New England Common Assessment Program 2011–12 Technical Report

July 2012

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. TEST DESIGN AND DEVELOPMENT 3
2.1. TEST SPECIFICATIONS 3
2.1.1. Criterion-Referenced Test 3
2.1.2. Item Types 3
2.1.3. Description of Test Design 4
2.2. READING TEST SPECIFICATIONS 4
2.2.1. Standards 4
2.2.2. Item Types 4
2.2.3. Test Design 5
2.2.4. Blueprints 6
2.2.5. Depth of Knowledge 7
2.2.6. Passage Types 8
2.3. MATHEMATICS TEST SPECIFICATIONS 8
2.3.1. Standards 8
2.3.2. Item Types 8
2.3.3. Test Design 9
2.3.4. Blueprints 9
2.3.5. Depth of Knowledge 10
2.3.6. Use of Calculators and Reference Sheets 11
2.4. WRITING TEST SPECIFICATIONS 12
2.4.1. Standards 12
2.4.2. Item Types 12
2.4.3. Test Design 13
2.4.4. Blueprints 13
2.5. TEST DEVELOPMENT PROCESS 15
2.5.1. Item Development 15
2.5.2. Item Reviews at Measured Progress 16
2.5.3. Item Reviews at State Level 17
2.5.4. Bias and Sensitivity Review 18
2.5.5. Reviewing and Refining 18
2.5.6. Item Editing 18
2.5.7. Item Selection and Operational Test Assembly 19
2.5.8. Operational Test Draft Review 20
2.5.9. Alternative Presentations 20
CHAPTER 3. TEST ADMINISTRATION 21
3.1. RESPONSIBILITY FOR ADMINISTRATION 21
3.2. ADMINISTRATION PROCEDURES 21
3.3. PARTICIPATION REQUIREMENTS AND DOCUMENTATION 21
3.4. ADMINISTRATOR TRAINING 22
3.5. DOCUMENTATION OF ACCOMMODATIONS 23
3.6. TEST SECURITY 23
3.7. TEST AND ADMINISTRATION IRREGULARITIES 24
3.8. TEST ADMINISTRATION WINDOW 24
3.9. NECAP SERVICE CENTER 25
CHAPTER 4. SCORING 26

4.1. SCORING OF STANDARD TEST ITEMS 26
4.1.1. Machine-Scored Items 26
4.1.2. Person-Scored Items 26
4.1.2.1. Scoring Location and Staff 27
4.1.2.2. Benchmarking Meetings with the NECAP State Specialists 28
4.1.2.3. Reader Recruitment and Qualifications 29
4.1.2.4. Methodology for Scoring Polytomous Items 29
4.1.2.5. Reader Training 30
4.1.2.6. Senior Quality Assurance Coordinator and Senior Reader Training 32
4.1.2.7. Monitoring of Scoring Quality Control and Consistency 32
CHAPTER 5. CLASSICAL ITEM ANALYSIS 37
5.1. CLASSICAL DIFFICULTY AND DISCRIMINATION INDICES 37
5.2. DIFFERENTIAL ITEM FUNCTIONING 40
5.3. DIMENSIONALITY ANALYSIS 41
CHAPTER 6. ITEM RESPONSE THEORY (IRT) SCALING AND EQUATING 46
6.1. ITEM RESPONSE THEORY (IRT) 48
6.2. ITEM RESPONSE THEORY RESULTS 50
6.3. EQUATING 51
6.4. EQUATING RESULTS 52
6.5. ACHIEVEMENT STANDARDS 53
6.6. REPORTED SCALED SCORES 54
CHAPTER 7. RELIABILITY 57
7.1. RELIABILITY AND STANDARD ERRORS OF MEASUREMENT 58
7.2. 2011–12 SUBGROUP RELIABILITY 59
7.3. REPORTING SUBCATEGORY RELIABILITY 59
7.4. INTERRATER CONSISTENCY 59
7.5. RELIABILITY OF ACHIEVEMENT LEVEL CATEGORIZATION 60
7.5.1. Accuracy and Consistency Results 62
CHAPTER 8. SCORE REPORTING 64
8.1. TEACHING YEAR VERSUS TESTING YEAR REPORTING 64
8.2. PRIMARY REPORTING DELIVERABLES 64
8.3. STUDENT REPORTS 65
8.4. ITEM ANALYSIS REPORTS 66
8.5. SCHOOL AND DISTRICT RESULTS REPORTS 67
8.6. SCHOOL AND DISTRICT SUMMARY REPORTS 71
8.7. SCHOOL AND DISTRICT STUDENT-LEVEL DATA FILES 71
8.8. ANALYSIS & REPORTING SYSTEM 72
8.8.1. Interactive Reports 72
8.8.1.1. Item Analysis Report 72
8.8.1.2. Achievement Level Summary 73
8.8.1.3. Released Items Summary Data 73
8.8.1.4. Longitudinal Data 73
8.8.2. User Accounts 74
8.9. DECISION RULES 74
8.9.1.1. Quality Assurance 74
CHAPTER 9. VALIDITY 76
9.1. QUESTIONNAIRE DATA 77
9.1.1. Difficulty of Assessment 77
9.1.1.1. Difficulty: Reading 77
9.1.1.2. Difficulty: Mathematics 78
9.1.1.3. Difficulty: Writing 79
9.1.2. Content 80

9.1.2.1. Content: Reading *80*
9.1.2.2. Content: Mathematics *82*
9.1.2.3. Content: Writing *84*
9.1.3. Homework 85
9.1.3.1. Homework: Reading *85*
9.1.3.2. Homework: Mathematics *87*
9.1.4. Performance in Courses 89
REFERENCES 91
APPENDICES 93
APPENDIX A COMMITTEE MEMBERSHIP
APPENDIX B PARTICIPATION RATES
APPENDIX C ACCOMMODATION FREQUENCIES BY CONTENT AREA
APPENDIX D TABLE OF STANDARD ACCOMMODATIONS
APPENDIX E ITEM-LEVEL CLASSICAL STATISTICS
APPENDIX F ITEM-LEVEL SCORE POINT DISTRIBUTIONS
APPENDIX G DIFFERENTIAL ITEM FUNCTIONING RESULTS
APPENDIX H ITEM RESPONSE THEORY CALIBRATION RESULTS
APPENDIX I TEST CHARACTERISTIC CURVES AND TEST INFORMATION FUNCTION PLOTS
APPENDIX J DELTA ANALYSES AND RESCORE ANALYSES
APPENDIX K α-PLOTS AND b-PLOTS
APPENDIX L PERFORMANCE LEVEL SCORE DISTRIBUTIONS
APPENDIX M RAW TO SCALED SCORE LOOK-UP TABLES
APPENDIX N SCALED SCORE DISTRIBUTIONS
APPENDIX O CLASSICAL RELIABILITIES
APPENDIX P INTERRATER AGREEMENT
APPENDIX Q DECISION ACCURACY AND CONSISTENCY RESULTS
APPENDIX R SAMPLE REPORTS
APPENDIX S 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 Maine, 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 Act (NCLB). The purposes of the tests are to provide: (a) data on student achievement in reading/English language arts and mathematics to meet the requirements of NCLB; (b) information to support program evaluation and improvement; and (c) information regarding student and school performance to both parents and the public. The tests are constructed to meet rigorous technical criteria, to include universal design elements and accommodations to allow all students access to test content, and to gather reliable student demographic information for accurate reporting. School improvement is supported by:

- providing a transparent test design through the elementary and middle school grade level expectations (GLEs), the high school grade span expectations (GSEs), distributions of emphasis, and practice tests;
- reporting results by GLE/GSE subtopics, released items, and subgroups; and
- hosting report interpretation workshops to foster understanding of results.

It is important to note that the NECAP tests in reading, mathematics, and writing are administered in the fall at the beginning of the school year and test student achievement based on the prior year's GLEs/GSEs. Student level results are provided to schools and families for use as one piece of evidence about progress and learning that occurred on the prior year's GLEs/GSEs. The results are a status report of a student's performance against GLEs/GSEs and 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 2011–12 NECAP tests. In October 2011, students in grades 3 through 8 and 11 participated in the administration of the NECAP tests in reading and mathematics. Students in grades 5, 8, and 11 also participated in writing. This report provides information about the technical quality of those tests, including a description of the processes used to develop, administer, and score the tests and to analyze the test results. This report is intended to serve as a guide for replicating and/or improving the procedures in subsequent years.

Though some parts of this technical report may be used by educated laypersons, the intended audience is 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, knowledge on more advanced topics is required.

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 specification and addresses all the intermediate steps that lead to final score reporting. Chapters 2 through 4 provide a description of the NECAP tests by covering the test design and development process, the administration of the tests, and scoring. Chapters 5 through 7 provide statistical and psychometric summaries, including chapters on item analysis, scaling and equating, and reliability. Chapter 8 is devoted to NECAP score reporting, and Chapter 9 is devoted to discussions on validity. Finally, the references cited throughout the report are provided, followed by the report appendices.

