	15.27221	15.18438	False	-0.80783
146924	0.62000	0.57000	11.77808	12.29450	False	1.98012
147033	0.56000	0.56000	12.39612	12.39612	False	-0.78186

continued

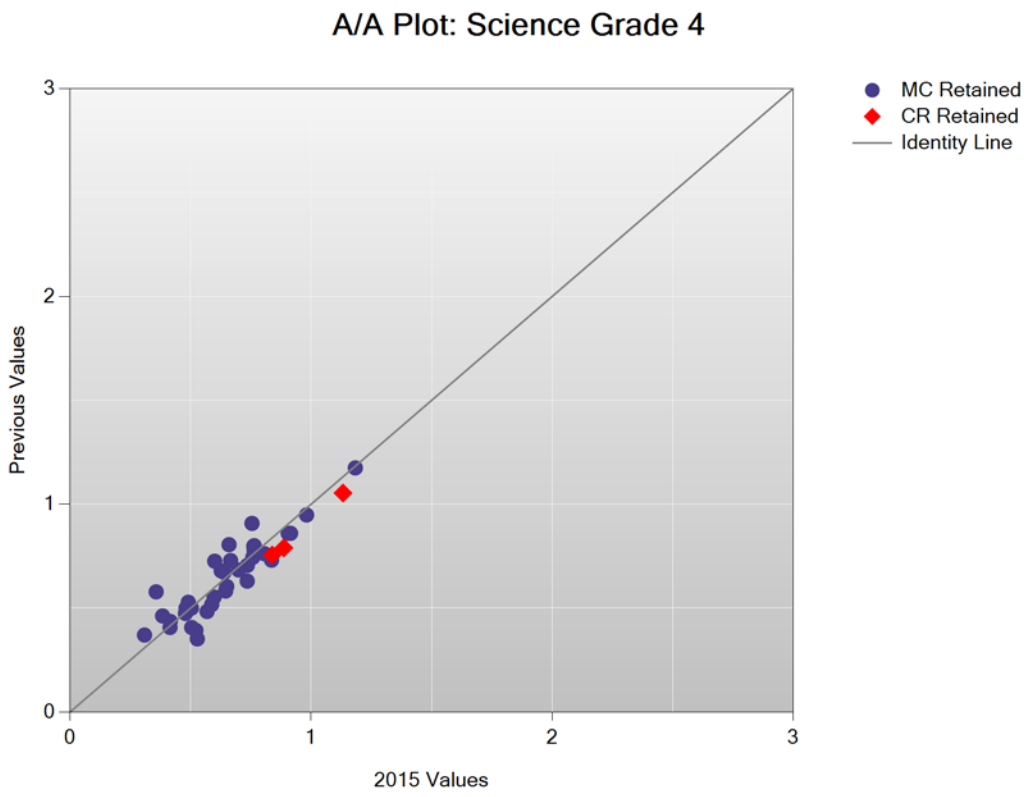
Item Number	Difficulty		Delta		Discard	Standard Deviation
	Old	New	Old	New		
169678	0.36000	0.38000	14.43384	14.22192	False	0.22961
176774	0.30000	0.25250	15.09760	15.66657	True	3.06897
176916	0.84000	0.82000	9.02217	9.33854	False	-0.00011
177100	0.50750	0.49250	12.92480	13.07520	False	-0.28683
219123	0.51000	0.51000	12.89972	12.89972	False	-0.89220
219519	0.71000	0.71000	10.78646	10.78646	False	-0.42916
242620	0.43000	0.43000	13.70550	13.70550	False	-1.06876
46019	0.51000	0.54000	12.89972	12.59827	False	1.18183
46096	0.58000	0.60000	12.19243	11.98661	False	0.67877
46148	0.65000	0.64000	11.45872	11.56616	False	-0.90362
46166	0.57000	0.58000	12.29450	12.19243	False	-0.05730
46179	0.70000	0.68000	10.90240	11.12920	False	-0.20433
47795	0.53000	0.54000	12.69892	12.59827	False	-0.15570
47884	0.71000	0.71000	10.78646	10.78646	False	-0.42916
47917	0.53000	0.52000	12.69892	12.79939	False	-0.67992
48005	0.55000	0.55000	12.49735	12.49735	False	-0.80404
48156	0.79000	0.77000	9.77432	10.04461	False	-0.15229
48357	0.39000	0.38000	14.11728	14.22192	False	-0.34037
48543	0.52000	0.53000	12.79939	12.69892	False	-0.17902
49902	0.48000	0.47000	13.20061	13.30108	False	-0.56999
49903	0.64000	0.64000	11.56616	11.56616	False	-0.60000
49922	0.62000	0.61000	11.77808	11.88272	False	-0.85291
49925	0.64000	0.64000	11.56616	11.56616	False	-0.60000
49930	0.37000	0.36000	14.32741	14.43384	False	-0.28211
49935	0.69000	0.67000	11.01660	11.24035	False	-0.20034
61142	0.71000	0.71000	10.78646	10.78646	False	-0.42916
62083	0.71000	0.71000	10.78646	10.78646	False	-0.42916
89333	0.46000	0.47000	13.40173	13.30108	False	-0.30969
89407	0.67000	0.62000	11.24035	11.77808	False	2.00886
89632	0.39000	0.40000	14.11728	14.01339	False	-0.44424
91901	0.52000	0.57000	12.79939	12.29450	False	2.60336

Table K-6. 2014–15 NECAP Science: Rescore Analyses Results—Grade 11

Item Number	Number of Score Categories	Average Score		Standard Deviation		Effect Size	Discard
		Old	New	Old	New		
146889	4	1.24938	1.28395	1.22456	1.20266	0.02823	False
176774	4	1.30049	1.13793	1.07651	1.07528	-0.15101	False
177100	4	2.07353	2.17647	0.78655	0.89125	0.13088	False

APPENDIX L—α-PLOTS AND b-PLOTS

Figure L-1. 2014–15 NECAP Science: Grade 4 Plots
Top: *α*-Plot Bottom: *b*-Plot



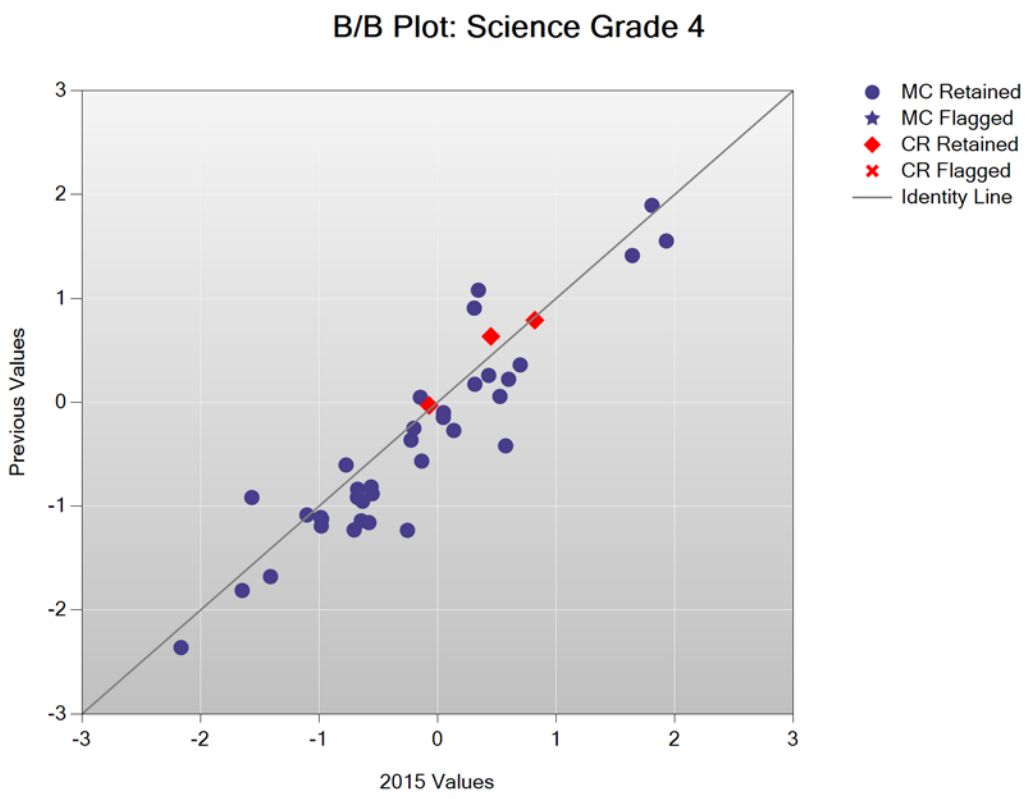
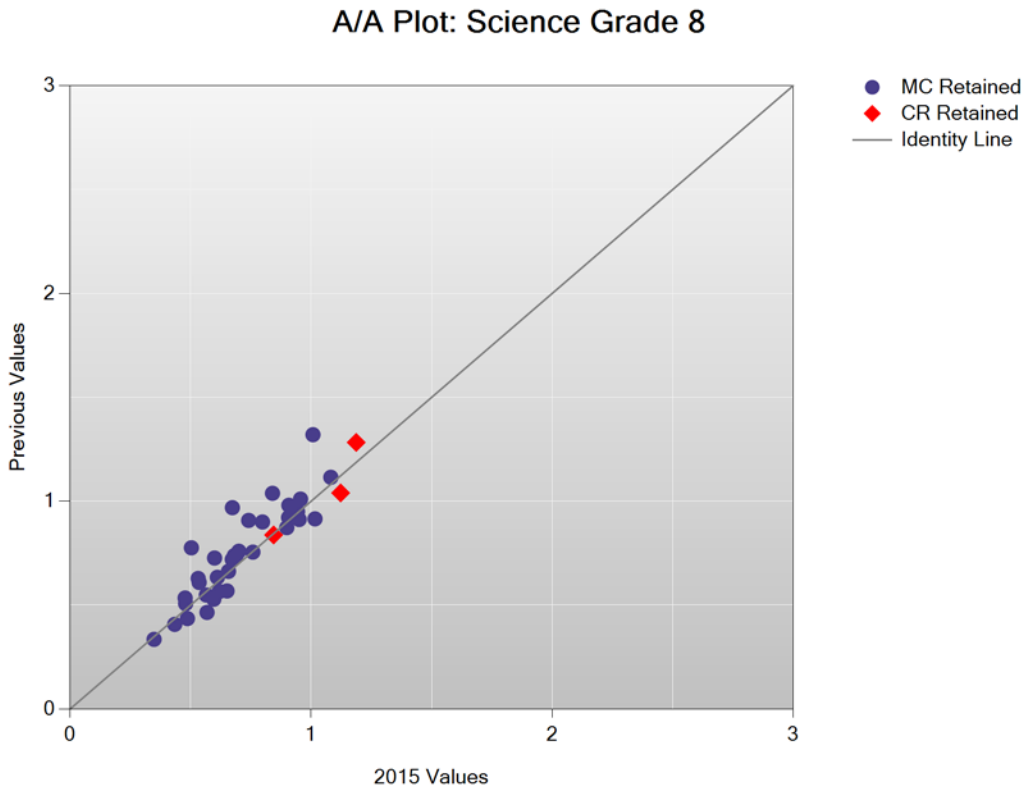


Figure L-2. 2014–15 NECAP Science: Grade 8 Plots
Top: α-Plot Bottom: b-Plot



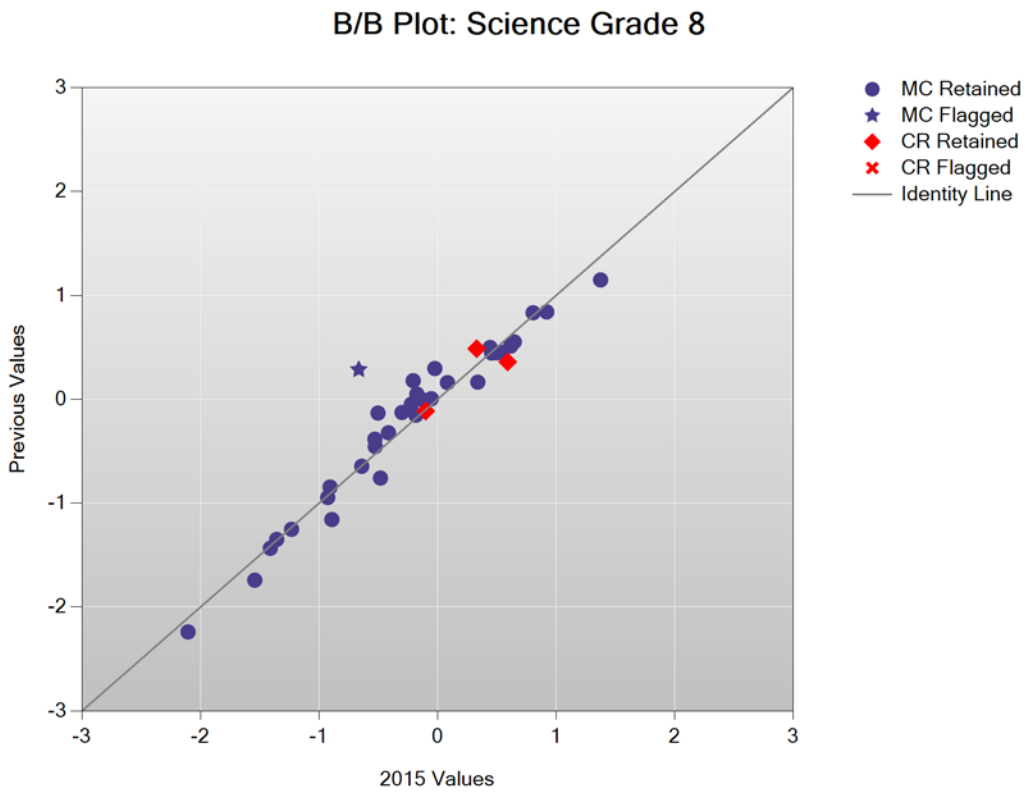
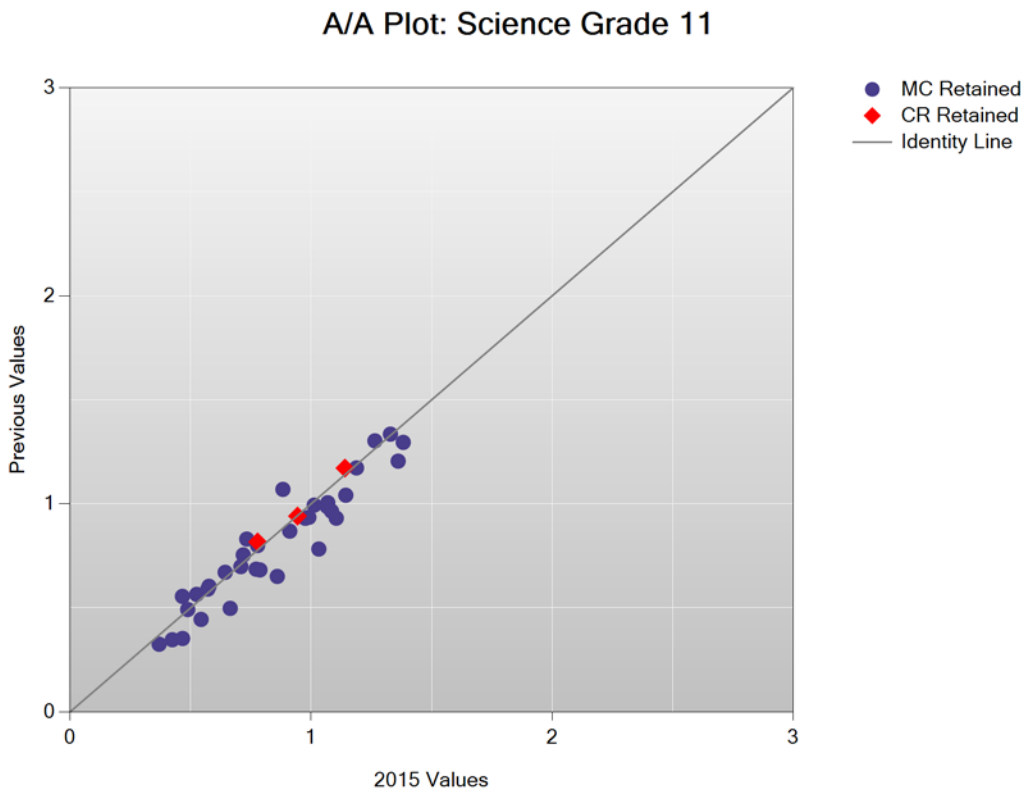
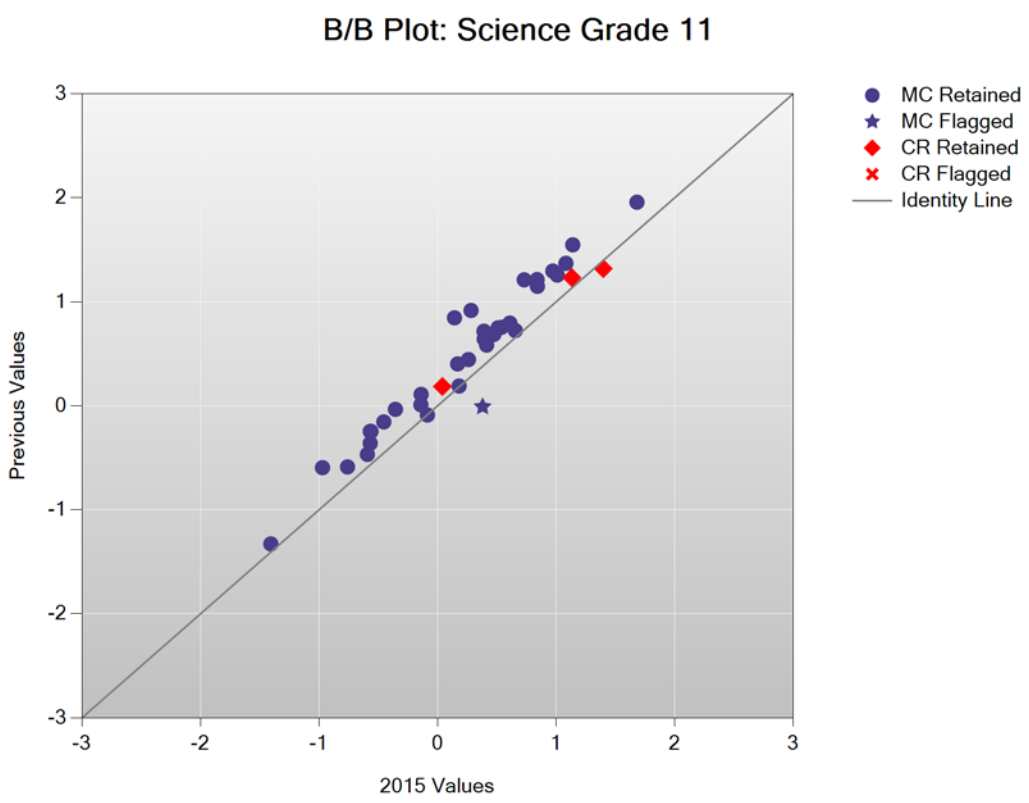


Figure L-3. 2014–15 NECAP Science: Grade 11 Plots
Top: *α*-Plot Bottom: *b*-Plot





APPENDIX M—ACHIEVEMENT LEVEL DISTRIBUTIONS

Table M-1. 2014–15 NECAP Science: Achievement Level Distributions by Grade

Grade	*Performance Level*	*Percent in Level*		
		2014–15	*2013–14*	*2012–13*
4	4	1.18	0.65	0.79
	3	43.84	43.17	45.70
	2	40.17	42.06	39.72
	1	14.81	14.11	13.79
8	4	0.48	0.59	0.60
	3	23.37	23.67	30.27
	2	49.31	52.27	47.45
	1	26.85	23.46	21.68
11	4	1.61	1.70	1.29
	3	30.95	27.99	28.90
	2	41.82	45.97	45.48
	1	25.62	24.34	24.33

APPENDIX N—RAW TO SCALED SCORE LOOKUP TABLES

Table N-1. 2014–15 NECAP Science: Raw Score to Scaled Score Correspondence
Grade 4

Raw Score	2015			2014		
	Scaled Score	Standard Error	Performance Level	Scaled Score	Standard Error	Performance Level
0	400	10.0	1	400	10.0	1
1	400	10.0	1	400	10.0	1
2	400	10.0	1	400	10.0	1
3	400	10.0	1	400	10.0	1
4	400	10.0	1	400	10.0	1
5	400	10.0	1	400	9.9	1
6	400	10.0	1	401	8.2	1
7	400	10.0	1	405	6.7	1
8	400	8.9	1	408	5.8	1
9	404	7.7	1	411	5.2	1
10	407	6.8	1	413	4.7	1
11	410	6.1	1	414	4.4	1
12	412	5.6	1	416	4.2	1
13	414	5.2	1	418	4.0	1
14	416	4.8	1	419	3.8	1
15	417	4.5	1	420	3.7	1
16	419	4.3	1	421	3.6	1
17	420	4.1	1	422	3.5	1
18	421	4.0	1	424	3.4	1
19	422	3.8	1	425	3.3	1
20	423	3.7	1	426	3.3	1
21	425	3.6	1	426	3.3	1
22	426	3.5	1	428	3.2	2
23	426	3.4	1	429	3.2	2
24	428	3.4	2	429	3.2	2
25	428	3.3	2	430	3.2	2
26	429	3.3	2	431	3.2	2
27	430	3.2	2	432	3.1	2
28	431	3.2	2	433	3.1	2
29	432	3.2	2	434	3.2	2
30	433	3.2	2	435	3.2	2
31	434	3.2	2	436	3.2	2
32	435	3.2	2	437	3.2	2
33	436	3.2	2	438	3.2	2
34	437	3.2	2	439	3.2	2
35	438	3.2	2	439	3.3	2
36	439	3.3	2	441	3.3	3
37	439	3.3	2	442	3.4	3
38	441	3.4	3	443	3.4	3
39	442	3.4	3	444	3.5	3
40	443	3.5	3	445	3.5	3
41	444	3.5	3	446	3.6	3
42	445	3.6	3	447	3.6	3
43	446	3.7	3	448	3.7	3
44	447	3.7	3	449	3.8	3

continued

Raw Score	2015			2014		
	Scaled Score	Standard Error	Performance Level	Scaled Score	Standard Error	Performance Level
45	448	3.8	3	451	3.9	3
46	450	3.9	3	452	4.0	3
47	451	4.0	3	453	4.0	3
48	452	4.2	3	455	4.1	3
49	454	4.3	3	456	4.2	3
50	456	4.4	3	458	4.3	3
51	457	4.6	3	459	4.4	3
52	459	4.7	3	461	4.5	3
53	461	4.9	3	462	4.7	3
54	463	5.1	4	465	4.8	4
55	465	5.4	4	467	5.1	4
56	468	5.7	4	469	5.3	4
57	471	6.1	4	472	5.7	4
58	474	6.6	4	475	6.2	4
59	479	7.3	4	478	6.8	4
60	479	7.3	4	479	7.0	4
61	479	7.3	4	479	7.0	4
62	479	7.3	4	479	7.0	4
63	480	7.3	4	480	7.0	4

Table N-2. 2014–15 NECAP Science: Raw Score to Scaled Score Correspondence Grade 8

Raw Score	2015			2014		
	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	806	7.1	1	803	7.9	1
8	810	5.6	1	807	6.0	1
9	812	4.8	1	810	5.1	1
10	814	4.2	1	812	4.4	1
11	816	3.8	1	814	4.0	1
12	818	3.5	1	815	3.7	1
13	819	3.3	1	817	3.5	1
14	820	3.1	1	818	3.3	1
15	821	2.9	1	819	3.2	1
16	822	2.8	1	820	3.1	1
17	823	2.7	1	821	3.0	1
18	824	2.6	1	822	2.9	1
19	825	2.6	1	823	2.9	1
20	825	2.5	1	824	2.8	1
21	826	2.4	1	825	2.7	1

continued

Raw Score	2015			2014		
	Scaled Score	Standard Error	Performance Level	Scaled Score	Standard Error	Performance Level
22	827	2.4	1	826	2.7	1
23	828	2.4	1	827	2.7	1
24	828	2.3	1	827	2.6	1
25	829	2.3	2	828	2.6	1
26	830	2.3	2	829	2.5	2
27	830	2.3	2	830	2.5	2
28	831	2.3	2	830	2.5	2
29	832	2.2	2	831	2.4	2
30	832	2.2	2	832	2.4	2
31	833	2.2	2	832	2.4	2
32	833	2.2	2	833	2.3	2
33	834	2.2	2	834	2.3	2
34	835	2.2	2	834	2.3	2
35	835	2.2	2	835	2.3	2
36	836	2.2	2	836	2.3	2
37	837	2.3	2	836	2.3	2
38	837	2.3	2	837	2.3	2
39	838	2.3	2	838	2.3	2
40	839	2.3	2	838	2.3	2
41	839	2.3	2	839	2.3	2
42	840	2.3	3	839	2.3	2
43	841	2.3	3	841	2.3	3
44	842	2.4	3	841	2.4	3
45	843	2.4	3	842	2.4	3
46	843	2.4	3	843	2.5	3
47	844	2.4	3	844	2.6	3
48	845	2.5	3	845	2.6	3
49	846	2.5	3	846	2.7	3
50	847	2.6	3	847	2.8	3
51	848	2.7	3	848	2.9	3
52	849	2.8	3	849	3.1	3
53	850	2.9	3	850	3.2	3
54	851	3.0	3	852	3.4	3
55	852	3.1	3	853	3.5	3
56	854	3.3	3	854	3.8	3
57	856	3.6	4	856	4.1	4
58	858	3.9	4	859	4.4	4
59	860	4.4	4	861	4.8	4
60	863	5.0	4	864	5.4	4
61	867	6.0	4	867	6.2	4
62	867	6.1	4	867	6.2	4
63	880	6.1	4	880	6.2	4

Table N-3. 2014–15 NECAP Science: Raw Score to Scaled Score Correspondence
Grade 11

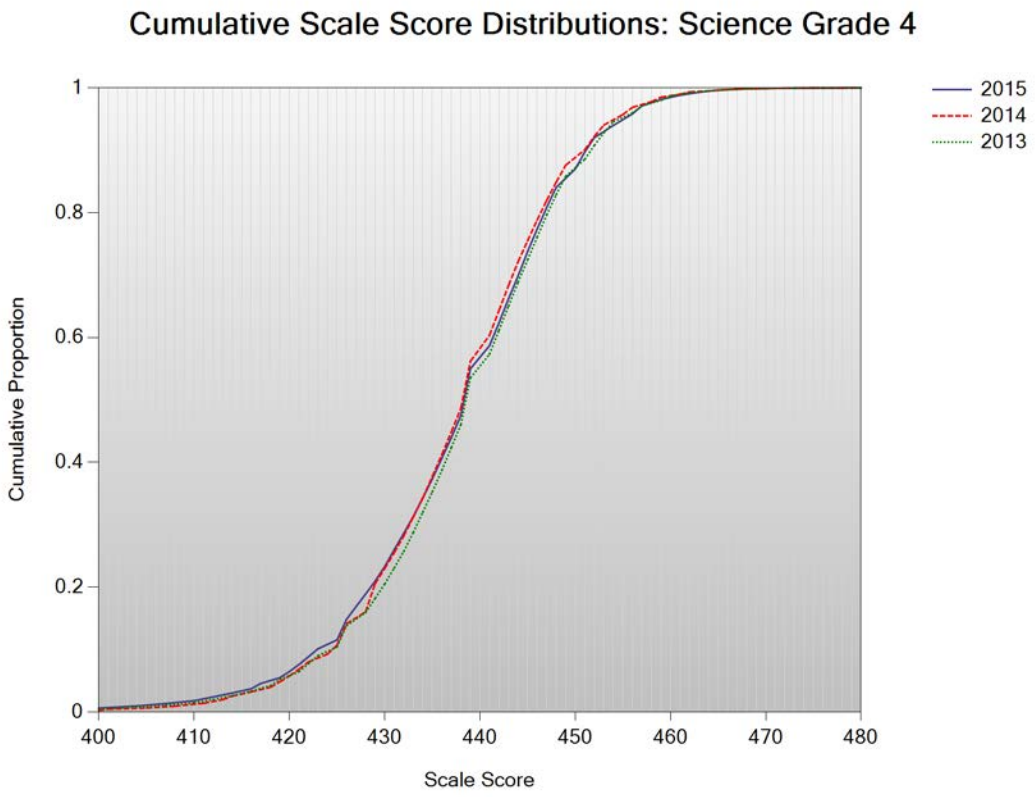
Raw Score	2015			2014		
	Scaled Score	Standard Error	Performance Level	Scaled Score	Standard Error	Performance Level
0	1100	10.0	1	1100	10.0	1
1	1100	10.0	1	1100	10.0	1
2	1100	10.0	1	1100	10.0	1
3	1100	10.0	1	1100	10.0	1
4	1100	10.0	1	1100	10.0	1
5	1100	10.0	1	1100	10.0	1
6	1102	10.0	1	1103	10.0	1
7	1108	7.0	1	1109	6.8	1
8	1112	5.6	1	1113	5.4	1
9	1115	4.8	1	1115	4.6	1
10	1117	4.3	1	1117	4.0	1
11	1119	4.0	1	1119	3.6	1
12	1120	3.7	1	1121	3.4	1
13	1122	3.5	1	1122	3.1	1
14	1123	3.3	1	1123	3.0	1
15	1124	3.1	1	1124	2.8	1
16	1125	3.0	1	1125	2.7	1
17	1126	2.9	1	1126	2.6	1
18	1127	2.8	1	1127	2.6	1
19	1128	2.7	1	1128	2.5	1
20	1129	2.6	1	1128	2.5	1
21	1130	2.6	2	1129	2.4	1
22	1131	2.5	2	1130	2.4	2
23	1132	2.5	2	1131	2.4	2
24	1132	2.5	2	1131	2.4	2
25	1133	2.5	2	1132	2.3	2
26	1134	2.4	2	1133	2.3	2
27	1135	2.4	2	1133	2.3	2
28	1135	2.4	2	1134	2.3	2
29	1136	2.4	2	1135	2.3	2
30	1137	2.4	2	1135	2.3	2
31	1137	2.4	2	1136	2.3	2
32	1138	2.4	2	1137	2.3	2
33	1139	2.4	2	1137	2.3	2
34	1139	2.3	2	1138	2.3	2
35	1140	2.3	3	1139	2.3	2
36	1141	2.3	3	1139	2.3	2
37	1141	2.3	3	1139	2.3	2
38	1142	2.3	3	1140	2.3	3
39	1143	2.3	3	1141	2.3	3
40	1144	2.3	3	1142	2.3	3
41	1144	2.2	3	1142	2.3	3
42	1145	2.2	3	1143	2.3	3
43	1146	2.3	3	1144	2.3	3
44	1146	2.3	3	1145	2.4	3
45	1147	2.3	3	1145	2.4	3

continued

Raw Score	2015			2014		
	Scaled Score	Standard Error	Performance Level	Scaled Score	Standard Error	Performance Level
46	1148	2.3	3	1146	2.4	3
47	1149	2.4	3	1147	2.4	3
48	1149	2.4	3	1148	2.5	3
49	1150	2.5	3	1148	2.5	3
50	1151	2.5	3	1149	2.5	3
51	1152	2.6	4	1150	2.6	3
52	1153	2.7	4	1151	2.7	3
53	1154	2.8	4	1152	2.8	4
54	1155	2.9	4	1153	2.9	4
55	1156	3.0	4	1154	3.0	4
56	1158	3.1	4	1156	3.1	4
57	1159	3.3	4	1157	3.4	4
58	1161	3.5	4	1159	3.6	4
59	1162	3.8	4	1161	4.0	4
60	1165	4.3	4	1163	4.7	4
61	1167	4.9	4	1167	5.7	4
62	1167	4.9	4	1167	5.7	4
63	1180	4.9	4	1180	5.7	4