CHAPTER 2. TEST DESIGN AND DEVELOPMENT

2.1. TEST SPECIFICATIONS

2.1.1. Criterion-Referenced Test

Items on the NECAP tests are developed specifically for those states participating in the NECAP and are directly linked to the NECAP Grade Level Expectations/Grade Span Expectations. These GLEs/GSEs are the basis for the reporting categories developed for each content area and are used to help guide the development of test items. Although each item is designed to measure a specific GLE/GSE, an item may address several GLEs/GSEs within a strand.

2.1.2. Item Types

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

Multiple-choice items were administered in grades 3 through 8 and 11 in reading and mathematics and in grades 5 and 8 in writing to provide breadth of coverage of the GLEs/GSEs. Because each requires 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, including, for example, word identification and vocabulary skills.

Short-answer items were administered in grades 3 through 8 and 11 in mathematics to assess students' skills and their ability to work with brief, well-structured problems with one solution or a very limited number of solutions. Each short-answer item requires approximately two to five minutes for most students to answer. The advantage of this item type is that it requires students to demonstrate knowledge and skills by generating, rather than merely selecting, an answer.

Constructed-response items typically require students to use higher-order thinking skills such as summary, evaluation, and analysis in constructing a satisfactory response. Each constructed-response item requires approximately 5 to 10 minutes for most students to complete. These items were administered in grades 3 through 8 and 11 in reading, in grades 5 through 8 and 11 in mathematics and in grades 5 and 8 in writing.

A common **writing prompt** was administered in grades 5, 8, and 11. In grade 11, an additional matrix prompt was administered to students. Students were given 45 minutes (plus additional time if necessary) to compose an extended response for the common prompt that was scored by two independent readers both on quality of the stylistic and rhetorical aspects of the writing and on the use of standard English conventions.

Approximately 25% of the common NECAP items were released to the public in 2011–12. The released NECAP items are posted on a Web site hosted by Measured Progress and on each Department of

Education Web site. Schools are encouraged to incorporate the use of released items in their instructional activities so that students will be familiar with the types of questions found on the NECAP tests.

2.1.3. Description of Test Design

The NECAP test is structured using both *common* and *matrix* items. Common items are taken by all students in a given grade level. Student scores are based only on common items. Matrix items are either new items included on the test for field-test purposes or equating items used to link one year's results to those of previous years. In addition, field-test and equating items are divided among the multiple forms of the test for each grade and content. The number of test forms varies by grade and content area but ranges between eight and nine forms. Each student takes only one form of the test and therefore answers a fraction of the field-test items. Equating and field-test items are not distinguishable to students and have a negligible impact on testing time. Because all students participate in the field test, an adequate sample size is provided to produce reliable data that can be used to inform item selection for future tests.

2.2. READING TEST SPECIFICATIONS

2.2.1. Standards

The test framework for reading in grades 3 through 8 was based on the NECAP GLEs, and each item on the NECAP tests was designed to measure a specific GLE. The test framework for reading in grade 11 was based on the NECAP GSEs, and each item on the NECAP test was designed to measure a specific GSE.

Reading comprehension is assessed on the NECAP test by items that are dually categorized by the type of text and by the level of comprehension measured. The level of comprehension is designated as either "Initial Understanding" or "Analysis and Interpretation." Word identification and vocabulary skills are assessed at each grade level primarily through multiple-choice items.

2.2.2. Item Types

The NECAP reading tests include multiple-choice and constructed-response items. Multiple-choice items require students to demonstrate a wide range of knowledge and skills, requiring one minute of response time per item. Constructed-response items are more complex, requiring 5 to 10 minutes of response time per item. Each type of item is worth a specific number of points in the student's total reading score, as shown in Table 2-1.

Table 2-1. 2011–12 NECAP: Reading Item Types

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

MC = multiple-choice; CR = constructed-response

2.2.3. Test Design

Table 2-2 summarizes the numbers and types of items that were used in the 2011–12 NECAP reading tests for grades 3 through 8. Note that, in reading, all students received the common items and one of either the equating or field-test forms. Each multiple-choice item was worth one point and each constructed-response item was worth four points.

Table 2-2. 2011–12 NECAP: Item Type and Number of Items—Reading Grades 3–8

	Long passages	*Short passages*	*Stand-alone MC*	*Total MC*	*Total CR*
Common	2	2	4	28	6
Matrix—Equating					
Forms 1–3	1	1	2	14	3
Matrix—FT					
Forms 4–7	1	1	2	14	3
Forms 8–9	1	1	2	14	3
Total per Student					
Forms 1–7	3	3	6	42	9
Forms 8–9	3	3	6	42	9

Long passages have 8 MC and 2 CR items; short passages have 4 MC and 1 CR items.
MC = multiple-choice; CR = constructed-response; FT = field test

Table 2-3 summarizes the numbers and types of items that were used in the 2011–12 NECAP reading test for grade 11. Note that, in reading, all students received the common items and one of either the equating or field-test forms. Each multiple-choice item was worth one point and each constructed-response item was worth four points.

Table 2-3. 2011–12 NECAP: Item Type and Number of Items—Reading Grade 11

	Long passages	*Short passages*	*Stand-alone MC*	*Total MC*	*Total CR*
Common	2	2	4	28	6
Matrix—Equating					
Forms 1–2	1	1	2	14	3
Matrix—FT					
Forms 3–8	1	1	2	14	3
Total per Student	3	3	6	42	9

Long passages have 8 MC and 2 CR items; short passages have 4 MC and 1 CR items;
MC = multiple-choice; CR = constructed-response; FT = field test

2.2.4. Blueprints

The distribution of emphasis for reading is shown in Table 2-4.

Table 2-4. 2011–12 NECAP: Distribution of Emphasis Across Reporting Subcategories in Terms of Targeted Percentage of Test by Grade—Reading Grades 3–8 and 11

Subcategory	*GLE/GSE grade (grade tested)*						
	2 (3)	*3 (4)*	*4 (5)*	*5 (6)*	*6 (7)*	*7 (8)*	*9–10 (11)*
Word Identification Skills and Strategies	20%	15%	0%	0%	0%	0%	0%
Vocabulary Strategies/Breadth of Vocabulary	20%	20%	20%	20%	20%	20%	20%
Initial Understanding of Literary Text	20%	20%	20%	20%	15%	15%	15%
Initial Understanding of Informational Text	20%	20%	20%	20%	20%	20%	20%
Analysis and Interpretation of Literary Text	10%	15%	20%	20%	25%	25%	25%
Analysis and Interpretation of Informational Text	10%	10%	20%	20%	20%	20%	20%
TOTAL	100%	100%	100%	100%	100%	100%	100%

Table 2-5 shows the content area subcategory reporting structure for reading and the maximum possible number of raw score points that students could earn. (With the exception of word identification/vocabulary items, reading items were reported in two ways: type of text and level of comprehension.) Note: because only common items are counted toward students' scaled scores, only common items are reflected in this table.

Table 2-5. 2011–12 NECAP: Reporting Subcategories and Possible Raw Score Points by Grade—Reading Grades 3–8 and 11

Subcategory	*Grade tested*						
	3	*4*	*5*	*6*	*7*	*8*	*11*
Word ID/Vocabulary	19	17	10	10	10	10	10
Type of Text							
Literary	16	18	21	21	21	21	21
Informational	17	17	21	21	21	21	21
Level of Comprehension							
Initial Understanding	21	20	19	23	18	18	17
Analysis and Interpretation	12	15	23	19	24	24	25
TOTAL	52	52	52	52	52	52	52

The total possible points in reading equals the sum of Word ID/Vocabulary points and the total points from either Type of Text or Level of Comprehension (since reading comprehension items are dually categorized by type of text and level of comprehension).

2.2.5. Depth of Knowledge

Each item on the NECAP tests in reading is assigned a DOK level according to the cognitive demand of the item. DOK is not synonymous with difficulty. The DOK level rates the complexity of the mental processing a student must use to answer the question. Each of the three levels is described in Figure 2-1.

Figure 2-1. 2011-12 NECAP: Depth of Knowledge—Reading

Level 1 (Recall)	This level requires students to receive or recite facts or to use simple skills or abilities. Oral reading that does not include analysis of the text as well as basic comprehension of a text is included. Items require only a shallow understanding of text presented and often consist of verbatim recall from text or simple understanding of a single word or phrase.
Level 2 (Skill/Concept)	This level includes the engagement of some mental processing beyond recalling or reproducing a response; it requires both comprehension and subsequent processing of text or portions of text. Intersentence analysis of inference is required. Some important concepts are covered but not in a complex way.
Level 3 (Strategic Thinking)	This level requires students to go beyond the text; however, they are still required to show understanding of the ideas in the text. Students may be encouraged to explain, generalize, or connect ideas. Standards and items involve reasoning and planning. Students must be able to support their thinking. Items may involve abstract theme identification, inference across an entire passage, or application of prior knowledge. Items may also involve more superficial connections between texts.

Table 2-6 lists the percentage of actual score points assigned to each depth-of-knowledge (DOK) level in reading.