APPENDIX O—SCALED SCORE DISTRIBUTIONS

Figure O-1. 2014–15 NECAP Science: Cumulative Distribution Plots
Top: Grade 4 Bottom: Grade 8



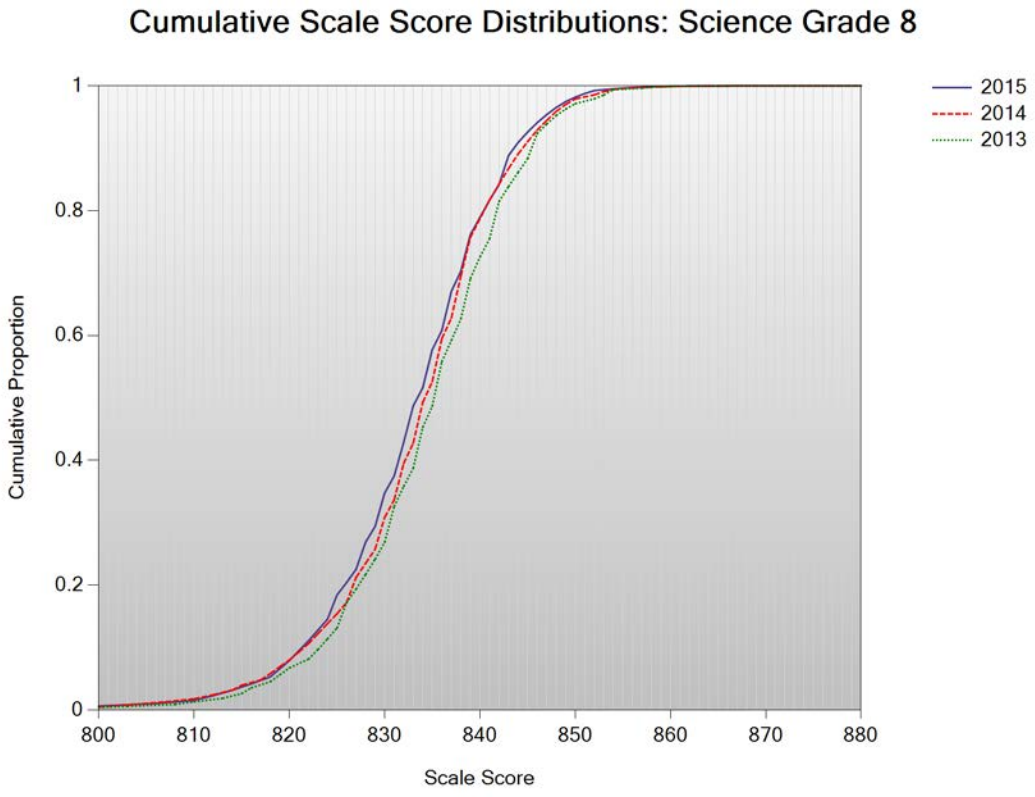
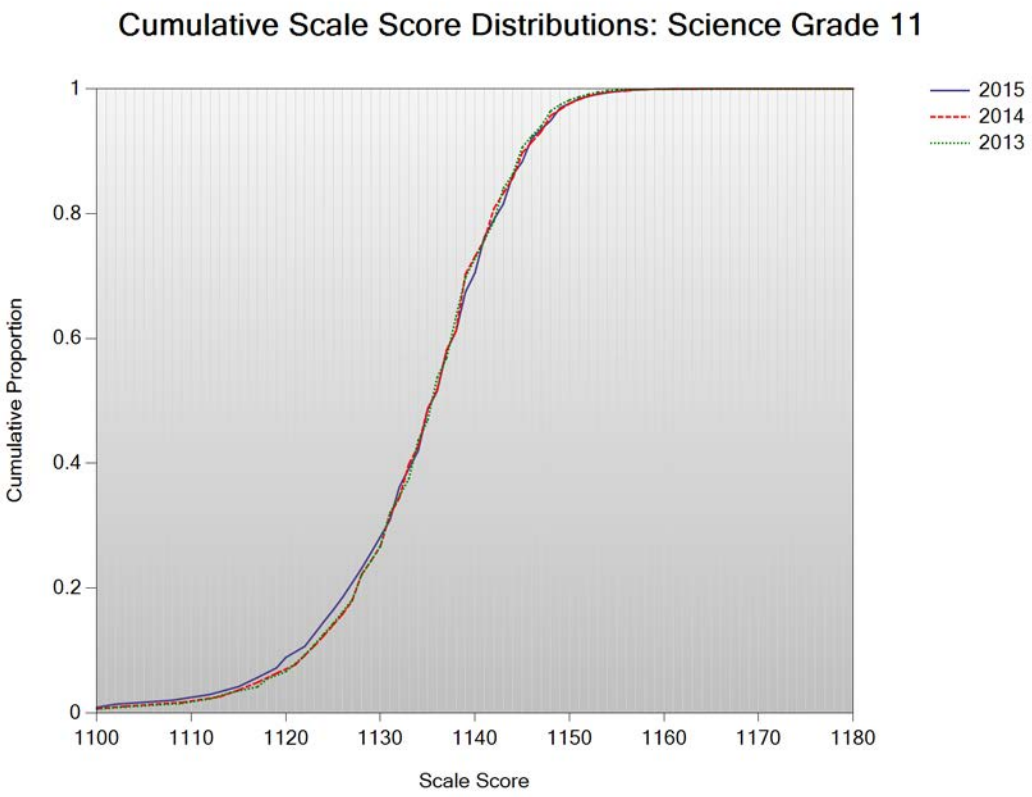


Figure O-2. 2014–15 NECAP Science: Cumulative Distribution Plot—Grade 11



APPENDIX P—CLASSICAL RELIABILITIES

Table P-1. 2014–15 NECAP Science: Subgroup Reliabilities—Grade 4

Description	*Number of Students*	*Raw Score*			*Alpha*	*Standard Error*
		Maximum	*Mean*	*Standard Deviation*		
All Students	29,673	63	35.03	10.32	0.88	3.59
Male	15,290	63	34.35	10.33	0.88	3.56
Female	14,365	63	35.76	10.26	0.88	3.61
Gender Not Reported	18	63	32.33	11.95	0.90	3.70
Hispanic or Latino	3,386	63	28.26	10.59	0.88	3.66
American Indian or Alaskan Native	128	63	30.74	10.79	0.88	3.69
Asian	899	63	36.11	10.76	0.89	3.60
Black or African American	1,219	63	27.87	10.17	0.87	3.65
Native Hawaiian or Pacific Islander	18	63	30.44	10.44	0.88	3.64
White (non-Hispanic)	22,921	63	36.48	9.64	0.86	3.56
Two or More Races (non-Hispanic)	1,071	63	33.44	10.91	0.89	3.60
No Primary race/Ethnicity Reported	31	63	31.61	11.14	0.89	3.67
Currently receiving LEP services	1,297	63	22.92	9.82	0.87	3.58
Former LEP student – monitoring year 1	699	63	31.71	9.20	0.84	3.69
Former LEP student – monitoring year 2	107	63	38.06	8.06	0.79	3.67
LEP: All Other Students	27,570	63	35.68	10.00	0.87	3.58
Students with an IEP	4,411	63	26.62	9.96	0.87	3.63
IEP: All Other Students	25,262	63	36.50	9.66	0.86	3.56
Economically Disadvantaged Students	11,818	63	30.51	10.23	0.87	3.64
SES: All Other Students	17,855	63	38.03	9.23	0.85	3.52
Migrant Students	14	63	28.64	11.42	0.90	3.62
Migrant: All Other Students	29,659	63	35.04	10.32	0.88	3.59
Students receiving Title 1 Services	8,659	63	30.00	10.43	0.88	3.66
Title 1: All Other Students	21,014	63	37.11	9.53	0.86	3.54
Plan 504	183	63	33.98	10.34	0.88	3.54
Plan 504: All Other Students	29,490	63	35.04	10.32	0.88	3.59

Table P-2. 2014–15 NECAP Science: Subgroup Reliabilities—Grade 8

Description	*Number of Students*	*Raw Score*			*Alpha*	*Standard Error*
		Maximum	*Mean*	*Standard Deviation*		
All Students	30,185	63	32.21	11.62	0.91	3.58
Male	15,542	63	31.58	11.96	0.91	3.53
Female	14,626	63	32.88	11.22	0.90	3.61
Gender Not Reported	17	63	30.12	10.75	0.88	3.72
Hispanic or Latino	3,095	63	24.02	11.12	0.90	3.51
American Indian or Alaskan Native	134	63	23.75	11.56	0.91	3.43
Asian	882	63	34.83	12.29	0.92	3.56
Black or African American	1,234	63	23.82	10.55	0.89	3.50
Native Hawaiian or Pacific Islander	31	63	28.00	11.60	0.91	3.52
White (non-Hispanic)	23,938	63	33.71	11.03	0.90	3.57
Two or More Races (non-Hispanic)	840	63	30.84	12.01	0.91	3.56
No Primary race/Ethnicity Reported	31	63	28.48	10.47	0.88	3.63

continued

Description	Number of Students	Raw Score			Alpha	Standard Error
		Maximum	Mean	Standard Deviation		
Currently receiving LEP services	957	63	16.80	8.74	0.86	3.23
Former LEP student – monitoring year 1	73	63	25.77	10.76	0.89	3.56
Former LEP student – monitoring year 2	93	63	24.57	10.22	0.87	3.65
LEP: All Other Students	29,062	63	32.75	11.35	0.90	3.58
Students with an IEP	4,450	63	20.99	9.37	0.87	3.38
IEP: All Other Students	25,735	63	34.15	10.85	0.89	3.57
Economically Disadvantaged Students	10,871	63	26.54	10.97	0.90	3.54
SES: All Other Students	19,314	63	35.40	10.73	0.89	3.55
Migrant Students	8	63				
Migrant: All Other Students	30,177	63	32.21	11.62	0.91	3.58
Students receiving Title 1 Services	3,919	63	23.86	10.85	0.90	3.49
Title 1: All Other Students	26,266	63	33.45	11.21	0.90	3.58
Plan 504	293	63	31.64	10.92	0.89	3.59
Plan 504: All Other Students	29,892	63	32.21	11.63	0.91	3.58

Table P-3. 2014–15 NECAP Science: Subgroup Reliabilities—Grade 11

Description	Number of Students	Raw Score			Alpha	Standard Error
		Maximum	Mean	Standard Deviation		
All Students	28,266	63	28.67	11.21	0.90	3.63
Male	14,218	63	28.16	11.60	0.90	3.58
Female	14,029	63	29.19	10.77	0.88	3.67
Gender Not Reported	19	63	20.74	12.73	0.93	3.29
Hispanic or Latino	2,537	63	21.79	10.19	0.88	3.51
American Indian or Alaskan Native	120	63	21.98	9.57	0.87	3.50
Asian	861	63	30.01	11.61	0.90	3.71
Black or African American	1,112	63	21.67	9.91	0.88	3.49
Native Hawaiian or Pacific Islander	30	63	28.70	10.28	0.87	3.66
White (non-Hispanic)	22,966	63	29.81	10.94	0.89	3.63
Two or More Races (non-Hispanic)	606	63	27.16	11.35	0.90	3.63
No Primary race/Ethnicity Reported	34	63	17.53	11.57	0.92	3.22
Currently receiving LEP services	617	63	14.35	6.42	0.77	3.11
Former LEP student – monitoring year 1	122	63	20.30	7.78	0.81	3.39
Former LEP student – monitoring year 2	202	63	21.59	8.89	0.85	3.46
LEP: All Other Students	27,325	63	29.08	11.08	0.89	3.63
Students with an IEP	3,548	63	18.05	8.41	0.85	3.28
IEP: All Other Students	24,718	63	30.19	10.73	0.88	3.65
Economically Disadvantaged Students	8,424	63	23.43	10.37	0.88	3.53
SES: All Other Students	19,842	63	30.89	10.81	0.89	3.65
Migrant Students	10	63	19.30	13.25	0.93	3.39
Migrant: All Other Students	28,256	63	28.67	11.21	0.90	3.63
Students receiving Title 1 Services	2,290	63	21.35	10.09	0.88	3.49
Title 1: All Other Students	25,976	63	29.31	11.08	0.89	3.63
Plan 504	257	63	28.92	11.63	0.90	3.69
Plan 504: All Other Students	28,009	63	28.67	11.21	0.90	3.63

Table P-4. 2014–15 NECAP Science: Reliabilities by Reporting Category

Grade	*Reporting Category*	*Number of Items*	*Raw Score*			*Alpha*	*Standard Error*
			Maximum	*Mean*	*Standard Deviation*		
4	ESS	12	15	8.69	3.25	0.63	1.98
	INQ	8	18	6.54	3.00	0.68	1.70
	LS	12	15	10.21	2.91	0.67	1.66
	PS	12	15	9.59	3.10	0.68	1.76
8	ESS	12	15	7.96	2.83	0.68	1.60
	INQ	8	18	8.53	4.43	0.83	1.81
	LS	12	15	7.91	3.30	0.71	1.79
	PS	12	15	7.81	2.97	0.67	1.70
11	ESS	12	15	7.09	3.05	0.65	1.79
	INQ	8	18	5.84	3.97	0.78	1.86
	LS	12	15	7.05	3.23	0.66	1.87
	PS	12	15	8.68	2.92	0.71	1.56

APPENDIX Q—INTERRATER CONSISTENCY

Table Q-1. 2014–15 NECAP Science: Item-Level Interrater Consistency Statistics by Grade