Table 2-6. 2011–12 NECAP: Depth of Knowledge in Terms of Percentage of Test by Grade—Reading Grades 3–8 and 11

DOK	*Grade*						
	3	*4*	*5*	*6*	*7*	*8*	*11*
Level 1	58%	40%	15%	12%	19%	17%	23%
Level 2	42%	60%	77%	88%	81%	75%	69%
Level 3	0%	0%	8%	0%	0%	8%	8%
TOTAL	100%	100%	100%	100%	100%	100%	100%

2.2.6. Passage Types

The reading passages on all the NECAP tests are broken down into the following categories:

- Literary passages, representing a variety of forms: modern narratives; diary entries; dramas; poetry; biographies; essays; excerpts from novels; short stories; and traditional narratives, such as fables, tall tales, myths, and folktales
- Informational passages/factual text, often dealing with areas of science and social studies. These passages are taken from such sources as newspapers, magazines, and book excerpts. Informational text could also be directions, manuals, recipes, etc. The passages are authentic texts selected from grade level appropriate reading sources that students would be likely to encounter in both classroom and independent reading.
- All passages are collected from published works.

2.3. MATHEMATICS TEST SPECIFICATIONS

2.3.1. Standards

The test framework for mathematics at grades 3 through 8 was based on the NECAP GLEs, and each item on the grades 3 through 8 NECAP tests was designed to measure a specific GLE. The test framework for mathematics at grade 11 was based on the NECAP GSEs, and each item on the grade 11 NECAP test was designed to measure a specific GSE. The mathematics items are organized into four content strands:

- Numbers and Operations: Students understand and demonstrate a sense of what numbers mean and how they are used. Students understand and demonstrate computation skills.
- Geometry and Measurement: Students understand and apply concepts from geometry. Students understand and demonstrate measurement skills.
- Functions and Algebra: Students understand that mathematics is the science of patterns, relationships, and functions. Students understand and apply algebraic concepts.
- Data, Statistics, and Probability: Students understand and apply concepts of data analysis. Students understand and apply concepts of probability.

Additionally, problem solving, reasoning, connections, and communication are embedded throughout the GLEs/GSEs.

2.3.2. Item Types

The NECAP mathematics tests include multiple-choice, short-answer, and constructed-response items. Short-answer items require students to perform a computation or solve a problem, requiring two to five minutes of response time per item. Constructed-response items are more complex, requiring 8 to 10 minutes

of response time per item. Each type of item is worth a specific number of points in the student's total mathematics score, as shown in Table 2-7.

Table 2-7. 2011–12 NECAP: Mathematics Item Types

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

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

2.3.3. Test Design

Table 2-8 summarizes the numbers and types of items that were used in the 2011–12 NECAP mathematics tests for grades 3 and 4, 5 through 8, and 11, respectively. Note that all students received the common items plus equating and field-test items in their forms. Each multiple-choice item was worth one point, each short-answer item was worth either one or two points, and each constructed-response item was worth four points. Except for grade 11, score points within a grade level were evenly divided so that multiple-choice items represented approximately 50% of the possible score points and short-answer and constructed-response items together represented approximately 50% of the possible score points. In grade 11, multiple-choice items represented approximately 40% of the possible score points and short-answer and constructed-response represented approximately 60% of the possible score points.

Table 2-8. 2011–12 NECAP: Item Type and Number of Items—Mathematics

Content Area and Grade	Common				Matrix–equating				Matrix–FT				Total per student			
	MC	SA1	SA2	CR	MC	SA1	SA2	CR	MC	SA1	SA2	CR	MC	SA1	SA2	CR
Mathematics 3–4	35	10	10		6	2	2		3	1	1		44	13	13	
Mathematics 5–8	32	6	6	4	6	2	2	1	3	1	1	1	41	9	9	6
Mathematics 11	24	12	6	4	4	2	1	1	4	2	1	1	32	16	8	6

MC = multiple-choice; SA1 = 1-point short-answer; SA2 = 2-point short-answer; FT = field-test
For grades 3–4 and 5–8, total of nine forms per grade; six contained unique matrix-equating items.
For grade 11, total of eight forms; six contained unique matrix-equating items.

2.3.4. Blueprints

The distribution of emphasis for NECAP content strands for mathematics is shown in Table 2-9.

Table 2-9. 2011–12 NECAP: Distribution of Emphasis in Terms of Target Percentage of Test by Grade—Mathematics Grades 3–8 and 11

Subcategory	*GLE/GSE grade (grade tested)*						
	2 (3)	*3 (4)*	*4 (5)*	*5 (6)*	*6 (7)*	*7 (8)*	*9–10 (11)*
Numbers and Operations	55%	50%	45%	40%	30%	20%	15%
Geometry and Measurement	15%	20%	20%	25%	25%	25%	30%
Functions and Algebra	15%	15%	20%	20%	30%	40%	40%
Data, Statistics, and Probability	15%	15%	15%	15%	15%	15%	15%
TOTAL	100%	100%	100%	100%	100%	100%	100%

Table 2-10 shows the subcategory reporting structure for mathematics and the maximum possible number of raw score points that students could earn. The goal for distribution of score points or balance of representation across the four content strands varies from grade to grade. Note: only common items are reflected in this table, as only they are counted toward students' scaled scores.

Table 2-10. 2011–12 NECAP: Reporting Subcategories and Possible Raw Score Points by Grade—Mathematics Grades 3–8 and 11

Subcategory	*Grade tested*						
	3	*4*	*5*	*6*	*7*	*8*	*11*
Numbers and Operations	35	32	30	27	20	13	9
Geometry and Measurement	10	13	13	16	16	17	20
Functions and Algebra	10	10	13	13	20	26	25
Data, Statistics, and Probability	10	10	10	10	10	10	10
TOTAL	65	65	66	66	66	66	64

2.3.5. Depth of Knowledge

Each item on the NECAP test in mathematics is assigned a DOK level according to the cognitive demand of the item. DOK is not synonymous with difficulty. The DOK level rates the complexity of the mental processing a student must use to solve a problem. Each of the three levels is described in Figure 2-2.

Figure 2-2. 2011–12 NECAP: Depth of Knowledge—Mathematics

Level 1 (Recalling Information and Carrying Out Simple Procedures)	This level requires the recall of a fact, definition, term, or simple procedure; the application of a formula; or the performance of a straight algorithmic procedure. Items at this level may require students to demonstrate a rote response.
Level 2 (Skill/Concept)	This level requires mental processing beyond that of a simple habitual response. These items often require students to make some decisions about how to approach a problem.
Level 3 (Strategic Thinking, Reasoning, Planning, Drawing Conclusions, and Using Concepts and Evidence)	This level requires students to develop a plan or sequence of steps. These items are more complex and abstract than the items at the previous two levels. These items may also have more than one possible answer and may require students to use evidence, make conjectures, or justify their answers.

Table 2-11 lists the percentage of total score points assigned to each level of DOK in mathematics.

Table 2-11. 2011–12 NECAP: Depth of Knowledge in Terms of Targeted Percentage of Test by Grade—Mathematics Grades 3–8 and 11

DOK	*Grade*						
	3	*4*	*5*	*6*	*7*	*8*	*11*
Level 1	18%	26%	27%	21%	29%	23%	36%
Level 2	72%	66%	59%	76%	62%	52%	58%
Level 3	9%	8%	14%	3%	9%	26%	6%
TOTAL	100%	100%	100%	100%	100%	100%	100%

2.3.6. Use of Calculators and Reference Sheets

The mathematics specialists from the New Hampshire, Rhode Island, Maine, and Vermont Departments of Education who designed the mathematics test acknowledge the importance of mastering arithmetic algorithms. At the same time, they understand that the use of calculators is a necessary and important skill. Calculators can save time and prevent error in the measurement of some higher-order thinking skills and, in turn, allow students to work on more sophisticated and intricate problems. For these reasons, it was decided that, at grades 3 through 8, calculators should be prohibited in the first of the three sessions of the NECAP mathematics tests and permitted in the remaining two sessions. It was decided that, at grade 11, calculators should be prohibited in the first of the two sessions and permitted in the second session.

Reference sheets are provided to students at grades 5–8 and 11. These sheets contain information, such as formulas, that students may need to answer certain test items. The reference sheets are published each year with the released items and have remained the same for several years over the various test administrations. Toolkits are provided to students at grades 3–5. These toolkits contain manipulatives to answer specific questions. The toolkits are designed for specific items and therefore change annually. They are published with the released items. All students in grades 3–8 and 11 receive rulers for use on the mathematics test. Students in grades 5–7 receive protractors for use on the mathematics test. Students may keep the rulers and protractors after test administration.

2.4. WRITING TEST SPECIFICATIONS

2.4.1. Standards

Grades 5 and 8

The test framework for grades 5 and 8 writing was based on the NECAP GLEs, and each item on the NECAP tests was designed to measure a specific GLE. The content standards for grades 5 and 8 writing identify four major genres that are assessed in the writing portion of the NECAP test each year:

- writing in response to literary text
- writing in response to informational text
- narratives
- informational writing (report/procedure text for grade 5 and persuasive essay for grade 8)

Grade 11

The test framework for grade 11 writing was based on the NECAP GSEs, and each item on the NECAP test was designed to measure a specific GSE. The content standards for grade 11 writing identify six genres:

- writing in response to literary text
- writing in response to informational text
- report writing
- procedural writing
- persuasive writing
- reflective writing

2.4.2. Item Types

The NECAP writing tests include multiple-choice items, constructed-response items, and extended-response writing prompts. At grades 5 and 8, multiple-choice items provide breadth of coverage of the GLEs/GSEs, requiring approximately one minute for most students to answer each item. Constructed-response items are more complex, requiring 5 to 10 minutes of response time per item. At grades 5, 8, and 11, students are required to respond to a writing prompt, receiving 45 minutes (plus additional time if necessary) to develop a response. Each type of item is worth a specific number of points in the student's total writing score, as shown in Table 2-12.

Table 2-12. 2011–12 NECAP: Writing Item Types

Item Type	*Possible Score Points*
MC	0 or 1
CR	0, 1, 2, 3, or 4
ER	0, 2–12

MC = multiple-choice; CR=constructed-response;
ER = extended-response writing prompt

2.4.3. Test Design

All items on the grades 5 and 8 writing tests were common. Each multiple-choice item was worth one point, each constructed-response item was worth four points, and the extended-response writing prompt was worth 12 points.

Table 2-13 summarizes the test design used in the 2011–12 NECAP writing test for grade 11. There were a total of eight forms: five equating forms and three field-test forms. Each grade 11 student responded to two different extended-response writing prompts, one common and either one matrix-equating or one field-test prompt depending on the form. The common prompt was worth 12 points.

Table 2-13. 2011–12 NECAP: Number of Items by Item Type and Number of Items —Writing Grade11

Common	*Matrix–Equating (5 Forms)*	*Matrix–Field Test (3 Forms)*
1 ER	1 ER	1 ER

ER = extended-response writing prompt.

2.4.4. Blueprints

Grades 5 and 8

The writing prompt and the three constructed-response items each address a different genre. In addition, structures of language and writing conventions are assessed through multiple-choice items but are also assessed through the extended-response writing prompts and the constructed-response items. The extended-response writing prompts and constructed-response items were developed with the following criteria as guidelines:

- The prompts must be interesting to students.
- The prompts must be accessible to all students (i.e., all students would have something to write about the topic).
- The prompts must generate sufficient text to be effectively scored.

The category reporting structure for grades 5 and 8 writing is shown in Table 2-14. The table provides the maximum possible number of raw score points that students could earn. The content category “Short

Responses" lists the total raw score points from the three constructed-response items; the reporting category "Extended Response" lists the total raw score points from the writing prompt.

Table 2-14. 2011–12 NECAP: Reporting Subcategory and Possible Raw Score Points Possible by Grade—Writing Grades 5 and 8

Subcategory	*Grade Tested*	
	5	*8*
Structures of Language and Writing Conventions	10	10
Short Response	12	12
Extended Response	12	12
TOTAL	34	34

Short response = constructed-response items; Extended response = writing prompt

Grade 11

The writing prompts (common, matrix-equating, and field-test), in combination, address each of the different genres. The prompts were developed using the following criteria as guidelines:

- The prompt must be interesting to students.
- The prompt must be accessible to all students (i.e., all students would have something to write about the topic).
- The prompt must generate sufficient text to be effectively scored.

For grade 11 writing, there is only one reporting category, "Extended Response," with a total possible raw score of 12 points. One hundred percent of the raw score points for writing was assigned to DOK Level 3.

Each item on the NECAP test in writing is assigned a DOK level according to the cognitive demand of the item. DOK is not synonymous with difficulty. The DOK level rates the complexity of the mental processing a student must use to answer the question. Each of the three levels is described in Figure 2-3:

Figure 2-3. 2011–12 NECAP: Depth of Knowledge—Writing

Level 1	This level requires the student to write or recite simple facts. This writing or recitation does not include complex synthesis or analysis but basic ideas.
Level 2	This level requires some mental processing. Students are beginning to connect ideas using a simple organizational structure. For example, students may be engaged in note-taking, outlining, or simple summaries.
Level 3	This level requires some higher-level mental processing. Students are engaged in developing compositions that include multiple paragraphs. These compositions may include complex sentence structure and may demonstrate some synthesis and analysis. Students show awareness of their audience and purpose through focus, organization, and the use of appropriate compositional elements. The use of appropriate compositional elements includes such things as addressing chronological order in a narrative or including supporting facts and details in an informational report.

Table 2-15 lists the percentage of actual score points assigned to each level of DOK in writing for grades 5 and 8.

Table 2-15. 2011–12 NECAP: Depth of Knowledge by Grade (in Percentage of Test)—Writing Grades 5 and 8

DOK	*Grade Tested*	
	5	*8*
Level 1	24%	21%
Level 2	41%	44%
Level 3	35%	35%
TOTAL	100%	100%

Table 2-16 lists the percentage of actual score points assigned to each level of DOK in writing for grade 11.

Table 2-16. 2011–12 NECAP: Depth of Knowledge in Terms of Percentage of Test,—Writing Grade 11

DOK	*Grade 11*
Level 1	0%
Level 2	0%
Level 3*	100%
TOTAL	100

* In grade 11, 100% of the writing test is assigned to DOK Level 3.

2.5. TEST DEVELOPMENT PROCESS

2.5.1. Item Development

Items used on the NECAP tests are developed and customized specifically for use on the NECAP tests and are consistent with NECAP GLE and GSE content standards. Measured Progress test developers work with Maine, New Hampshire, Rhode Island, and Vermont educators to verify the alignment of items to the appropriate NECAP content standards.

The development process combined the expertise of Measured Progress test developers and committees of educators to help ensure items meet the needs of the NECAP tests. All items used on the common portions of the NECAP tests were reviewed by a committee of content experts and by a committee of bias experts. Tables 2-17 through 2-20 show the number of items developed within each content area for the 2011–2012 NECAP tests.

Table 2-17. 2011–12 NECAP: Annual English Language Arts Item Development—Grades 3–8

Passages	*MC*	*CR*
5 long passages (divided by literary and informational)	80	15
5 short passages (divided by literary and informational)	40	10
Standalones	20	0
10 total passages	140	25

MC = multiple-choice; CR = constructed-response

Table 2-18. 2011–12 NECAP: Annual English Language Arts Item Development—Grade 11

Passages	*MC*	*CR*
5 long passages (divided by literary and informational)	80	15
5 short passages (divided by literary and informational)	40	10
Standalones	20	0
10 total passages	140	25

MC = multiple-choice; CR = constructed-response

Table 2-19. 2011–12 NECAP: Annual Writing Item Development—Grade 11

Grade	*ER*
11	6

ER = extended-response writing prompt

Table 2-20. 2011–12 NECAP: Annual Mathematics Item Development—Grades 3–8 and 11

Grade	*MC*	*SA1*	*SA2*	*CR*
3	27	9	9	0
4	27	9	9	0
5	27	9	9	9
6	27	9	9	9
7	27	9	9	9
8	27	9	9	9
11	46	24	20	14

MC = multiple-choice; SA1 = 1-point short-answer; SA2 = 2-point short-answer; CR = constructed-response

2.5.2. Item Reviews at Measured Progress

For the internal item review, the lead Measured Progress test developer within the content area performed the following activities:

- review of the formatted items, open-response scoring guides, and any reading selections and graphics
- evaluation of item "integrity," content, and structure; appropriateness to designated content area; format; clarity; possible ambiguity; answer cueing; appropriateness and quality of reading selections and graphics; and appropriateness of scoring guide descriptions and distinctions (in relation to each item and across all items within the guide)
- ensuring that, for each item, there was only one correct answer
- consideration of scorability and evaluation as to whether the scoring guide adequately addressed performance on the item

Fundamental questions the lead developer considered, but was not limited to, included the following:

- What is the item asking?
- Is the key the only possible key? (Is there only one correct answer?)
- Is the open-response item scorable as written? (Were the correct words used to elicit the response defined by the guide?)
- Is the wording of the scoring guide appropriate and parallel to the item wording?
- Is the item complete (i.e., includes scoring guide, content codes, key, grade level, DOK, and identified contract)?
- Is the item appropriate for the designated grade level?

2.5.3. Item Reviews at State Level

Item Review Committees (IRCs) were formed by the states to provide an external review of items. The committees included teachers, curriculum supervisors, and higher education faculty from all four states, with committee members serving rotating terms. (A list of IRC member names and affiliations is included in Appendix A.) The committees' role is to review test items for the NECAP, provide feedback, and make recommendations about which items should be selected for program use. The 2011–12 NECAP IRCs for each content area in grade levels 3 through 8 and 11 met in the spring of 2010. Committee members reviewed the entire set of embedded field-test items proposed for the 2011–12 operational test and made recommendations about selecting, revising, or eliminating specific items from the item pool. Members reviewed each item against the following criteria:

- GLE/GSE alignment
  - Is the test item aligned to the appropriate GLE/GSE?
  - If not, which GLE/GSE or grade level is more appropriate?
- Content Correctness
  - Are the items and distractors correct with respect to content accuracy and developmental appropriateness?
  - Are the scoring guides consistent with GLE/GSE wording and developmental appropriateness?
- Depth of Knowledge[1]
  - Are the items coded to the appropriate DOK?
  - If consensus cannot be reached, is there clarity around why the item might be on the borderline of two levels?