Grade	Item Number	Number of		Percent		Correlation	Percent of Third Scores
		Score Categories	Examinee Scores	Exact	Adjacent		
4	141411	5	670	70.60	28.06	0.88	1.34
	263753	3	617	79.58	20.10	0.81	0.32
	263754	3	615	91.06	8.94	0.87	0.00
	263757	4	627	70.97	26.63	0.77	2.39
	263761	3	597	81.57	17.25	0.75	1.17
	263763	3	701	68.19	29.10	0.58	2.28
	263765	3	594	87.88	11.95	0.80	0.17
	263766	3	593	89.71	10.29	0.80	0.00
	263767	4	611	69.56	29.30	0.67	1.15
	50385	5	664	64.01	32.98	0.80	2.86
	99017	5	621	74.56	19.65	0.88	5.48
8	219230	5	599	59.43	38.23	0.73	2.17
	263771	3	631	82.73	16.64	0.77	0.63
	263773	4	614	73.45	24.59	0.83	1.14
	263780	3	628	87.10	12.90	0.89	0.00
	263781	3	627	68.10	28.23	0.71	3.83
	263782	4	602	75.08	23.09	0.87	1.83
	263784	3	610	75.90	22.79	0.74	1.31
	263785	3	609	84.89	12.32	0.78	2.79
	263786	3	604	87.58	12.25	0.69	0.17
	60005	5	601	64.73	33.61	0.59	1.50
	82184	5	584	65.58	28.42	0.80	5.48
11	176772	5	553	76.31	22.24	0.75	1.08
	241745	5	567	63.84	33.33	0.82	2.82
	242637	5	540	79.44	18.52	0.90	2.04
	262508	3	555	66.67	32.61	0.58	0.72
	262515	3	552	71.01	27.90	0.58	0.91
	262516	3	558	82.80	17.20	0.76	0.00
	262517	4	544	85.48	12.68	0.93	1.65
	262523	3	540	74.81	24.44	0.74	0.56
	262524	3	574	67.07	28.75	0.52	4.18
	262527	3	539	69.39	27.27	0.58	3.15
	262529	4	540	75.00	24.07	0.81	0.93

APPENDIX R—DECISION ACCURACY AND CONSISTENCY RESULTS

Table R-1. 2014–15 NECAP Science: Summary of Decision Accuracy (and Consistency) Results by Subject and Grade—Overall and Conditional on Performance Level

Grade	*Overall*	*Kappa*	*Conditional on Level*			
			Substantially Below Proficient	*Partially Proficient*	*Proficient*	*Proficient with Distinction*
4	0.81 (0.73)	0.58	0.79 (0.69)	0.78 (0.72)	0.84 (0.77)	0.80 (0.53)
8	0.83 (0.77)	0.64	0.86 (0.80)	0.83 (0.78)	0.82 (0.73)	0.61 (0.26)
11	0.82 (0.74)	0.61	0.84 (0.78)	0.79 (0.72)	0.84 (0.76)	0.69 (0.37)

Table R-2. 2014–15 NECAP Science: Summary of Decision Accuracy (and Consistency) Results by Subject and Grade—Conditional on Cutpoint

Grade	*Substantially Below Proficient / Partially Proficient*			*Partially Proficient / Proficient*			*Proficient / Proficient with Distinction*		
	Accuracy (Consistency)	*False*		*Accuracy (Consistency)*	*False*		*Accuracy (Consistency)*	*False*	
		Positive	*Negative*		*Positive*	*Negative*		*Positive*	*Negative*
4	0.94 (0.92)	0.03	0.03	0.89 (0.85)	0.07	0.04	0.98 (0.97)	0.02	0
8	0.93 (0.90)	0.04	0.04	0.92 (0.89)	0.05	0.03	0.99 (0.98)	0.01	0
11	0.92 (0.90)	0.04	0.04	0.91 (0.87)	0.06	0.04	0.98 (0.98)	0.01	0

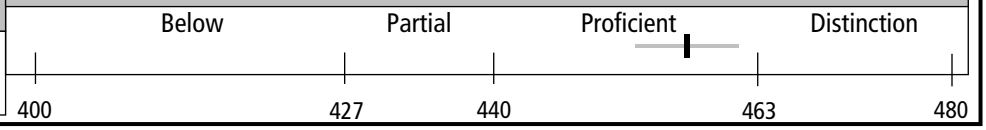
APPENDIX S—SAMPLE REPORTS

Student	Grade	School	District	State
Travis L Arms	4	Demonstration School 1	Demonstration District A	NH

Spring 2015 - Grade 4 NECAP Science Test Results

Achievement Level	Scaled Score
Proficient	457

This Student's Achievement Level and Scaled Score



Interpretation of Graphic Display

The line (**I**) represents the student's score. The bar (▬) surrounding the score represents the probable range of scores for the student if he or she were to be tested many times. This statistic is called the standard error of measurement. See the reverse side for the achievement level descriptions.

This Student's Achievement Level Compared to Other End of Grade 4 Students by School, District, and State

	Student	School	District	State
Proficient with Distinction		0%	0%	1%
Proficient	✓	50%	53%	48%
Partially Proficient		35%	38%	40%
Substantially Below Proficient		15%	9%	11%

This Student's Performance in Science Domains

	Possible Points	Student	Average Points Earned			
			School	District	State	Students at Beginning of Proficient
Physical Science	15	**14**	9.9	10.1	9.9	8.9-12.4
Earth Space Science	15	**13**	9.4	9.3	9.0	7.5-11.5
Life Science	15	**13**	10.1	10.7	10.5	9.4-12.7
Inquiry	18	**11**	6.2	7.0	6.8	5.1-8.5

Description of the Inquiry Task

There are many interesting and essential facts, formulas, and processes that students should know across the three content domains of science. But science is more than content. Inquiry skills are skills that all students should have in addition to the content. Inquiry skills are the ability to formulate questions and hypothesize, plan investigations and experiments, conduct investigations and experiments, and evaluate results. These are the broad areas that constitute scientific inquiry. Content from Physical Science, Earth Space Science, and Life Science forms the basis of each NECAP Science Inquiry Task. Instead of measuring student knowledge of content, inquiry tasks measure the student's ability to make connections, express ideas, and provide evidence of scientific thinking.

The grade 4 inquiry task, *Percolation*, required students to read a story about two students investigating why train tracks can become crooked over time. The students in the story discuss the changes that occur to water bottles and walkways when water freezes into ice. They think about the different earth materials that may be beneath the train tracks, and develop a research question based on their experience with the effects that ice has on other materials. In the inquiry task investigation students use a percolation model to investigate how much water different earth materials can hold. Students worked with partners to complete the task and then answered questions on their own.

About The New England Common Assessment Program

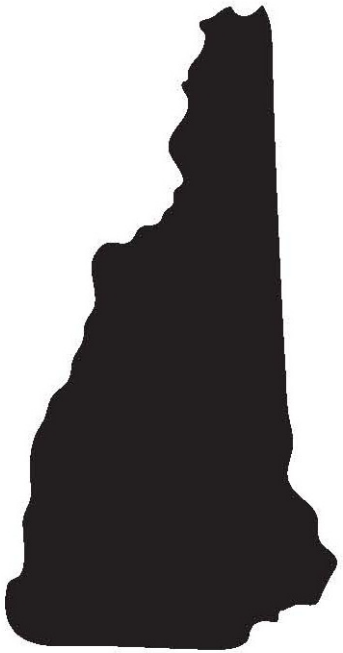
This report highlights results from the Spring 2015 New England Common Assessment Program (NECAP) science tests. The NECAP tests are administered to students in New Hampshire, Rhode Island, and Vermont as part of each state's statewide assessment program. NECAP science test results are used primarily for program evaluation, school improvement and public reporting. Achievement level results are used in the state accountability system required under No Child Left Behind (NCLB). More detailed school and district results are used by schools to help improve curriculum and instruction. Individual student results are used to support information gathered through classroom instruction and assessments.

The NECAP science tests are administered to students in grades 4, 8, and 11.The tests are designed to measure student performance on standards developed and adopted by the three states. Specifically, the tests are designed to measure the content and skills that students are expected to have as they complete the K-4, 5-8, and 9-11 grade spans-in other words, the content and skills that students have learned through the end of the tested grade.

Each test contains a mix of multiple-choice and constructed-response questions. Constructed-response questions require students to develop their own answers to questions. The science test also includes an inquiry session that requires students to answer questions based on results of an actual scientific investigation.

This report contains a variety of school- and/or district-, and state-level assessment results for the NECAP science tests administered at a grade level. Achievement level distributions and mean scaled scores are provided for all students tested as well as for subgroups of students classified by demographics or program participation. The report also contains comparative information on school and district performance on four specific science domains.

In addition to this report of grade level results, schools and districts will also receive Item Analysis Reports, released item support materials, and student-level data files containing NECAP results. Districts will also receive a Summary Report that will show results for all district schools. Together, these reports and data constitute a rich source of information to support local decisions in curriculum, instruction, assessment, and professional development. Over time, this information can also strengthen the school's and district's evaluation of their ongoing improvement efforts.



Spring 2015
Grade 4
NECAP Science Test

District Results

District: Demonstration District A

Code: DEM-DEA

Spring 2015 - Grade 4 NECAP Science Test

Grade Level Summary Report

District: Demonstration District A
State: New Hampshire
Code: DEM-DEA

Schools and districts administered all NECAP tests to every enrolled student with the following exceptions: students who participated in the alternate assessment for the 2014-15 school year, students who withdrew from the school after May 4, 2015, students who enrolled in the school after May 4, 2015, students for whom a special consideration was granted through the state Department of Education, and other students for reasons not approved. On this page, and throughout this report, results are only reported for groups of students that are larger than nine (9).

PARTICIPATION in NECAP	Number			Percentage		
	School	District	State	School	District	State
Students enrolled on or after May 4		**70**	**13,116**		**100**	**100**
	Science			Science		
Students tested		**68**	**12,961**		**97**	**99**
With an approved accommodation		14	2,539		21	20
Current LEP Students		2	318		3	2
With an approved accommodation		1	135		50	42
IEP Students		9	2,078		13	16
With an approved accommodation		7	1,458		78	70
Students not tested in NECAP		**2**	**155**		**3**	**1**
State Approved		2	113		100	73
Alternate Assessment		1	105		50	93
Withdrew After May 4		0	0		0	0
Enrolled After May 4		0	0		0	0
Special Consideration		1	8		50	7
Other		0	42		0	27

NECAP RESULTS

	District													State												
	Enrolled	NT Approved	NT Other	Tested	Level 4		Level 3		Level 2		Level 1		Mean Scaled Score	Tested	Level 4	Level 3	Level 2	Level 1	Mean Scaled Score	Tested	Level 4	Level 3	Level 2	Level 1	Mean Scaled Score
	N	N	N	N	N	%	N	%	N	%	N	%		N	%	%	%	%		N	%	%	%	%	
SCIENCE	70	2	0	68	0	0	36	53	26	38	6	9	440	12,961	1	48	40	11	440						

Level 4 = Proficient with Distinction; Level 3 = Proficient; Level 2 = Partially Proficient; Level 1 = Substantially Below Proficient
Note: Throughout this report, percentages may not total 100 since each percentage is rounded to the nearest whole number.
Note: Some numbers may have been left blank because fewer than ten (10) students were tested.

Spring 2015 - Grade 4 NECAP Science Test

Science Results

District: Demonstration District A
State: New Hampshire
Code: DEM-DEA

Proficient with Distinction (Level 4)
Students performing at this level demonstrate the knowledge and skills as described in the content standards for this grade span. Errors made by these students are few and minor and do not reflect gaps in knowledge and skills.
(Scaled Score 463–480)

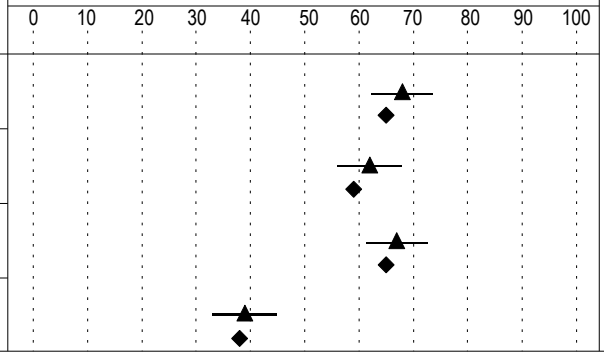
Proficient (Level 3)
Students performing at this level demonstrate the knowledge and skills as described in the content standards for this grade span with only minor gaps. It is likely that any gaps in knowledge and skills demonstrated by these students can be addressed by the classroom teacher during the course of classroom instruction.
(Scaled Score 440–462)

Partially Proficient (Level 2)
Students performing at this level demonstrate gaps in knowledge and skills as described in the content standards for this grade span. Additional instructional support may be necessary for these students to achieve proficiency on the content standards.
(Scaled Score 427–439)

Substantially Below Proficient (Level 1)
Students performing at this level demonstrate extensive and significant gaps in knowledge and skills as described in the content standards for this grade span. Additional instructional support is necessary for these students to achieve proficiency on the content standards.
(Scaled Score 400–426)

	Enrolled	NT Approved	NT Other	Tested	Level 4		Level 3		Level 2		Level 1		Mean Scaled Score
	N	N	N	N	N	%	N	%	N	%	N	%	
School													
2012-13													
2013-14													
2014-15													
Cumulative Total													
District													
2012-13	80	3	1	76	2	3	35	46	29	38	10	13	439
2013-14	80	2	1	77	0	0	37	48	28	36	12	16	437
2014-15	**70**	**2**	**0**	**68**	**0**	**0**	**36**	**53**	**26**	**38**	**6**	**9**	**440**
Cumulative Total	230	7	2	221	2	1	108	49	83	38	28	13	439
State													
2012-13	13,919	165	26	13,728	93	1	6,855	50	5,416	39	1,364	10	440
2013-14	13,948	102	47	13,799	84	1	6,238	45	6,055	44	1,422	10	439
2014-15	**13,116**	**113**	**42**	**12,961**	**154**	**1**	**6,160**	**48**	**5,245**	**40**	**1,402**	**11**	**440**
Cumulative Total	40,983	380	115	40,488	331	1	19,253	48	16,716	41	4,188	10	440

Science Domain	Total Possible Points
Physical Science	31
Earth Space Science	31
Life Science	31
Inquiry	18



Spring 2015 - Grade 4 NECAP Science Test

Disaggregated Science Results

District: Demonstration District A
State: New Hampshire
Code: DEM-DEA

REPORTING CATEGORIES	District													State											
	Enrolled	NT Approved	NT Other	Tested	Level 4		Level 3		Level 2		Level 1		Mean Scaled Score	Tested	Level 4	Level 3	Level 2	Level 1	Mean Scaled Score	Tested	Level 4	Level 3	Level 2	Level 1	Mean Scaled Score
	N	N	N	N	N	%	N	%	N	%	N	%		N	%	%	%	%		N	%	%	%	%	
All Students	**70**	**2**	**0**	**68**	**0**	**0**	**36**	**53**	**26**	**38**	**6**	**9**	**440**	**12,961**	**1**	**48**	**40**	**11**	**440**						
Gender																									
Male	41	1	0	40	0	0	21	53	15	38	4	10	439	6,701	1	46	42	11	439						
Female	29	1	0	28	0	0	15	54	11	39	2	7	442	6,260	2	49	39	10	440						
Not Reported	0	0	0	0										0											
Race/Ethnicity																									
Hispanic or Latino	7	0	0	7										769	<1	26	49	24	433						
Not Hispanic or Latino																									
American Indian or Alaskan Native	1	0	0	1										45	2	33	42	22	437						
Asian	2	0	0	2										443	3	56	32	9	442						
Black or African American	1	0	0	1										249	0	20	48	31	431						
Native Hawaiian or Pacific Islander	1	0	0	1										6											
White	56	2	0	54	0	0	31	57	21	39	2	4	442	11,121	1	49	40	9	440						
Two or more races	2	0	0	2										328	2	47	37	15	439						
No Race/Ethnicity Reported	0	0	0	0										0											
LEP Status																									
Current LEP student	2	0	0	2										318	0	12	46	42	428						
Former LEP student - monitoring year 1	2	0	0	2										262	1	32	54	13	436						
Former LEP student - monitoring year 2	1	0	0	1										32	3	63	31	3	444						
All Other Students	65	2	0	63	0	0	34	54	24	38	5	8	441	12,349	1	49	40	10	440						
IEP																									
Students with an IEP	10	1	0	9										2,078	<1	19	51	30	432						
All Other Students	60	1	0	59	0	0	35	59	20	34	4	7	442	10,883	1	53	38	7	441						
SES																									
Economically Disadvantaged Students	23	1	0	22	0	0	7	32	10	45	5	23	435	3,911	<1	29	50	21	435						
All Other Students	47	1	0	46	0	0	29	63	16	35	1	2	443	9,050	2	55	36	7	442						
Migrant																									
Migrant Students	1	0	0	1										3											
All Other Students	69	2	0	67	0	0	36	54	25	37	6	9	441	12,958	1	48	40	11	440						

Level 4 = Proficient with Distinction; Level 3 = Proficient; Level 2 = Partially Proficient; Level 1 = Substantially Below Proficient

Note: Some numbers may have been left blank because fewer than ten (10) students were tested.

About The New England Common Assessment Program

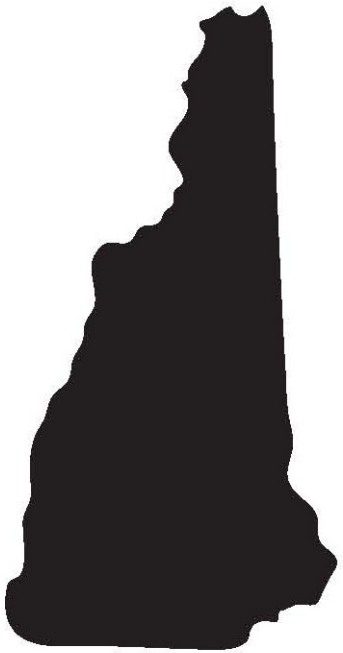
This report highlights results from the Spring 2015 New England Common Assessment Program (NECAP) science tests. The NECAP tests are administered to students in New Hampshire, Rhode Island, and Vermont as part of each state's statewide assessment program. NECAP science test results are used primarily for program evaluation, school improvement and public reporting. Achievement level results are used in the state accountability system required under No Child Left Behind (NCLB). More detailed school and district results are used by schools to help improve curriculum and instruction. Individual student results are used to support information gathered through classroom instruction and assessments.

The NECAP science tests are administered to students in grades 4, 8, and 11.The tests are designed to measure student performance on standards developed and adopted by the three states. Specifically, the tests are designed to measure the content and skills that students are expected to have as they complete the K-4, 5-8, and 9-11 grade spans-in other words, the content and skills that students have learned through the end of the tested grade.

Each test contains a mix of multiple-choice and constructed-response questions. Constructed-response questions require students to develop their own answers to questions. The science test also includes an inquiry session that requires students to answer questions based on results of an actual scientific investigation.

This report contains a variety of school- and/or district-, and state-level assessment results for the NECAP science tests administered at a grade level. Achievement level distributions and mean scaled scores are provided for all students tested as well as for subgroups of students classified by demographics or program participation. The report also contains comparative information on school and district performance on four specific science domains.

In addition to this report of grade level results, schools and districts will also receive Item Analysis Reports, released item support materials, and student-level data files containing NECAP results. Districts will also receive a Summary Report that will show results for all district schools. Together, these reports and data constitute a rich source of information to support local decisions in curriculum, instruction, assessment, and professional development. Over time, this information can also strengthen the school's and district's evaluation of their ongoing improvement efforts.



Spring 2015
Grade 4
NECAP Science Test

School Results

School: Demonstration School 3

District: Demonstration District A

Code: DEM-DEA-DEMO3

Spring 2015 - Grade 4 NECAP Science Test

Grade Level Summary Report

School: Demonstration School 3
District: Demonstration District A
State: New Hampshire
Code: DEM-DEA-DEMO3

Schools and districts administered all NECAP tests to every enrolled student with the following exceptions: students who participated in the alternate assessment for the 2014-15 school year, students who withdrew from the school after May 4, 2015, students who enrolled in the school after May 4, 2015, students for whom a special consideration was granted through the state Department of Education, and other students for reasons not approved. On this page, and throughout this report, results are only reported for groups of students that are larger than nine (9).

PARTICIPATION in NECAP	Number			Percentage		
	School	District	State	School	District	State
Students enrolled on or after May 4	**4**	**70**	**13,116**	**100**	**100**	**100**
	Science			Science		
Students tested	**3**	**68**	**12,961**	**75**	**97**	**99**
With an approved accommodation	1	14	2,539	33	21	20
Current LEP Students	0	2	318	0	3	2
With an approved accommodation	0	1	135		50	42
IEP Students	0	9	2,078	0	13	16
With an approved accommodation	0	7	1,458		78	70
Students not tested in NECAP	**1**	**2**	**155**	**25**	**3**	**1**
State Approved	1	2	113	100	100	73
Alternate Assessment	1	1	105	100	50	93
Withdrew After May 4	0	0	0	0	0	0
Enrolled After May 4	0	0	0	0	0	0
Special Consideration	0	1	8	0	50	7
Other	0	0	42	0	0	27

NECAP RESULTS

	School													District						State					
	Enrolled	NT Approved	NT Other	Tested	Level 4		Level 3		Level 2		Level 1		Mean Scaled Score	Tested	Level 4	Level 3	Level 2	Level 1	Mean Scaled Score	Tested	Level 4	Level 3	Level 2	Level 1	Mean Scaled Score
	N	N	N	N	N	%	N	%	N	%	N	%		N	%	%	%	%		N	%	%	%	%	
SCIENCE	4	1	0	3										68	0	53	38	9	440	12,961	1	48	40	11	440

Level 4 = Proficient with Distinction; Level 3 = Proficient; Level 2 = Partially Proficient; Level 1 = Substantially Below Proficient
Note: Throughout this report, percentages may not total 100 since each percentage is rounded to the nearest whole number.
Note: Some numbers may have been left blank because fewer than ten (10) students were tested.

Spring 2015 - Grade 4 NECAP Science Test

Science Results

School: Demonstration School 3
District: Demonstration District A
State: New Hampshire
Code: DEM-DEA-DEMO3

Proficient with Distinction (Level 4)
Students performing at this level demonstrate the knowledge and skills as described in the content standards for this grade span. Errors made by these students are few and minor and do not reflect gaps in knowledge and skills.
(Scaled Score 463–480)

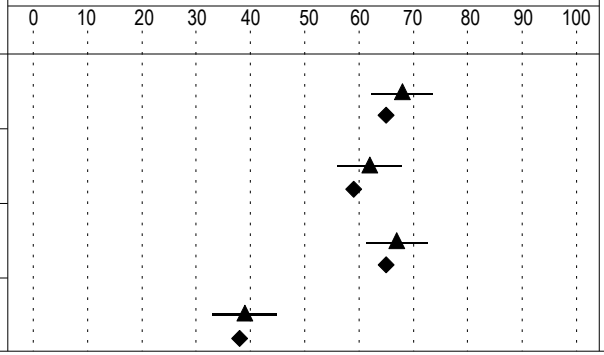
Proficient (Level 3)
Students performing at this level demonstrate the knowledge and skills as described in the content standards for this grade span with only minor gaps. It is likely that any gaps in knowledge and skills demonstrated by these students can be addressed by the classroom teacher during the course of classroom instruction.
(Scaled Score 440–462)

Partially Proficient (Level 2)
Students performing at this level demonstrate gaps in knowledge and skills as described in the content standards for this grade span. Additional instructional support may be necessary for these students to achieve proficiency on the content standards.
(Scaled Score 427–439)

Substantially Below Proficient (Level 1)
Students performing at this level demonstrate extensive and significant gaps in knowledge and skills as described in the content standards for this grade span. Additional instructional support is necessary for these students to achieve proficiency on the content standards.
(Scaled Score 400–426)

	Enrolled	NT Approved	NT Other	Tested	Level 4		Level 3		Level 2		Level 1		Mean Scaled Score
	N	N	N	N	N	%	N	%	N	%	N	%	
School													
2012-13	4	0	0	4									
2013-14	4	0	0	4									
2014-15	**4**	**1**	**0**	**3**									
Cumulative Total													
District													
2012-13	80	3	1	76	2	3	35	46	29	38	10	13	439
2013-14	80	2	1	77	0	0	37	48	28	36	12	16	437
2014-15	**70**	**2**	**0**	**68**	**0**	**0**	**36**	**53**	**26**	**38**	**6**	**9**	**440**
Cumulative Total	230	7	2	221	2	1	108	49	83	38	28	13	439
State													
2012-13	13,919	165	26	13,728	93	1	6,855	50	5,416	39	1,364	10	440
2013-14	13,948	102	47	13,799	84	1	6,238	45	6,055	44	1,422	10	439
2014-15	**13,116**	**113**	**42**	**12,961**	**154**	**1**	**6,160**	**48**	**5,245**	**40**	**1,402**	**11**	**440**
Cumulative Total	40,983	380	115	40,488	331	1	19,253	48	16,716	41	4,188	10	440

Science Domain	Total Possible Points
Physical Science	31
Earth Space Science	31
Life Science	31
Inquiry	18



Spring 2015 - Grade 4 NECAP Science Test

Disaggregated Science Results

School: Demonstration School 3
District: Demonstration District A
State: New Hampshire
Code: DEM-DEA-DEMO3

REPORTING CATEGORIES	School													District						State					
	Enrolled	NT Approved	NT Other	Tested	Level 4		Level 3		Level 2		Level 1		Mean Scaled Score	Tested	Level 4	Level 3	Level 2	Level 1	Mean Scaled Score	Tested	Level 4	Level 3	Level 2	Level 1	Mean Scaled Score
	N	N	N	N	N	%	N	%	N	%	N	%		N	%	%	%	%		N	%	%	%	%	
All Students	**4**	**1**	**0**	**3**										**68**	**0**	**53**	**38**	**9**	**440**	**12,961**	**1**	**48**	**40**	**11**	**440**
Gender																									
Male	3	1	0	2										40	0	53	38	10	439	6,701	1	46	42	11	439
Female	1	0	0	1										28	0	54	39	7	442	6,260	2	49	39	10	440
Not Reported	0	0	0	0										0						0					
Race/Ethnicity																									
Hispanic or Latino	0	0	0	0										7						769	<1	26	49	24	433
Not Hispanic or Latino																									
American Indian or Alaskan Native	1	0	0	1										1						45	2	33	42	22	437
Asian	0	0	0	0										2						443	3	56	32	9	442
Black or African American	0	0	0	0										1						249	0	20	48	31	431
Native Hawaiian or Pacific Islander	0	0	0	0										1						6					
White	3	1	0	2										54	0	57	39	4	442	11,121	1	49	40	9	440
Two or more races	0	0	0	0										2						328	2	47	37	15	439
No Race/Ethnicity Reported	0	0	0	0										0						0					
LEP Status																									
Current LEP student	0	0	0	0										2						318	0	12	46	42	428
Former LEP student - monitoring year 1	0	0	0	0										2						262	1	32	54	13	436
Former LEP student - monitoring year 2	0	0	0	0										1						32	3	63	31	3	444
All Other Students	4	1	0	3										63	0	54	38	8	441	12,349	1	49	40	10	440
IEP																									
Students with an IEP	1	1	0	0										9						2,078	<1	19	51	30	432
All Other Students	3	0	0	3										59	0	59	34	7	442	10,883	1	53	38	7	441
SES																									
Economically Disadvantaged Students	2	1	0	1										22	0	32	45	23	435	3,911	<1	29	50	21	435
All Other Students	2	0	0	2										46	0	63	35	2	443	9,050	2	55	36	7	442
Migrant																									
Migrant Students	0	0	0	0										1						3					
All Other Students	4	1	0	3										67	0	54	37	9	441	12,958	1	48	40	11	440

Level 4 = Proficient with Distinction; Level 3 = Proficient; Level 2 = Partially Proficient; Level 1 = Substantially Below Proficient

Note: Some numbers may have been left blank because fewer than ten (10) students were tested.

APPENDIX T—INTERACTIVE REPORTS

CONFIDENTIAL

Spring 2015 - Grade 08 NECAP Tests

Item Analysis Report - Science

School: Demonstration School 1
District: Demonstration District A
State: New Hampshire
Code: DEA-DEMO1

		Released Inquiry Task								Total Test Results						
	Item Number	1	2	3	4	5	6	7	8	Domain Points Earned				Total Points Earned	Scaled Score	Achievement Level
	Science Domain	INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ	Physical Science	Earth Space Science	Life Science	Inquiry			
	Inquiry Construct	8	8	2	13	8	4	1	13							
	Depth of Knowledge Code	2	2	2	3	2	2	3	3							
	Item Type	SA	CR	SA	SA	CR	SA	SA	SA							
Name/Student ID	Total Possible Points	2	3	2	2	3	2	2	2	**15**	**15**	**15**	**18**	**63**		
Benitez, America V	D0810059	2	1	2	1	1	1	1	1	8	8	11	10	37	837	2
Bennett, Clinton	D0810007	1	0	1	1	2	0	1	1	7	9	9	7	32	833	2
Boyd, Tucker H	D0810048	1	2	1	1	2	0	0	1	8	10	4	8	30	832	2
Branco, Madison L	D0810071	1	2	1	0	2	0	1	1	6	7	8	8	29	832	2
Cady, Dustin F	D0810045	2	2	1	2	1	0	1	0	6	6	8	9	29	832	2
Contreras, Orlando M	D0810011	2	2	2	1	2	1	1	1	15	9	11	12	47	844	3
Cooper, Hannah	D0810018	1	2	2	2	0	1	2	1	6	4	7	11	28	831	2
Dukes, Desean M	D0810080	1	1	1	1	1	0	0	1	5	7	6	6	24	828	1
Enriquez, Luis	D0810077	0	1	0	1	1	0	0	0	5	6	4	3	18	824	1
Ettlinger, Matilda	D0810023	1	0	0	1	0	0	1	1	5	3	8	4	20	825	1
Eustace, Meghan A	D0810036	0	0	0	0	0	0	0	0	7	6	6	0	19	825	1
Frost, Tyler	D0810078	1	2	1	0	0	0	0	1	6	6	5	5	22	827	1
Galarza, Junshuannet	D0810008	2	2	1	2	2	1	0	0	5	5	4	10	24	828	1
Goodwin, Matthew D	D0810032	1	1	1	1	0	0	1	1	3	5	5	6	19	825	1
Harris, Tyler S	D0810054	2	2	1	0	1	0	1	1	9	8	7	8	32	833	2
Herter, Rebecca C	D0810074	0	2	0	1	2	1	2	1	12	10	8	9	39	838	2
Hilton, Nicholas J	D0810070	1	2	2	1	2	1	2	1	11	10	12	12	45	843	3
Landry, Jade	D0810031	0	2	1	1	0	0	1	1	9	9	10	6	34	835	2
Larue, Trai K	D0810020	0	1	1	2	2	1	1	1	11	12	10	9	42	840	3
Masten, Dakota	D0810016									0	0	0	0	0		A
Mehigan, Jessica	D0810055	2	0	1	1	0	0	1	0	4	6	6	5	21	826	1
Michel, Alex F	D0810003	1	1	0	0	1	0	0	1	3	1	2	4	10	814	1
Michel, Alissa	D0810051	2	3	0	2	3	1	2	1	8	11	11	14	44	842	3
Mills, William M	D0810046	2	2	2	2	2	1	1	1	14	12	14	13	53	850	3
Moore, Rebecca	D0810026									8	10	8	0	26	830	2
Morin, Tyler	D0810025	1	0	0	0	0	0	0	1	2	4	2	2	10	814	1
Nickerson, Sarah M	D0810022	1	1	0	1	1	0	1	1	9	9	7	6	31	833	2
Norris, Guiseppe H	D0810027	1	3	2	2	0	2	1		9	8	8	11	36	836	2
Olmstead, Tyler R	D0810052	0	0	0	0	1	0	1	1	6	5	3	3	17	823	1
Palenciaarriola, Diana	D0810064	0	2	1	0	2	0	0	1	8	4	3	6	21	826	1