[1] NECAP employed the work of Dr. Norman Webb to guide the development process with respect to Depth of Knowledge. Test specification documents identified ceilings and targets for Depth of Knowledge coding.

- Universal Design
  - Is there an appropriate use of simplified language? (Does it not interfere with the construct being assessed?)
  - 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 Brailled?
  - Is the item language clear?
  - Is the item language accurate (syntax, grammar, conventions)?

2.5.4. Bias and Sensitivity Review

Bias review is an essential part of the development process. NECAP passages and items were reviewed by a committee of teachers, English language learner 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 member names and affiliations is included in Appendix A.) Passages and items were examined for issues that might offend or dismay students, teachers, or parents. Including such groups in the development of test items and materials can prevent many unduly controversial issues and can allay unfounded concerns before the test forms are produced.

2.5.5. Reviewing and Refining

Test developers presented item sets to the IRCs who then recommended which items should be included in the embedded field-test portions of the test. The Maine, New Hampshire, Rhode Island, and Vermont Departments of Education content specialists made the final selections with the assistance of Measured Progress test developers at a final face-to-face meeting.

2.5.6. Item Editing

Measured Progress editors reviewed and edited the items to ensure uniform style (based on *The Chicago Manual of Style*, 15th edition) and adherence to sound testing principles. These principles included the stipulation that items

- were correct with regard to grammar, punctuation, usage, and spelling;
- were written in a clear, concise style;
- contained unambiguous explanations to students, detailing what is required to attain a maximum score;

- were written at a reading level that would allow the student to demonstrate his or her knowledge of the tested subject matter, regardless of reading ability;
- exhibited high technical quality in terms of psychometric characteristics;
- had appropriate answer options or score point descriptors; and
- were free of potentially sensitive content.

2.5.7. Item Selection and Operational Test Assembly

At Measured Progress, test assembly is the sorting and laying out of item sets into test forms. Criteria considered during this process for the 2011–12 NECAP included the following:

- *Content coverage / match to test design*. The Measured Progress test developers completed an initial sorting of items into sets based on a balance of reporting categories across sessions and forms, as well as a match to the test design (e.g., number of multiple-choice, short-answer, and constructed-response items).
- *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 across forms.
- *Visual balance*. Item sets were reviewed to ensure that each set reflected similar length and "density" of selected items (e.g., length/complexity of reading selections, number of graphics).
- *Option balance*. Each item set was checked to verify that it contained a roughly equivalent number of key options (As, Bs, Cs, and Ds).
- *Name balance*. Item sets were reviewed to ensure that a diversity of student names was used.
- *Bias*. Each item set was reviewed to ensure fairness and balance based on gender, ethnicity, religion, socioeconomic status, and other factors.
- *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 (a graphic or reading selection), consideration was given both 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 served to minimize the amount of "page flipping" required of students.
- *Relationship between forms*. Although embedded field-test items differ from form to form, they must take up the same number of pages in each form so that sessions and content areas begin on the same page in every form. Therefore, the number of pages needed for the longest form often determined the layout of each form.
- *Visual appeal*. The visual accessibility of each page of the form was always taken into consideration, including such aspects as the amount of "white space," the density of the text, and the number of graphics.

2.5.8. Operational Test Draft Review

Any changes made by a test construction specialist were reviewed and approved by a lead developer. After a form was laid out in what was considered its final form, it was reviewed to identify any final considerations, including the following:

- *Editorial changes*. All text was scrutinized for editorial accuracy, including consistency of instructional language, grammar, spelling, punctuation, and layout (based on Measured Progress's publishing standards and *The Chicago Manual of Style*, 15th edition).
- *"Keying" items*. Items were reviewed for any information that might "key" or provide information that would help to answer another item. Decisions about moving keying items are based on the severity of the "key-in" and the placement of the items in relation to each other within the form.
- *Key patterns*. The final sequence of keys was reviewed to ensure that their order appeared random (i.e., no recognizable pattern and no more than three of the same key in a row).

2.5.9. Alternative Presentations

Common items in each grade-level test were translated into Braille by a subcontractor that specializes in test materials for blind and visually impaired students. In addition, Form 1 for each grade was enlarged to a 20-point font to create a large-print version of the test for visally impaired students.

CHAPTER 3. TEST ADMINISTRATION

3.1. RESPONSIBILITY FOR ADMINISTRATION

The *2011 NECAP Principal/Test Coordinator Manual* indicated that principals and/or their designated NECAP test coordinators were responsible for the proper administration of the NECAP. Uniformity of administration procedures from school to school was ensured by using manuals that contained explicit directions and scripts to be read aloud to students by test administrators.

3.2. ADMINISTRATION PROCEDURES

Principals and/or the schools' designated NECAP test coordinators were instructed to read the *Principal/Test Coordinator Manual* before testing and to be familiar with the instructions provided in the grade-level *Test Administrator Manual*. The *Principal/Test Coordinator Manual* included a section highlighting aspects of test administration that were new for the year and checklists to help prepare for testing. The checklists outlined tasks to be performed by school staff before, during, and after test administration. In addition to these checklists, the *Principal/Test Coordinator Manual* described the testing material sent to each school and how to inventory it, track it during administration, and return it after testing was complete. The *Test Administrator Manual* included checklists for the administrators to use to prepare themselves, their classrooms, and the students for the administration of the tests. The *Test Administrator Manual* contained sections that detailed the procedures to be followed for each test session and instructions for preparing the material before the principal/test coordinator returned it to Measured Progress.

3.3. PARTICIPATION REQUIREMENTS AND DOCUMENTATION

The Departments of Education's intent is for *all* students in grades 3 through 8 and 11 to participate in the NECAP through standard administration, administration with accommodations, or alternate assessment. Furthermore, any student who is absent during any session of the NECAP tests is expected to take a make-up test within the three-week testing window.

Schools were required to return a Student Answer Booklet for every enrolled student in the grade level, with the exception of students who took an alternate assessment in the previous school year. Students who were alternately assessed in the 2010–11 school year were not required to participate in the NECAP in 2011–12. On those occasions when it was deemed impossible to test a particular student, school personnel were required to inform their Department of Education. A grid was included on the Student Answer Booklet that listed the approved reasons why a booklet could be returned blank for one or more sessions of the test:

- Student is new to the United States after October 1, 2010, and is LEP (reading and writing only)

  A. First-year LEP students who took the ACCESS test of English language proficiency, as scheduled in their states, were not required to take the reading and writing tests in 2011; however, these students were required to take the mathematics test in 2011.

- Student withdrew from school after October 1, 2011

  B. If a student withdrew after October 1, 2011, 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 October 1, 2011

  C. If a student enrolled after October 1, 2011, and was unable to complete all of the test sessions before the end of the test administration window, school personnel were instructed to code this reason on the student's answer booklet.

- State-approved special consideration

  D. Each state 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.

- Student was enrolled in school on October 1, 2011, and did not complete test for reasons other than those listed above.

  E. If a student was not tested for a reason other than those stated above, school personnel were instructed to code this reason on the student's answer booklet. These "Other" categories were considered "not state-approved."

Appendix B lists the participation rates of the four NECAP states combined in reading, mathematics, and writing.

3.4. ADMINISTRATOR TRAINING

In addition to distributing the *Principal/Test Coordinator Manual* and *Test Administrator Manual*, the Maine, New Hampshire, Rhode Island, and Vermont Departments of Education, along with Measured Progress, conducted test administration workshops in regional locations in each state to inform school personnel about the NECAP tests and to provide training on the policies and procedures regarding administration of the tests. A test administration workshop was also conducted via an online webinar for each state. These live webinars were recorded so that test coordinators and test administrators could view them at a time that was convenient for them. A link was provided to each state for their recorded workshop presentation in order for it to be added to the Department of Education Web site for school personnel to access. Lastly, an audio PowerPoint workshop presentation was prerecorded and provided to each state for inclusion on their Department of Education Web site.

3.5. DOCUMENTATION OF ACCOMMODATIONS

The *Principal/Test Coordinator Manual* and *Test Administrator Manual* provided directions for coding 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 first list of allowable accommodations was created by the three original NECAP states (New Hampshire, Rhode Island, and Vermont) at the beginning of the program in 2004. The list was later reviewed and revised in 2009 when the state of Maine joined the program. The four NECAP states worked together to change the coding system, revise existing accommodations, and add or delete certain accommodations. The resulting Table of Standard Test Accommodations is divided into accommodations for timing, setting, presentation, and response. Each accommodation is listed with details on how to deliver it to students. A *NECAP Accommodations Guide* was also produced to provide additional details on planning for and implementing accommodations. This guide was available on each state's Department of Education Web site. The states collectively made the decision that accommodations would continue to be made available to all students based on individual need regardless of disability status. Decisions regarding accommodations were to be made by the student's educational team on an individual basis and were to be consistent with those used during the student's regular classroom instruction. Making accommodations decisions for a group 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 Table of Standard Test Accommodations, schools were instructed to contact the Department of Education in advance of testing for specific instructions for coding in the "Other Accommodations (O)" and/or "Modifications (M)" sections.