CONFIDENTIAL

Spring 2015 - Grade 08 NECAP Tests

Item Analysis Report - Science

School: Demonstration School 1
District: Demonstration District A
State: New Hampshire
Code: DEA-DEMO1

		Released Inquiry Task								Total Test Results						
	Item Number	1	2	3	4	5	6	7	8	Domain Points Earned				Total Points Earned	Scaled Score	Achievement Level
	Science Domain	INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ	Physical Science	Earth Space Science	Life Science	Inquiry			
	Inquiry Construct	8	8	2	13	8	4	1	13							
	Depth of Knowledge Code	2	2	2	3	2	2	3	3							
	Item Type	SA	CR	SA	SA	CR	SA	SA	SA							
Name/Student ID	Total Possible Points	2	3	2	2	3	2	2	2	**15**	**15**	**15**	**18**	**63**		
Parker, Malik C	D0810062	1	3	2	1	3	1	1	1	11	11	12	13	47	844	3
Potvin, Jasmine R	D0810075	0	2	2	1	3	2	0	1	11	8	10	11	40	839	2
Powers, Alexandra	D0810065	1	2	1	2	3	1	2	1	7	7	9	13	36	836	2
Prato, Jack P	D0810068	1	2	2	1	2	0	1	1	11	9	8	10	38	837	2
Price, Caleb R	D0810014	1	3	0	1	1	1	0	1	6	7	9	8	30	832	2
Reid, Ie A	D0810029	1	2	0	1	0	0		1	8	8	5	5	26	830	2
Rhoadsdoyle, Jamison	D0810042	1	3	1	2	2	0	0	1	8	8	6	10	32	833	2
Rockwell, Sierra L	D0810019	2	1	1	0	1	0	0	1	6	5	2	6	19	825	1
Rogers, Kyle	D0810056	0	2	1	0	1	2	1	0	9	9	6	7	31	833	2
Smithw, Brittney A	D0810058	1	0	0	0	0	1	0	1	7	9	4	3	23	828	1
Snyder, Geoffrey R	D0810076	1	3	0	1	2	0	1	1	9	10	9	9	37	837	2

Item Number	1	2	3	4	5	6	7	8				
Percent Correct/Average Score: Group	1.0	1.6	0.9	0.9	1.2	0.5	0.7	0.8	7.7	7.5	7.2	7.6
Percent Correct/Average Score: School	1.0	1.6	0.9	0.9	1.2	0.5	0.7	0.8	7.7	7.5	7.2	7.6
Percent Correct/Average Score: District	1.0	1.7	1.0	1.0	1.4	0.6	0.9	0.8	8.3	8.3	8.1	8.5
Percent Correct/Average Score: State	1.0	1.6	1.1	1.1	1.5	0.7	0.9	0.9	8.1	8.3	8.2	8.7

Achievement Level Summary

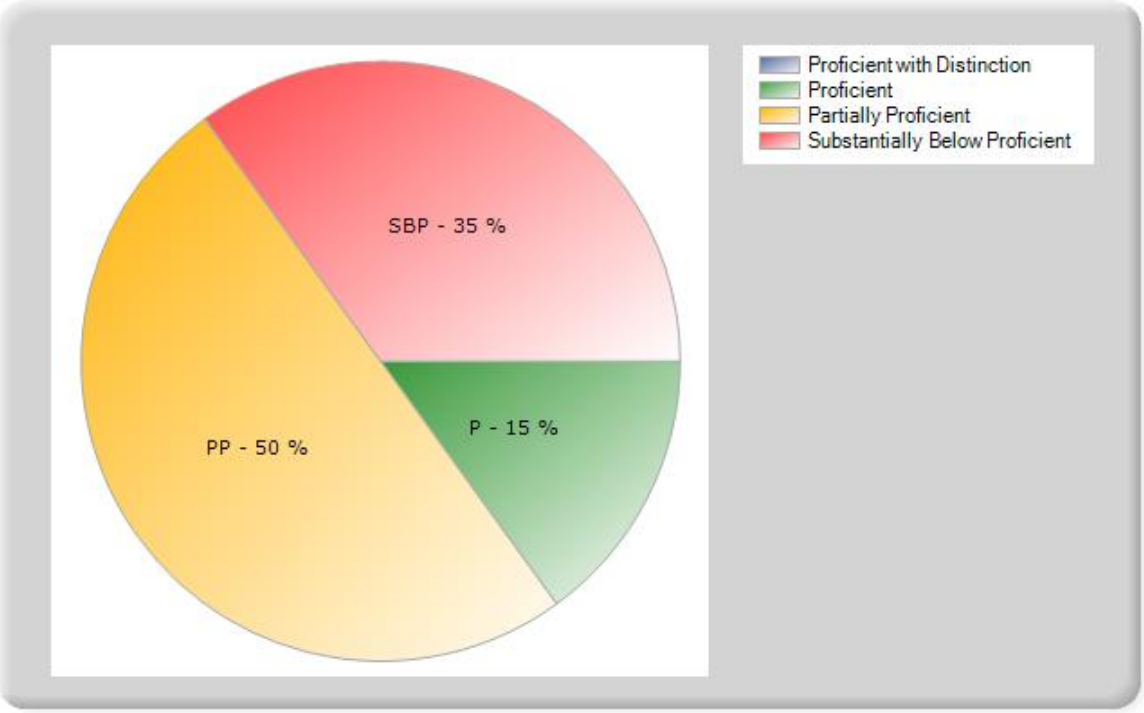
District: Demonstration District A

School: Demonstration School 1

Grade: 08

Date: 11/11/2015 1:26:53 PM

Science



Achievement Level	Count	Percentage %*
Proficient with Distinction	0	0
Proficient	6	15
Partially Proficient	20	50
Substantially Below Proficient	14	35

*Percentages may not total exactly 100% due to applied rounding.

Science Released Items Summary Data

District: Demonstration District A
School: Demonstration School 1
Grade: 08
Date: 11/11/2015 1:27:01 PM

Multiple Choice

Released Item	Domain	Assessment Target	Correct (#)	A (#)	B (#)	C (#)	D (#)	IR (#)	Correct Response
1	PS	1-4	21	15	21	3	1	0	B
2	PS	2-6	21	17	1	1	21	0	D
3	PS	3-8	32	32	5	1	2	0	A
5	ESS	1-3	30	2	4	30	3	1	C
6	ESS	2-6	24	6	4	24	6	0	C
7	ESS	2-8	23	8	23	2	6	1	B
8	LS	1-3	20	6	20	11	3	0	B
9	LS	2-7	15	7	10	8	15	0	D
10	LS	4-10	29	5	29	2	4	0	B

Constructed Response

Released Item	Domain	Assessment Target	Point Value	Average Score
4	PS	2-7	4	1.1

Inquiry Task

Released Item	Domain	Inquiry Construct	Point Value	Average Score
1	INQ	8	2	1.0
2	INQ	8	3	1.6
3	INQ	2	2	0.9
4	INQ	13	2	0.9
5	INQ	8	3	1.2
6	INQ	4	2	0.5
7	INQ	1	2	0.7
8	INQ	13	2	0.8

CONFIDENTIAL

Longitudinal Data Report

Student Name
Drew Bradbury

Year	Enrolled Grade	School Name	Administration	Test Name	Content Area	Score	Achievement Level
1011	04	Demonstration School 1	NECAP Fall 2010	Grade 04 Mathematics	mat	456	Proficient with Distinction
1011	04	Demonstration School 1	NECAP Fall 2010	Grade 04 Reading	rea	455	Proficient
1112	04	Demonstration School 2	NECAP Fall 2011	Grade 04 Mathematics	mat	419	Substantially Below Proficient
1112	04	Demonstration School 2	NECAP Fall 2011	Grade 04 Reading	rea	439	Partially Proficient
1213	04	Demonstration School 1	NECAP Fall 2012	Grade 04 Mathematics	mat	421	Substantially Below Proficient
1213	04	Demonstration School 1	NECAP Fall 2012	Grade 04 Reading	rea	431	Partially Proficient
1314	04	Demonstration School 1	NECAP Fall 2013	Grade 04 Mathematics	mat	446	Proficient
1314	04	Demonstration School 1	NECAP Fall 2013	Grade 04 Reading	rea	434	Partially Proficient

Note: This report returns as many years of NECAP data as are available for this student beginning with 08-09.

APPENDIX U—ANALYSIS AND REPORTING DECISION RULES

Data Analysis and Static Reporting Decision Rules
NECAP
Spring 14-15 Administration

This document specifies rules for data analysis and static reporting requirements. The final student level data set used for analysis and reporting is described in the "Data Processing Specifications." This document is considered a draft until the NECAP State Departments of Education (DOE) signs off. If there are rules that need to be added or modified after said sign-off, DOE sign off will be obtained for each rule. Details of these additions and modifications will be in the Addendum section.

I. General Information

NECAP is administered in the fall and spring. This document incorporates fall and spring rules so that changes are carried to future administrations. In the spring, students are reported based on the spring school/district (referred to as testing school/district). In the spring, students are not reported based on the teaching school. Rules pertaining to the teaching school/district can be ignored for spring administrations. For more information regarding discode, schcode, sprdiscode, sprschcode, senddiscode, and sprsenddiscode, please refer to the data processing specifications and demographic data specification.

This document is the official rules for the current reporting administration.

A. Spring Tests Administered

Grade	Subject	Test items used for Scaling	Item Reporting Categories (Subtopic and Subcategory Source)
04	Science	Common	Cat3
08	Science	Common	Cat3
11	Science	Common	Cat3

B. Reports Produced:

1. Student Report
   a. Testing School District
      I. Parent Copy
      II. School Copy
2. Interactive Reporting (Only the data analysis requirements are outlined in this document)
   a. Item Analysis
   b. Achievement Level Summary
   c. Item Information
   d. Student Longitudinal
3. Grade Level School/District/State Results
   a. Testing School District
4. District/State Summary
   a. Testing School District
5. Writing Prompt CDs

C. Files Produced:

1. Preliminary State Results
2. State Student Released Item Data
3. State Student Raw Data

4. State Student Scored Data
5. District Student Data
6. School Student Data
7. Common Item Information
8. Grade Level Results Report Disaggregated and Historical Data
9. State Standard Deviations and Average Scaled Scores
10. Grade Level Results Report Participation Category Data
11. Grade Level Results Report Subtopic Data
12. Summary Results Data
13. Released Item Percent Responses Data
14. Invalidated Students Original Score
15. Student Questionnaire Summary
16. TCTA Questionnaire Raw Data
17. TCTA Questionnaire Frequency Distribution
18. Scaled Score Lookup
19. Subtopic Average Points Earned (For Program Management)
20. Item Stats for Inquiry Task Items (For Program Management)
21. Memo Shipping files (For Program Management)
22. Released Item Info File (For Program Management)
23. CD print file.

D. School Type*:*

Testing School Type: SchType	**Source: ICORE SubTypeID**	**Description**	**States**
PUB	1,12,13	Public School	ME, NH, RI, VT
CHA	11	Charter School	NH, RI, ME
PSP	19	Public Special Purpose	ME
PSE	15	Public Special Education	ME
INS	7	Institution	VT
OTH	9	Other	VT
OOD	4	Out-of-District Private Providers	NH
OUT	8	Out Placement	RI
PSN	23	Private Special Purpose	ME
BIG	6	Private with >60% Publicly Funded	ME
PRI	3	Private School	RI, VT

School Type Impact on Data Analysis and Reporting				
Level	**Testing**		**Teaching (Fall Only)**	
	Impact on Analysis	**Impact on Reporting**	**Impact on Analysis**	**Impact on Reporting**
Student	n/a	Report students based on testing discode and schcode. District data will be blank for students tested at BIG, PSN, PRI, OOD, OUT, INS, or OTH schools. Always print tested year state data.	n/a	n/a
School	Do not exclude any students based on school type using testing school code for aggregations	Generate a report for each school with at least one student enrolled using the tested school aggregate denominator. District data will be blank for BIG, PSN, PRI, OOD, OUT, INS, or OTH schools. Always print tested year state data.	Exclude students who do not have a teaching school code.	Generate a report for each school with at least one student enrolled using the teaching school aggregate denominator. District data will be blank for BIG, PSN, PRI, OOD, OUT, INS, or OTH schools. Always print tested year state data.
District	For OUT, OOD, BIG, and PSN schools, aggregate using the sending district. If OUT, OOD, BIG, or PSN student does not have a sending district, do not include in aggregations. Do not include students tested at PRI, INS, or OTH schools	Generate a report for each district with at least one student enrolled using the tested district aggregate denominator. Always report tested year state data.	For OUT, OOD, BIG, and PSN teaching schools, aggregate using the spring sending district. If OUT, OOD, BIG, or PSN teaching school student does not have a teaching sending district, do not include in aggregations. Do not include students taught at PRI, INS, or OTH schools	Generate a report for each district with at least one student enrolled using the teaching district aggregate denominator. Always report tested year state data.
State	Do not include students tested at PRI schools for NH and RI. Include all students for VT and ME.	Always report testing year state data.	n/a	n/a

E. Student Status

StuStatus	Description
1	Homeschooled
2	Privately Funded
3	Exchange Student
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 District 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 district names.
School	Exclude all students with a StuStatus value of 1, 2 or 3.	Students with a StuStatus value of 1, 2 or 3 are excluded from Interactive Reporting.
District	Exclude all students with a StuStatus value of 1, 2 or 3.	n/a
State	Exclude all students with a StuStatus value of 1, 2, 3 or 4.	n/a.

F. Requirements To Report Aggregate Data(Minimum N)

Calculation Description	Rule
Number and Percent at each achievement level, mean score by disaggregated category and aggregate level	If the number of tested students included in the denominator is less than 10, then do not report.
Content Area Subcategories Average Points Earned based on common items only by aggregate level	If the number of tested students included in the denominator is less than 10, then do not report.
Aggregate data on Item Analysis report	No required minimum number of students
Number and Percent of students in a participation category by aggregate level	No required minimum number of students
Content Area Subtopic Percent of Total Possible Points and Standard Error Bar and Grade 11 Writing Distribution of Score Points Across Prompts	If any item was not administered to at least one tested student included in the denominator or the number of tested students included in the denominator is less than 10, then do not report
Content Area Cumulative Total Enrollment, Not tested, Tested, Number and Percent at each achievement level, mean score	Suppress all cumulative total data if at least one reported year has fewer than 10 tested students. Spring: The reported years are 1213, 1314, and 1415.

G. Special Forms*:*

1. Form 00 is created for students whose matrix scores will be ignored for analysis. Such students include Braille or administration issues resolved by program management.

H. Other Information

1. NH, RI, and VT participate in NECAP testing for Grades 03-08 and 11. ME only participates in NECAP testing for Grades 03-08.
2. Grade 12 students are allowed to participate in the NECAP Grade 11 test under the following circumstances: RI students trying to improve prior NECAP score, and NH, RI, and VT students taking the NECAP Grade 11 test for the first time.
    a. RI students trying to improve are identified as StuGrade=12 and Grade=11. They only receive a student report. They are not listed on a roster or included in any aggregations. Do not print tested school and district aggregate data on the student report.
    b. For students taking NECAP for the first time the StuGrade in the student demographics file will be 11 and the remaining decision rules apply.
3. Plan504 data not available for NH and VT; therefore 504 Plan section will be suppressed for NH and VT.
4. To calculate Title1 data for writing using Title1rea variable.
5. Title 1 data are not available for VT; therefore Title 1 section will be suppressed for VT.
6. Title 1 Science data are not available for NH; therefore, Title 1 section will be suppressed for NH on Science specific reports. Title 1 Reading and Math data are available for NH and should not be suppressed.
7. Testing level is defined by the variables discode and schcode. Teaching level is defined by the variables sprdiscode and sprschcode. Every student will have testing district and school codes. In the fall, some students will have a teaching school code and some students will have a teaching district code. In the spring, no students will have a teaching school/district.
8. A non-public district code is a district code associated with a school that is type BIG, PSN, PRI, OOD, OUT, INS, or OTH. Non-public testing sending district codes will be ignored. . For example: For RI, senddiscode of 88 is ignored. For NH, senddiscode of 000 is ignored.
9. Only students with a testing school type of OUT, OOD, BIG, or PSN are allowed to have a testing sending district code. Testing sending district codes will be blanked for students at any other testing school types.
10. Only students with a teaching school type of OUT, OOD, BIG, or PSN are allowed to have a spring sending district code. Spring sending district codes will be blanked for students at any other teaching school types.
11. If students have a teaching district code and no teaching school, then ignore teaching district codes that are associated with schools that are BIG, PSN, PRI, OOD, OUT, INS, or OTH.

II. Student Participation / Exclusions

A. Test Attempt Rules by content area

1. A multiple choice item has been answered by a student if the response is A, B, C, D, or * (*=multiple responses)
2. An open response item has been answered if it is not scored blank ‘B’

B. Session Attempt Rules by content area

1. A session was attempted if any multiple choice item or non-field test open response item has been answered in the session. (Use original item responses – see special circumstances section II.F)

C. Not Tested Reasons by content area

1. Not Tested State Approved Alternate Assessment

a. If a student links to the demographic file has content area not tested status of "Not Tested State Approved Alternate Assessment" is identified as "Not Tested State Approved Alternate Assessment" for the content area.

b. If a student is identified as receiving an alternate assessment achievement level, then the student's record will be updated as outlined in the NECAP1415StudentDemographicFileDescription.doc.

2. Not Tested State Approved First Year LEP (reading and writing only)

a. If a student links to the demographic file has content area not tested status of "Not Tested State Approved First Year LEP" or does not link to the demographic file has content area "First Year LEP blank or partially blank reason" marked, then the student is identified as "Not Tested State Approved First Year LEP".

3. Not Tested State Approved Special Consideration

a. If a student links to the demographic data file has content area "Not Tested State Approved Special Consideration" indicated or does not link to the demographic data file and has content area "Special Consideration blank or partially blank reason" marked, then the student is identified as "Not Tested State Approved Special Consideration".

4. Not Tested State Approved Withdrew After

a. If a student links to the demographic data file has content area not tested status of "Not Tested Withdrew After" and at least one content area session was not attempted or does not link to the demographic file has content area "Withdrew After blank or partially blank reason" marked and at least one content area session was not attempted, then the student is identified as "Not Tested State Approved Withdrew After".

5. Not Tested State Approved Enrolled After

a. If a student links to the demographic data file has content area not tested status of "Not Tested Enrolled After" and at least one content area session was not attempted or does not link to the demographic file has content area "Enrolled After blank or partially blank reason" marked and at least one content area session was not attempted, then the student is identified as "Not Tested State Approved Enrolled After".

6. Not Tested Other

a. If content area test was not attempted, the student is identified as "Not Tested Other".

D. Not Tested Reasons Hierarchy by content area: if more than one reason for not testing at a content area is identified then select the first category indicated in the order of the list below.

1. Not Tested State Approved Alternate Assessment
2. Not Tested State Approved Special Consideration
3. Not Tested State Approved Enrolled After
4. Not Tested State Approved Withdrew After
5. Not Tested Other

E. Special Circumstances by content area

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 Measured Progress by the DOE and used in reporting..

a. If sciaccomM1 is marked, then mark sciInvSes3.

b. If sciaccomM3 is marked, then mark sciInvSes1, sciInvSes2, and sciInvSes3.

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. The school and district averages will be suppressed for the impacted subcategories on the student report. 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, achievement level, writing annotations (where applicable). 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" and has sciInvSes3 marked, then all science session 1 item responses will be reported as a blank.

F. Student Participation Status Hierarchy by content area
    1. Not Tested: State Approved Alternate Assessment
    2. Not Tested: State Approved Special Consideration
    3. Not Tested: State Approved Enrolled After
    4. Not Tested: State Approved Withdrew After
    5. Not Tested: Other
    6. Tested Damaged SRB
    7. Tested with Non-Standard Accommodations
    8. Tested Incomplete
    9. Tested

G. Student Participation Summary

Participation Status	Description	Raw Score (*)	Scaled Score (&)	Ach. Level	Student Report Ach. Level Text	Roster Ach. Level Text
Z	Tested Damaged SRB(**)	✓	✓	✓	Substantially Below Proficient, Partially Proficient, Proficient, or Proficient with Distinction	1,2,3, or 4
A	Tested	✓	✓	✓	Substantially Below Proficient, Partially Proficient, Proficient, or Proficient with Distinction	1,2,3, or 4
B	Tested Incomplete(%)	✓	✓	✓	Substantially Below Proficient, Partially Proficient, Proficient, or Proficient with Distinction	1,2,3, or 4
C	Tested with Non-Standard Accommodations (%%)	✓	✓	✓	Substantially Below Proficient, Partially Proficient, Proficient, or Proficient with Distinction	1,2,3, or 4
D	Not Tested State Approved Alternate Assessment	✓			Alternate Assessment	A
E	Not Tested State Approved First Year LEP (Reading and Writing only)	✓			First Year LEP	L
F	Not Tested State Approved Enrolled After	✓			Spring: Enrolled After May [date]	E
G	Not Tested State Approved Withdrew After	✓			Spring: Withdrew After May [date]	W
H	Not Tested State Approved Special Consideration	✓			Special Consideration	S
I	Not Tested Other	✓			Not Tested	N

If a student has a participation status of Alternate Assessment for all subjects assessed at the grade level, a Parent Letter is not produced.

(*) 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.

III. Calculations

A. Rounding

1. All percents are rounded to the nearest whole number
2. All mean scaled scores are rounded to the nearest whole number
3. All mean raw scores are rounded to the nearest tenth.

4. Content Area Subcategories: Average Points Earned (student report): round to the nearest tenth.

5. Round non-multiple choice average item scores to the nearest tenth.

B. 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, subtopic percent of total possible points and standard error, subtopic distribution across writing prompts, 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.

C. Raw scores

1. For all analyses, non-response for an item by a tested student is treated as a score of 0. Items identified as damaged (response of 'X') will be excluded for student identified as "Tested Damaged SRB".

2. Content Area Total Points: Sum the points earned by the student for the common items.

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. If the score is not numeric ('B'), then 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.

5. For all writing prompt scores, the final score of record is the sum of scorer 1 and scorer 2. If both scorers give the student a B, then the final score is B. If both scorers give the student an O or F, then the final score is 0.

E. Scaling

1. Scale Form creation

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.

F. SubTopic Item Scores

1. Identify the Subtopic

a. Spring: NECAP science item information is stored in IABS, including inquiry items.

I. Program management provided Data Analysis with "IABS Export Codes for NECAP SCI Reporting.doc" which contains the crosswalk between IABS item information and reporting.

II. Program management provided Data Analysis with "2010 IABS_Released ItemsSCI for Tara.xls" which contains released item order. Inquiry items are listed at the end in the order they are in the test booklet.

2. Student Content Area Subcategories (student report): Subtopic item scores at the student level is the sum of the points earned by the student for the common items in the subtopic.

3. Content Area Subtopic (grade level results report): Subtopic scores are based on all unique common and matrix items.

a. Percent of Total Possible Points:

I. For each unique common and matrix item calculate the average student score as follows: (sum student item score/number of tested students administered the item).

II. 100 * (Sum the average score for items in the subtopic)/(Total Possible Points for the subtopic) rounded to the nearest whole number.

b. Standard Error Bar: Before multiplying by 100 and rounding the Percent of Total Possible points (ppe) calculate standard error for school, district and state: 100* (square root (((ppe)*(1-ppe)/number of tested students)) rounded to the nearest tenth. For the lower bound and upper bound round the Percent of Total Possible Points +/- Rounded Standard Error to the nearest hundredth.

G. Cumulative Total

1. Include the yearly results where the number tested is greater than or equal to 10

2. Cumulative total N (Enrolled, Not Tested Approved, Not Tested Other, Tested, at each achievement level) is the sum of the yearly results for each category where the number tested is greater than or equal to 10.

3. Cumulative percent for each achievement level is 100*(Number of students at the achievement level cumulative total / number of students tested cumulative total) rounded to the nearest whole number.

4. Cumulative mean scaled score is a weighted average. For years where the number tested is greater than or equal to 10, (sum of (yearly number tested * yearly mean scaled score)) / (sum of yearly number tested) rounded to the nearest whole number.

H. Participation

1. For participation calculate the number and percent of students in each of the following categories by school, district, and state according to schtype and stustatus decision rules.

2. Note that a student is tested with approved accommodations if one is tested, has a non-M accommodation marked, and does not have the M2 or M3 accommodation marked for that subject.

   a. For Students Enrolled, Students Tested, and Students Not Tested the denominator will be the number of students enrolled

   b. For Students Tested with approved Accommodations, Current LEP Students Tested (LEP=1), and IEP Students Tested the denominator will be the number of students tested.

   c. For Current LEP Students Tested with approved accommodations (LEP=1 the denominator will be the number of current LEP students tested.

   d. For IEP Students Tested with approved accommodations the denominator will be the number of IEP students tested.

   e. For Students Not Tested State Approved and Not Tested Other the denominator will be the number of students not tested.

   f. For Students Not Tested Alternate Assessment, First Year LEP, Withdrew After October 1, Enrolled After October 1, and Special Considerations the denominator will be the number of students not tested state approved.

I. Average Points Earned Students at Proficient Level (Range)

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

IV. Report Specific Rules

A. *Student Report*

1. Student header Information

   a. If "FNAME" or "LNAME" is not missing then print "FNAME MI LNAME". Otherwise, print "No Name Provided".

   b. Print the student's single digit tested grade

   c. For school and district name do the following.

      I. For students with a stustatus value of 0 or 4, print the abbreviated tested school and district ICORE name based on school type decision rules.

      II. Otherwise, for the school and district names print the "Description" in the StuStatus table presented earlier in this document.

   d. Print "ME", "NH","RI", or "VT" for state.

2. Test Results by content area

   a. Always display the cut scores in the graphic display.

   a. For students identified as "Not Tested", print the not tested reason in the achievement level, leave scaled score and graphic display blank.

   b. For students identified as tested for the content area then do the following

      I. Print the complete achievement level name the student earned

      II. Print the scaled score the student earned

III. Print a vertical black bar for the student scaled score with gray horizontal bounds in the graphic display

IV. For students identified as "Tested with a non-standard accommodation" for a content area, print '**' after the content area earned achievement level and after student points earned for each subcategory.

V. For students identified as "Tested Incomplete" for a content area, place a section symbol after content area earned scaled score.

3. This Student's Achievement Compared to Other Students by content area

   a. For tested students, print a check mark in the appropriate achievement level in the content area student column. For not tested students leave student column blank

   b. For percent of students with achievement level by school, district, and state print aggregate data based on student status, StuGrade, school type and minimum N rules.

4. This Student's Performance in Content Area Subcategories by content area

   a. Always print total possible points and students at proficient average points earned range.

   b. For students identified as not tested then leave student scores blank

   c. For students identified as tested do the following

      I. Print school, district, and state aggregate data for subcategories based on student status, StuGrade, school type and minimum N rules.

      II. For students identified as "Tested Damaged SRB" do not report student, school, and district aggregate data for subcategories that have at least one damaged item. Print Points Possible and state aggregate data.

      III. Otherwise, always print student subcategory scores

      IV. If the student is identified as tested with a non-standard accommodation for the content area then place '**" after the student points earned for each subcategory.

5. Footer information

   a. Footnotes

      I. If the student received a participation status of "Tested with a non-standard accommodation" for any content area then print "**Student received no credit for parts of the test that were administered under non-standard conditions."

      II. If the student received a participations status of "Tested Incomplete" for any content area then print "§This score should be viewed with caution because the student did not complete all parts of the test."

      III. If both footnotes should appear, the print I. above II.

   b. For NH the SAU, district, and school codes should appear at the bottom right of the page separated by '-'.

   c. For ME, RI, and VT district and school codes should appear at the bottom right of the page separated by '-'.

B. Grade Level School/District/State Results

1. Reports are run by testing state, testing district, testing school using the aggregate school and district codes described in the school type table.

2. Exclude students based on stugrade=12, student status, school type and participation status decision rules for aggregations.

3. The reports will be collated as follows:

a. Page 1 is the Title page.

b. Page 2 is the Participation Results

c. Page 3 is the Historical and Subtopic Results

d. Page 4 is the Disaggregated Results

4. Report Header Information

a. "Spring YYYY Grade XX NECAP Science Tests" where XX is the single digit grade level and YYYY is the year, will print as the title.

b. Teaching level reports will have the following subtitle: "Grade XX-1 Students in (YYYY-1)-(YYYY)".

c. Testing level reports will have the following subtitle: "Grade XX Students in (YYYY)-(YYYY+1)".

d. Use abbreviated school and district name from ICORE based on school type decision rules.

e. Print "Maine", "New Hampshire", "Rhode Island", or "Vermont" to reference the state. The state graphic is printed on the first page.

f. For NH print SAU, district, and school codes separated by '-' for Code on first page for school level. Print SAU and district codes separated by '-'for the district level. Print the full state name for the state level.

g. For ME, RI, and VT print district and school codes separated by '-' for Code on first page for the school level. Print the district code for the district level. Print the full state name for the state level.

5. For achievement level and participation category data if the number of students in an achievement level or participation category does not equal 0, and the percent of students is 0 then format the percent as <1.

6. Report Section: Participation in NECAP

a. For testing level reports always print number and percent based on school type decision rules.

b. For the teaching level reports leave the section blank.

7. Report Section: NECAP Results by content area

a. For the testing level report always print based on minimum N-size and school type decision rules.

b. For the teaching level report leave Enrolled, NT Approved, and NT Other blank. Print Tested, number and percent at each achievement level, mean scaled score based on minimum N-size and school type decision rules.

8. Report Section: Historical NECAP Results by content area

a. For tested level report always print current year, prior years, and cumulative total results based on minimum N-size and school type decision rules.

b. For teaching level report leave Enrolled, NT Approved, and NT Other blank. Print Tested, number and percent at each achievement level, mean scores based on minimum N-size and school type decision rules.

c. Bold current year data.