Appendix C shows the accommodation frequencies by content area for the October 2011 NECAP test administration. The accommodation codes (T1–4, S1–2, P1–11, R1–7, O1–2, and M1–M3) are defined in the Table of Standard Test Accommodations, which can be found in Appendix D.

3.6. TEST SECURITY

Maintaining test security is critical to the success of the NECAP and the continued partnership among the four 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 school's test coordinator and/or principal immediately. The test coordinator or the principal or both were responsible for immediately reporting the concern to the District Superintendent and the State Assessment Director at the Department of Education. Test security was also strongly emphasized during test administration webinars as well as workshops that were conducted in all four states. The four states also required principals to log on to a secure Web site to complete the *Principal's Certification of Proper Test Administration* form for each grade level tested at their

school. Principals were requested 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 they were returning to Measured Progress. Principals were instructed to submit the form by entering a unique password, which acted as their digital signature. By signing and submitting the form, the principal was certifying that the tests were administered according to the test administration procedures outlined in the *Principal/Test Coordinator Manual* and *Test Administrator Manual*, that the security of the tests was maintained, that no secure material was duplicated or in any way retained in the school, and that all test materials had been accounted for and returned to Measured Progress.

3.7. TEST AND ADMINISTRATION IRREGULARITIES

There were several test irregularities that occurred during the 2011 NECAP test administration. One of the irregularities involved an issue with a reading field-test item in grade 4 while others were attributed to printing issues.

A reading field-test item associated with a field-test passage that was used in test form 8 and test form 9 contained a reference to a numbered paragraph in the passage. The issue is that students were asked to reference a specific paragraph by number but the paragraph was not numbered in the passage printed in the test booklet. Students were still able to respond to the item because the paragraph referenced was accessible in the passage. However, Measured Progress and the NECAP states decided not to score the item. Since this item was a field-test item, there was no affect on NECAP reports.

During test administration, a printing error was discovered in a grade 3 integrated test booklet and in a grade 4 integrated test booklet. Both of these printing errors were reported to Measured Progress by a school calling the NECAP Service Center. The two defective booklets were produced by the same print vendor. In the grade 3 test booklet, pages 3 through 10 were duplicated. In this case, the school was instructed to have the student skip those duplicate pages and complete the remaining items in the session. The school was also instructed to place the defective booklet in a special handling envelope for return to Measured Progress. In the grade 4 test booklet, pages 19 through 26 were printed in reverse order. The pages went from 1 to 18, then 26 to 19, followed by pages 27 to 84 which is the end of the test booklet. In this case, the school reported this after the student had completed testing. The school reported that the student was able to respond to all of the items and record the responses in the correct spaces in the test booklet. Therefore, NECAP reports were not affected by these irregularities. Measured Progress discussed these issues with the print vendor and is continuing to work with the company to ensure the print quality of the test booklets in the future.

3.8. TEST ADMINISTRATION WINDOW

The test administration window was October 3–25, 2011.

3.9. NECAP SERVICE CENTER

To provide additional support to schools before, during, and after testing, Measured Progress operates the NECAP Service Center. The support of a service center is essential to the successful administration of any statewide test program. It provides a centralized location to which individuals in the field can call, using a toll-free number, to ask specific questions or report any problems they may be experiencing. Representatives are responsible for receiving, responding to, and tracking calls, then routing issues to the appropriate person(s) for resolution. All calls are logged into a database that includes notes regarding the issue and resolution of each call.

The service center was staffed year-round and was available to receive calls from 8:00 a.m. to 4:00 p.m. Monday through Friday. Extra representatives were available as needed, beginning approximately two weeks before the start of the testing window and ending two weeks after the end of the testing window to assist with handling the additional call volume.

CHAPTER 4. SCORING

4.1. SCORING OF STANDARD TEST ITEMS

Upon receipt of used NECAP answer booklets following testing, Measured Progress scanned all student responses, along with student identification and demographic information. Imaged data for multiple-choice responses were machine-scored. Images of open-response items were processed and organized by iScore, a 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) and typed responses submitted according to applicable test accommodations were physically reviewed and scored on an individual basis by trained, qualified readers. 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.

4.1.1. Machine-Scored Items

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

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

4.1.2. Person-Scored Items

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

Through iScore, qualified readers at computer terminals accessed electronically scanned images of student responses. Readers evaluated each response and recorded each score via keypad or mouse entry

through the iScore system. When a reader finished one response, the next response appeared immediately on the computer screen.

Imaged responses from all answer booklets were sorted into item-specific groups for scoring purposes. Readers reviewed responses from only one item at a time; however, imaged responses from a student's entire booklet were always available for viewing when necessary, and the physical booklet was also available to the chief reader on-site. (Chief reader and other scoring roles are described in the section that follows.)

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

4.1.2.1. Scoring Location and Staff

Scoring Location

The iScore database, its operation, and its administrative controls are all based in Dover, New Hampshire. Table 4-1 presents the locations where 2011–12 NECAP test item responses by grade and content area were scored.

Table 4-1. 2011–12 NECAP: Operational Scoring Locations by Content Area and Grade

Content area	*Grade*	*Dover, NH*	*Menands, NY*	*Longmont, CO*
Mathematics	3	X		
	4			X
	5			X
	6			X
	7			X
	8			X
	11			X
Reading	3	X		
	4		X	
	5		X	
	6		X	
	7		X	
	8			X
	11			X
Writing	5		X	
	8			X
	11			X

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

Staff Positions

The following staff members were involved with scoring the 2011–12 NECAP responses:

- The NECAP scoring project manager, an employee of Measured Progress, was located in Dover, New Hampshire, and oversaw communication and coordination of scoring across all scoring sites.
- The iScore operational manager and iScore administrators, employees of Measured Progress, were located in Dover, New Hampshire, and coordinated technical communication across all scoring sites.
- A chief reader in each content area (mathematics, reading, and writing) ensured consistency of scoring across all scoring sites for all grades tested in that content area. Chief readers also provided read-behind activities (defined in a later section) for quality assurance coordinators (QACs). Chief readers are employees of Measured Progress.
- Numerous QACs, selected from a pool of experienced senior readers for their ability to score accurately and their ability to instruct and train readers, participated in benchmarking activities for each specific grade and content area. QACs provided read-behind activities (defined in a later section) for senior readers at their sites. The ratio of QACs and senior readers to readers was approximately 1:11.
- Numerous senior readers, selected from a pool of skilled and experienced readers, provided read-behind activities (defined in a later section) for the readers at their scoring tables (2–12 readers at each table). The ratio of QACs and senior readers to readers was approximately 1:11.
- Readers at scoring sites scored operational and field-test NECAP 2011–12 student responses. Recruitment of readers is described in Section 4.1.2.3.

4.1.2.2. Benchmarking Meetings with the NECAP State Specialists

In preparation for implementing NECAP scoring guidelines, Measured Progress scoring staff prepared and facilitated benchmarking meetings held with NECAP state specialists from their respective Departments of Education. The purpose of these meetings was to establish guidelines for scoring NECAP items during the current field-test scoring session and for future operational scoring sessions.

Several dozen student responses for each item that chief readers identified as illustrative midrange examples of the respective score points were selected. Chief readers presented these responses to the NECAP content specialists during benchmarking meetings and worked collaboratively with them to finalize an authoritative set of score point exemplars for each field-test item. As a matter of practice, these sets are included in the scoring training materials each time an item is administered.

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

4.1.2.3. Reader Recruitment and Qualifications

For scoring the 2011–12 NECAP, Measured Progress actively sought a diverse scoring pool representative of the population of the four NECAP states. The broad range of reader backgrounds included scientists, editors, business professionals, authors, teachers, graduate school students, and retired educators. Demographic information (e.g., gender, race, educational background)about readers was electronically captured for reporting.

Although a four-year college degree or higher was preferred, readers were required to have successfully completed at least two years of college and to have demonstrated knowledge of the content area they scored. This permitted recruiting readers currently enrolled in a college program, a sector of the population with relatively recent exposure to current classroom practices and trends in their fields. In all cases, potential readers were required to submit documentation (e.g., résumé and/or transcripts) of their qualifications.

Table 4-2 summarizes the qualifications of the 2011–12 NECAP scoring leadership and readers.

Table 4-2. 2011–12 NECAP: Qualifications of Scoring Leadership and Readers—Fall Administration

Scoring Responsibility	*Educational Credentials*				*Total*
	Doctorate	*Master's*	*Bachelor's*	*Other*	
Scoring Leadership	7.6%	29.7%	55.7%	7.0%	100.0%
Readers	5.1%	29.4%	54.6%	10.9%	100.0%

Scoring Leadership = chief readers, QACs, and senior readers
*9 QACs/senior readers had an associate's degree and 2 had at least 48+ college credits.
**77 readers had an associate's degree and 53 had at least 48+ college credits.

Readers were either temporary Measured Progress employees or were secured through temporary employment agencies. All readers were required to sign a nondisclosure/confidentiality agreement.