9. Report Section: Subtopic Results by content area

a. For testing and teaching level reports always print based on minimum N-size and school type decision rules

10. Report Section: Disaggregated Results by content area

    a. For testing level report always print based on minimum N-size and school type decision rules.

    b. For teaching level report leave Enrolled, NT Approved, and NT Other blank. Print Tested, number and percent at each achievement level, mean scores based on minimum N-size and school type decision rules.

C. School/District/State Summary(School Level is run in the Fall Only)

1. Report Header Information

    a. Use abbreviated school and district name from ICORE based on school type decision rules.

    b. Print “Maine”, “New Hampshire”, “Rhode Island”, or “Vermont” to reference the state.

    c. For NH print SAU, district, and school codes separated by ‘-‘ for Code on first page for school level. Print SAU and district codes separated by ‘-‘for the district level. Print the full state name for the state level.

    d. For ME, RI, and VT print district and school codes separated by ‘-‘ for Code on first page for the school level. Print the district code for the district level. Print the full state name for the state level.

2. Reports are run by testing state, testing district, testing school (Fall Only) using the aggregate school and district codes described in the school type table

3. Exclude students based on StuGrade=12, student status, school type and participation status decision rules for aggregations.

4. For achievement level and participation category data if the number of students in an achievement level or participation category does not equal 0, and the percent of students is 0 then format the percent as <1.

5. For testing level report print entire aggregate group across grades tested and list grades tested results based on minimum N-size and school type decision rules. Mean scores across the grades is not calculated.

6. For the teaching level report leave Enrolled, NT Approved, and NT Other blank. Print Tested, number and percent at each achievement level, mean scaled score based on minimum N-size and school type decision rules. Mean scores across the grades is not calculated.

7. Printed Grade Column

    a. For the all grades row, display the school, district, or state name.

    b. For grades 3-8 and 11 rows print Grade X (no leading zero).

V. Data Requirements Interactive Reporting

A. Student Level

1. Refer to Sections II and III. D for decision rules on how student test data will be stored.

2. Students will be loaded into the Interactive System based off of the Interactive flag in tblStuDemo. Students with Interactive flag set to 0 will not be loaded into the system. Students with Interactive set to 1 will be loaded.

    a. Students with StuStatus value of 1, 2 or 3 or RI StuGrade=12 will have the Interactive flag set to 0.

    b. All others will have Interactive=1.

3. The Included flag will determine which students are included in school level aggregations. Students with Included=0 are excluded from all aggregations. Students with Included=2 will be included in Performance Level aggregations and excluded from raw score aggregations (item, subcategory, and total raw score). Students with Included=1 will be included in all school level aggregations.

    a. Students with a Not Tested Participation Status, StuStatus=1, 2, or 3, or RI StuGrade=12 will have their Included flag set to 0.

    b. Students who do fall into the above group and have Participation Status of Tested Damaged SRB will have their Included flag set to 2.

    c. All other students will have their Included flag set to 1.

4. Longitudinal Data

    a. Only students with a valid StudentID and Interactive flag=1 will be loaded.

    b. The complete achievement level name or not tested reason will be stored .

B. Aggregate Level

1. Data Analysis will compute Item Averages for the whole group only at the testing and teaching (Fall only) School and District Levels.

2. Data Analysis will compute Item Averages for all of the filter combinations that exist at the State Level.

3. Data Analysis will create a lookup table with all of the possible filter combinations. It will contain the variable Filter with length 5. Each position represents one of the filter variables. It will contain all the possible combinations of the values plus nulls for when variables are not selected. The first position will be Gender, second Ethnic, third IEP, fourth LEP, and fifth EconDis.

4. Data Analysis will compute Item Averages, Achievement Level Summary, and Item Summary data for the filter combinations for a sample of schools for quality assurance reveiw.

    a. For this sample, percents will be rounded to the nearest whole number and open response average scores will be rounded to the nearest tenth.

    b. For the Item Summary data, item responses other than A, B, C, and D will be counted in the IR column.

VI. Data File Rules

In the file names GR refers to the two digit grade (03-08, 11), YYYY refers to the year, DDDDD refers to the district code, and SS refers to two letter state code. Refer to the tables at the end of this section for filenames and layouts. Teaching level data files will be produced in the Fall Only.

A. Preliminary State Results

1. A PDF file will be created for each state containing preliminary state results for each grade and subject and will list historical state data for comparison.

2. The file name will be SSPreliminaryResultsDATE.pdf

B. State Student Released Item Data

1. A CSV file will be created for each state for grades 3-8 and one for grade 11.

2. One CSV file will be created for each state in the Spring.

3. Accommodation Flags

    a. If the student has at least 1 standard accommodation marked (excluding M) for a given subject then set [sub]STDaccom flag to ‘1’. Otherwise set it to ‘0’.

b. For each group of accommodations (S, T, P, R, and O) if a student has any accommodation in that group marked set [sub]Accom[group]='1'. Otherwise set it to '0'.

c. If a student has the M2 accommodation marked, then set [sub]AccomM2='1'. Otherwise set it to '0'.

d. If a student has the M3 accommodation marked, then set [sub]AccomM3='1'. Otherwise set it to '0'.

4. Exclusion Rules

a. NH: If the student has a testing school type of 'PRI' or StuStatus is 1,2 ,3,or 4 then exclude the student

b. RI: If testing school type is PRI and teaching school type is PRI or blank, then exclude the student.

c. VT: Do not exclude any students

d. ME: If the student has a StuStatus is 1,2 ,3,or 4 then exclude the student

C. State Student Raw Data

1. A CSV file will be created for each state by grade span. The grade spans are 3-4, 5-8, and 11. In the spring, all grades will be combined.

2. If the student has at least 1 standard accommodation marked (excluding M) for a given subject then set [sub]STDaccom flag to '1'. Otherwise set it to '0'.

3. Exclusion Rules

a. NH: If the student has a testing school type of 'PRI' or StuStatus is 1,2, 3,or 4 then exclude the student

b. RI: If testing school type is PRI and teaching school type is PRI or blank, then exclude the student.

c. VT: Do not exclude any students

d. ME: If the student has a StuStatus is 1,2 ,3,or 4 then exclude the student.

D. State Student Scored Data

1. A CSV file will be created for each state including all grades.

2. Exclusion Rules

a. NH: If the student has a testing school type of 'PRI' or StuStatus is 1,2 3,or 4 then exclude the student

b. RI: If testing school type is PRI and teaching school type is PRI or blank, then exclude the student.

c. VT: Do not exclude any students

d. ME: If the student has a StuStatus is 1,2 ,3,or 4 then exclude the student.

E. District Student Data

1. Testing and teaching CSV files will be created for each state and grade and district.

2. Students with the Discode or SendDiscode will be in the district grade specific CSV file for the testing year.

3. For ME, NH, and RI only public school districts will receive district data files. (Districts with at least one school with schoolsubtypeID=1, 11, 19, or 15 in ICORE)

4. Accommodation Flags

a. If the student has at least 1 standard accommodation marked (excluding M) for a given subject then set [sub]STDaccom flag to '1'. Otherwise set it to '0'.

b. For each group of accommodations (S, T, P, R, and O) if a student has any accommodation in that group marked set [sub]Accom[group]='1'. Otherwise set it to '0'.

c. If a student has the M2 accommodation marked, then set [sub]AccomM2='1'. Otherwise set it to '0'.

d. If a student has the M3 accommodation marked, then set [sub]AccomM3='1'. Otherwise set it to '0'.

5. Exclusion Rules

a. NH & RI: If the student has a StuStatus value of 1,2, or 3 then exclude the student

b. VT: If the student has a StuStatus value of 1, then exclude the student.

c. ME: If the student has a StuStatus is 1, 2, or 3 then exclude the student.

F. School Student Data

1. Testing and teaching CSV files will be created for each state and grade and school.

2. Students with the SchCode will be in the school grade specific CSV file for the testing year.

3. Accommodation Flags

a. If the student has at least 1 standard accommodation marked (excluding M) for a given subject then set [sub]STDaccom flag to '1'. Otherwise set it to '0'.

b. For each group of accommodations (S, T, P, R, and O) if a student has any accommodation in that group marked set [sub]Accom[group]='1'. Otherwise set it to '0'.

c. If a student has the M2 accommodation marked, then set [sub]AccomM2='1'. Otherwise set it to '0'.

d. If a student has the M3 accommodation marked, then set [sub]AccomM3='1'. Otherwise set it to '0'.

4. Exclusion Rules

a. NH & RI: If the student has a StuStatus value of 1,2 or 3, then exclude the student

b. VT: If the student has a StuStatus value of 1, then exclude the student.

c. ME: If the student has a StuStatus is 1, 2, or 3 then exclude the student.

G. Common Item Information

1. An excel file will be created containing item information for common items: grade, subject, released item number, item analysis heading data, raw data item name, item type, key, and point value.

H. State Standard Deviations and Averages Scaled Scores

1. A csv file will be created for each state containing the standard deviations and average scale scores for disaggregated subgroups by subject.

2. Exclude students based on state aggregation StuGrade, StuStatus, and SchType decision rules.

3. Data will be suppressed based on minimum N-size and report type decision rules.

4. Average scaled score will be rounded to the nearest whole number. Standard deviations will be rounded to the nearest tenth.

I. Grade Level Results Report Disaggregated and Historical Data

1. Teaching and testing CSV files will be created for each state containing the grade level results disaggregated and historical data.

2. Data will be suppressed based on minimum N-size and report type decision rules.
3. Private schools are excluded from NH & RI files.

J. Grade Level Results Report Participation Category Data

1. Testing CSV file will be created for each state containing the grade level results participation data.
2. Private schools are excluded from NH & RI files.

K. Grade Level Results Report Subtopic Data

1. Teaching and testing CSV files will be created for each state containing the grade level results subtopic.
2. Data will be suppressed based on minimum N-size and report type decision rules.
3. Private schools are excluded from NH & RI files.

L. Summary Results Data

1. Teaching and testing CSV files will be created for each state containing the school, district and state summary data.
2. Data will be suppressed based on minimum N-size and report type decision rules.
3. Private schools are excluded from NH & RI files.

M. Released Item Percent Responses Data

1. The CSV files will only contain state level aggregation for released items.
2. CSV files will be created for each state and grade containing the released item analysis report state data.

N. Invalidated Students Original Score

1. A CSV file will be created for each state including all grades.
2. Original raw scores for students whose responses were invalidated for reporting will be provided.
3. Exclusion Rules
    a. NH: If the student has a testing school type of 'PRI' or StuStatus is 1,2, 3, or 4 then exclude the student
    b. RI: If testing school type is PRI and teaching school type is PRI or blank, then exclude the student.
    c. VT: Do not exclude any students
    d. ME: If the student has a StuStatus is 1,2 ,3,or 4 then exclude the student.

O. Student Questionnaire Summary

1. One CSV file will be created for each state containing percent of students at each response, percent of students at each achievement level, and average scaled score, by student questionnaire response.
2. Only include students who are included in state level aggregations.
3. Data will be suppressed based on minimum N-size and report type decision rules.

P. TCTA Questionnaire Raw Data

1. One CSV file will be created for each state containing raw TC Questionnaire data.
2. One CSV file will be created for each state containing raw TA Questionnaire data.

Q. TCTA Questionnaire Frequency Distribution

1. One CSV file will be created for each state containing the distribution of responses of TC Questionnaire raw data.
2. One CSV file will be created for each state containing the distribution of responses of TA Questionnaire raw data.

R. Scaled Score Lookup

1. One CSV file and one excel file will be created containing the scaled score lookup data.

S. Subtopic Average Points Earned (For Program Management)

1. One excel file will be created containing four worksheets. The first worksheet contains the total possible points for each subtopic as reported on the item analysis report and the range for students who are just proficient. The remaining three worksheets contain state average subtopic scores as reported on the item analysis report.
2. Program management uses this file to create a document which is provided to the schools.

T. Item Stats for Inquiry Task Items (For Program Management)

1. Since Inquiry Task Items are not stored in IABS, one CSV file will be created containing item stats for Inquiry Task items.
2. All three states are included in the calculations.

U. Memo Shipping Files (For Program Management)

1. Provide PM in excel list of schools and districts that tested regardless of grade.

V. CD Print File

A. Spring Table Data File Deliverables

Data File	**Layout**	**File Name**
Preliminary State Results	N/A	Included in Equating Report
State Student Released Item Data	NECAP1415SpringStudentReleasedItemLayout.xls	NECAP1415SpringStateStudentReleasedItem.csv
State Student Raw Data	NECAP1415SpringStateStudentRawLayout.xls	NECAP1415SpringStateStudentRaw.csv
State Student Scored Data	NECAP1415SpringStateStudentScoredLayout.xls	NECAP1415SpringStateStudentScored.csv
District Student Data	NECAP1415SpringStudentReleasedItemLayout.xls	NECAP1415SpringDistrictSlice[GR]_[District Code].csv
School Student Data	NECAP1415SpingStudentReleasedItemLayout.xls	NECAP1415SpringSchoolSlice[GR]_[District Code][School Code].csv
Common Item Information	NECAP1415SpringCommonItemInformationLayout.xls	NECAP1415SpringCommonItemInformation.csv
State Standard Deviations and Average Scaled Scores	NECAP1415SpingStateStandardDeviationsLayout.xls	NECAP1415SpringStateStandardDeviations.csv
Grade Level Results Report Disaggregated and Historical	NECAP1415SpringResultsReport DisaggregatedandHistoricaLayout.xls	NECAP1415SpringResultsReportDisaggregatedandHistorical.csv

Data		
Grade Level Results Report Participation Category Data	NECAP1415SpringResultsReport ParticipationLayout.xls	NECAP1415SpringResultsReportParticipation.csv
Grade Level Results Report Subtopic Data	NECAP1415SpringResultsReport SubtopicLayout.xls	NECAP1415SpringResultsReportSubtopic.csv
Summary Results Data	NECAP1415SpringSummaryResultsLayout.xls	NECAP1415SpringSummaryResults.csv
Released Item Percent Responses Data	NECAP1415SpringReleasedItemPercentResponsesLayout.xls	NECAP1415SpringReleasedItemPercentResponses.csv
Invalidated Students Original Score	NECAP1415SpringStateInvalidatedStudent OriginalScoredLayout.xls	NECAP1415SpringStateInvalidatedStudent OriginalScored.csv
Student Questionnaire Summary	NECAP1415SpringStudentQuestionnaireSummaryLayout.xls	NECAP1415SpringStudentQuestionnaireSummary.csv
TCTA Questionnaire Raw Data	NECAP1415SpringTCQuestionnaireRawLayout.xls NECAP1415SpringTAQuestionnaireRawLayout.xls	NECAP1415SpringTCQuestionnaireRaw.csv NECAP1415SpringTAQuestionnaireRaw.csv
TCTA Questionnaire Frequency Distribution	NECAP1415SpringTCTAQuestionnaireFreqLayout.xls	NECAP1415SpringTCTAQuestionnaireFreq.csv
Scaled Score Lookup	NECAP1415SpringScaleScoreLookupLayout.xls	NECAP1415SpringScaleScoreLookup.xls NECAP1415SpringScaleScoreLookup.csv
Subtopic Average Points Earned (For Project Management)	N/A	NECAP1415SpringSubtopicAvgPointsEarned.xls
Item Stats for Inquiry Task Items (For Program Management)	N/A	NECAP1415SpringInquiryItemStats.csv
Memo Shipping Files (For Program Management)	N/A	TBD