4.1.2.4. Methodology for Scoring Polytomous Items

Possible Score Points

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

Table 4-3. 2011–12 NECAP: Possible Score Points for Polytomous Item Types

Polytomous Item Type	*Possible Score Point Range*
Writing prompt	0–6
Constructed-response	0–4
2-point short-answer (SA2)	0–2
1-point short-answer (SA1)	0–1
Non-scorable items	0

Non-Scorable Items

Readers could designate a response as non-scorable 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)—*see note below.*
- Response was written in the wrong location (seemed to be a legitimate answer to a different question)—*see note below.*
- 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.

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

Scoring Procedures

Scoring procedures for polytomous items included both single scoring and double scoring. Single-scored items were scored by one reader. Double-scored items were scored independently by two readers, whose scores were tracked for "interrater agreement." (For further discussion of double scoring and interrater agreement, see Section 5.1.2.7 and Appendix P.)

4.1.2.5. Reader Training

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

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

Following review of an item's scoring guide, readers reviewed or scored the particular response set organized for that training: anchor sets, training sets, and qualifying sets. (These are defined below.)

During training, readers could highlight or mark hard copies of the anchor and training sets (as well as the first qualifying sets after the qualification round), even if all or part of the set was also presented online.

Anchor Set

Readers first reviewed an anchor set of exemplary responses for an item. This is a set approved by the reading, writing, and mathematics content specialists representing the four NECAP state Departments of Education. Responses in anchor sets are 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 client and Measured Progress scoring staff. Each set contains one client-approved sample response per score point considered to be a midrange exemplar. The set included a second sample response if there was more than one plausible way to illustrate the merits and intent of a score point.

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

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

Training Set

Next, readers practiced applying the scoring guide and anchors to responses in the training set. The training set typically included 10 to 15 student responses designed to help establish both the full score point range and the range of possible responses within each score point. The training set often included unusual responses that were less clear or solid (e.g., shorter than normal, employing atypical approaches, simultaneously containing very low and very high attributes, and written in ways difficult to decipher). Responses in the training set were presented in randomized score point order.

After readers independently read and scored a training set response, trainers would poll readers or use online training system reports to record their initial range of scores. Trainers then led a group discussion of one or two responses, directing readers' attentions to difficult scoring issues (e.g., the borderline between two score points). Throughout the training, trainers modeled how to discuss scores by referring to the anchor set and to scoring guides.

Qualifying Set

After the training set had been completed, readers were required to score responses accurately and reliably in qualifying sets composed of constructed-response items, writing prompts, and all 2-point short-answer items for grades 3 and 4 mathematics. The 10 responses in each qualifying set were selected from an array of responses that clearly illustrated the range of score points for that item as reviewed and approved by the state specialists. Hard copies of the responses were also made available to readers after the qualification round so that they could make notes and refer back during the post-qualifying discussion.

To be eligible to live-score one of the above items, readers were required to demonstrate scoring accuracy rates of at least 80% exact agreement (i.e., to exactly match the predetermined score on at least eight of the ten responses) and at least 90% exact or adjacent agreement (i.e., to exactly match or be within one

score point of the predetermined score on nine or ten of the ten responses), except 70% and 90%, respectively, for 6-point writing-prompt responses. In other words, readers were allowed one discrepant score (i.e., one score of 10 that was more than 1score point from the predetermined score) provided they had at least eight exact scores (seven for writing-prompt items).

To be eligible to score 1-point short-answer mathematics items (which were benchmarked "right" or "wrong") and 2-point short-answer mathematics items for grades 5–8 and 11, readers had to qualify on at least one other mathematics item for that grade.

Retraining

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

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

4.1.2.6. Senior Quality Assurance Coordinator and Senior Reader Training

QACs and select senior readers were trained in a separate training session immediately prior to reader training. In addition to a discussion of the items and their responses, QAC and senior reader training included greater detail on the client's rationale behind the score points than that covered with regular readers in order to better equip QACs and senior readers to handle questions from the readers.

4.1.2.7. Monitoring of Scoring Quality Control and Consistency

Readers were monitored for continued accuracy and consistency throughout the scoring process, using the following methods and tools (which are defined in this section):

- embedded committee-reviewed responses (CRRs)
- read-behind procedures
- double-blind scoring
- recalibration sets
- scoring reports

It should be noted that any reader whose accuracy rate fell below the expected rate for a particular item and monitoring method was retrained on that item. Upon approval by the QAC or chief reader as appropriate (see below), the reader was allowed to resume scoring. Readers who met or exceeded the expected accuracy rates continued scoring.

Furthermore, the accuracy rate required of a reader to *qualify* to score responses live was stricter than that required to *continue* to score responses live. The reason for the difference is that an "exact score" in double-blind scoring requires that *two* readers choose the same score for a response (in other words, it is dependent on peer agreement), whereas an "exact score" in qualification requires only that a *single* reader match a score pre-established by scoring leadership. The use of multiple monitoring techniques is critical toward monitoring reader accuracy during the process of live scoring.

Embedded CRRs

CRRs are previously scored responses that are loaded ("embedded") by scoring leadership into iScore and distributed "blindly" to readers during scoring. Embedded CRRs may be chosen either before or during scoring, and are inserted into the scoring queue so that they appear the same as all other live student responses.

Between five and thirty embedded CRRs were distributed at random points throughout the first full day of scoring to ensure that readers were sufficiently calibrated at the beginning of the scoring period. Individual readers often received up to 20 embedded CRRs within the first 100 responses scored and up to 10 additional responses within the next 100 responses scored on that first day of scoring.

Any reader who fell below the required scoring accuracy rate was retrained before being allowed by the QAC to continue scoring. Once allowed to resume scoring, scoring leadership carefully monitored these readers by increasing the number of read-behinds (defined in the next section).

Embedded CRRs were employed for all constructed-response items. They were not used for writing 6-point extended-response items, because these are 100% double-blind scored (defined below). Embedded CRRs were also not used for math 2-point short-answer items, because read-behind and double-blind techniques are more informative and cost-effective for these items.

Read-Behind Procedures

Read-behind scoring refers to scoring leadership (usually a senior reader) scoring a response after a reader has already scored the response. The practice was applied to all open-response item types.

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

The senior reader entered his or her score into iScore before being allowed to see the reader's score. The senior reader then compared the two scores, and the score of record (i.e., the reported score) was determined as follows:

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

scoring consultation with the senior reader, after which the QAC determined whether or when the reader could resume scoring.)

- If the scores were discrepant (i.e., differed by more than one point), the senior reader's score became the score of record. (This triggered an individual consultation for the reader with the senior reader, after which the QAC determined whether or when the reader could resume scoring on that item.)

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

Table 4-4. 2011–12 NECAP: Examples of Read-Behind Scoring Resolutions

Reader Score	*QAC/SR Score*	*Score of Record*
4	4	4
4	3	3*
4	2	2*

* QAC/senior reader's score

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

In addition to regular read-behinds, scoring leadership could choose to do read-behinds on any reader at any point during the scoring process to gain an immediate, real-time "snapshot" of a reader's accuracy.

Double-Blind Scoring

Double-blind scoring refers to two readers independently scoring a response without knowing whether the response was to be double-blind scored. The practice was applied to all open-response item types. Table 4-5 shows by which method(s) the responses to both common and equating open-response item types for each operational test were scored.

Table 4-5. 2011–12 NECAP: Frequency of Double-Blind Scoring by Grade and Content

Grade	*Content Area*	*Responses Double-Blind Scored*
3–8, 11	Reading	2% randomly
3–8, 11	Mathematics	2% randomly
5, 8, 11	Writing (ER)	100%
5, 8	Writing (CR)	2% randomly
All	Unreadable responses	100%
All	Blank responses	100%

If there was a discrepancy (a difference greater than one score point) between double-blind scores, the response was placed into an arbitration queue. Arbitration responses were reviewed by scoring leadership (senior Reader or QAC) without knowledge of the two readers' scores. Scoring leadership assigned the final

score. Appendix P provides the NECAP 2011–12 percentages of agreement between readers for each common item for each grade and content area.

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

Recalibration Sets

To determine whether readers were still calibrated to the scoring standard, they 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.

- Readers who were discrepant on two of the 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.
- Readers who were discrepant on only one of the five responses of the first recalibration set, and/or exact on three, were retrained by their senior reader by discussing the recalibration set responses in terms of the score point descriptions and the original anchor set. After this retraining, such readers began scoring operational responses under the proviso that the reader's scores for that day and that item would be kept only if the reader was exact on all five of the five responses of the second recalibration set administered at the shift midpoint. The QAC determined whether or when these Readers had received enough retraining to resume scoring operational responses. Scoring leadership also carefully monitored the accuracy of such readers by significantly increasing the number of their read-behinds.
- Readers 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 set 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 readers would have achieved an overall 80%-exact and 90%-adjacent standard for that item for that day.

The scoring project manager voided all scores posted on that item for that day by readers who did not meet the accuracy requirement. Responses associated with voided scores were reset and redistributed to readers with demonstrated accuracy for that item.

Recalibration sets were employed for all constructed-response items. They were not used for writing 6-point extended-response items, which were 100% double-blind scored. They were also not used for 2-point short-answer items, for which read-behind and double-blind techniques are more informative and cost-effective.

Scoring Reports

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

CHAPTER 5. CLASSICAL ITEM ANALYSIS

As noted in Brown (1983), "A test is only as good as the items it contains." A complete evaluation of a test's quality must include an evaluation of each item. Both *Standards for Educational and Psychological Testing* (AERA et al., 1999) 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 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 statistics, and (4) dimensionality analyses. The item analyses presented here are based on the statewide administration of NECAP in fall 2011. 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.)

5.1. 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 and one-point short-answer 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 2-point short-answer items, for which students can receive scores of 0, 1, or 2, and constructed-response items, which are worth 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 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 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 of 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 approximately 0.4 and 0.7 for the lower mathematics grades and between about 0.6 and 0.8 for reading, writing, and the higher mathematics grades. However, on a standards-referenced assessment such as NECAP, it may be appropriate to include some items with very low or very high item difficulty values to ensure sufficient content coverage.

A desirable characteristic of an item is for higher-ability students to perform better on the item than lower-ability students do. The correlation between student performance on a single item and total test score is a commonly used measure of this characteristic of the item. Within classical test theory, the item-test correlation is referred to as the item's discrimination, because it indicates the extent to which successful performance on an item discriminates between high and low scores on the test. For constructed-response items, the item discrimination index used was the Pearson product-moment correlation; for multiple-choice items, the corresponding statistic is commonly referred to as a point-biserial correlation. The theoretical range of these statistics is –1.0 to 1.0, with a typical observed range from 0.2 to 0.7.

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 content area and grade is presented in Table 5-1. Note that the statistics are presented for all items as well as by item type (multiple-choice, short-answer, constructed-response, and, for writing, writing prompt). Note also that only a single writing prompt is administered in grades 5 and 8; thus, it is not possible to calculate standard deviations of the difficulty and discrimination values. Furthermore, because the grade 11 writing test consists solely of a single prompt, no discrimination values or standard deviations could be calculated. The mean difficulty and discrimination values shown in the table are within generally acceptable and expected ranges.

Table 5-1. 2011–12 NECAP: Summary of Item Difficulty and Discrimination Statistics by Content Area and Grade

Content Area	*Grade*	*Item Type*	*Number of Items*	*p-Value*		*Discrimination*	
				Mean	*Standard Deviation*	*Mean*	*Standard Deviation*
Mathematics	3	ALL	55	0.68	0.16	0.43	0.10
		MC	35	0.71	0.16	0.41	0.09
		SA	20	0.62	0.16	0.47	0.10
	4	ALL	55	0.66	0.15	0.45	0.08
		MC	35	0.70	0.16	0.42	0.07
		SA	20	0.60	0.12	0.50	0.06
	5	ALL	48	0.57	0.16	0.43	0.10
		CR	4	0.52	0.12	0.62	0.03
		MC	32	0.60	0.16	0.40	0.08
		SA	12	0.51	0.16	0.47	0.10

continued

Content Area	Grade	Item Type	Number of Items	p-Value		Discrimination	
				Mean	Standard Deviation	Mean	Standard Deviation
Mathematics	6	ALL	48	0.54	0.16	0.45	0.11
		CR	4	0.44	0.05	0.66	0.07
		MC	32	0.55	0.17	0.39	0.08
		SA	12	0.57	0.11	0.51	0.08
	7	ALL	48	0.53	0.17	0.44	0.13
		CR	4	0.38	0.08	0.63	0.03
		MC	32	0.58	0.18	0.37	0.10
		SA	12	0.45	0.11	0.54	0.11
	8	ALL	48	0.54	0.17	0.45	0.12
		CR	4	0.45	0.11	0.66	0.03
		MC	32	0.58	0.17	0.40	0.08
		SA	12	0.47	0.17	0.53	0.08
	11	ALL	46	0.41	0.18	0.48	0.13
		CR	4	0.31	0.04	0.67	0.07
		MC	24	0.49	0.19	0.40	0.10
		SA	18	0.32	0.14	0.54	0.09
Reading	3	ALL	34	0.72	0.16	0.45	0.08
		CR	6	0.54	0.23	0.51	0.06
		MC	28	0.76	0.11	0.43	0.08
	4	ALL	34	0.70	0.12	0.44	0.08
		CR	6	0.57	0.16	0.51	0.08
		MC	28	0.73	0.09	0.42	0.07
	5	ALL	34	0.70	0.17	0.44	0.08
		CR	6	0.41	0.03	0.55	0.07
		MC	28	0.76	0.11	0.41	0.06
	6	ALL	34	0.73	0.16	0.41	0.11
		CR	6	0.46	0.03	0.61	0.04
		MC	28	0.79	0.10	0.37	0.06
	7	ALL	34	0.70	0.14	0.42	0.13
		CR	6	0.49	0.05	0.65	0.02
		MC	28	0.74	0.11	0.37	0.07
	8	ALL	34	0.70	0.15	0.41	0.13
		CR	6	0.50	0.05	0.66	0.03
		MC	28	0.74	0.12	0.36	0.07
	11	ALL	34	0.69	0.12	0.44	0.13
		CR	6	0.56	0.06	0.70	0.02
		MC	28	0.72	0.12	0.39	0.06
Writing	5	ALL	14	0.72	0.18	0.40	0.14
		CR	3	0.45	0.05	0.59	0.03
		MC	10	0.82	0.03	0.33	0.06
		ER	1	0.44		0.62	
	8	ALL	14	0.71	0.11	0.43	0.15
		CR	3	0.60	0.04	0.64	0.04
		MC	10	0.76	0.08	0.35	0.04
		ER	1	0.50		0.67	
	11	ALL	1	0.52			
		ER	1	0.52			

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, for mathematics, students in higher grade levels found their items more difficult than students in lower grades found theirs while, for reading, difficulty indices were more consistent across grades.

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 E; 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 was negative. While it is not inappropriate to include items with low discrimination values or with very high or very low item difficulty values to ensure that content is appropriately covered, there were very few such cases on NECAP. Item-level score point distributions are provided for open-response items in Appendix F; for each item, the percentage of students who received each score point is presented.

5.2. DIFFERENTIAL ITEM FUNCTIONING

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 construct-irrelevant, factors. *Standards for Educational and Psychological Testing* (AERA et al., 1999) includes similar guidelines. As part of the effort to identify such problems, an evaluation of the NECAP items was conducted in terms of differential item functioning (DIF) statistics.

For NECAP, 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. In order to calculate DIF statistics, a minimum of 200 students must be in each comparison group.

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 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 2011–12 NECAP tests, seven subgroup comparisons were evaluated for DIF:

- male versus female
- no disability versus disability
- non-economically disadvantaged versus economically disadvantaged
- non-LEP versus LEP
- White versus Asian
- White versus Black
- White versus Hispanic

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

5.3. Dimensionality Analysis

Because tests are constructed with multiple content area subcategories and their associated knowledge and skills, the potential exists for a large number of dimensions being invoked beyond the common primary dimension. Generally, the subcategories are highly correlated with each other; therefore, the primary dimension they share typically explains an overwhelming majority of variance in test scores. In fact, the

[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 Department of Education to determine whether to include the flagged item in a future operational test administration.

presence of just such a dominant primary dimension is the psychometric assumption that provides the foundation for the unidimensional item response theory (IRT) models that are used for calibrating, linking, scaling, and equating the NECAP 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 2011–12 NECAP common items for mathematics, reading, and writing are reported below. (Note: only common items were analyzed since they are used for score reporting, and grade 11 writing was not analyzed because it consisted of a single assessment task.)

The dimensionality analyses were conducted using the nonparametric IRT-based methods DIMTEST (Stout, 1987; Stout, Froelich, & Gao, 2001) and DETECT (Zhang & Stout, 1999). Both of these methods use as their basic statistical building block the estimated average conditional covariances for item pairs. A conditional covariance is the covariance between two items conditioned on total score for the rest of the test, and the average conditional covariance is obtained by averaging overall 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, nonrandom patterns of positive and negative conditional covariances are indicative of multidimensionality.

DIMTEST is a hypothesis-testing procedure for detecting violations of local independence. The data are first randomly divided into a training sample and a cross-validation sample. Then an exploratory analysis of the conditional covariances is conducted on the training sample data to find the cluster of items that displays the greatest evidence of local dependence. The cross-validation sample is then used to test whether the conditional covariances of the selected cluster of items 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. As with DIMTEST, the data are first randomly divided into a training sample and a cross-validation sample (these samples are drawn independent of those used with DIMTEST). The training sample is used to find a set of mutually exclusive and collectively exhaustive clusters of items that best fits 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 are used with the cross-validation sample data to average the conditional covariances: within-cluster conditional covariances are summed, from this sum the between-cluster conditional covariances are subtracted, this difference is divided by the total number of item pairs, and this average is multiplied by 100 to yield an index of the average violation of local independence for an item pair. DETECT values less than 0.2 indicate very weak multidimensionality (or near unidimensionality), values of

0.2 to 0.4 indicate weak to moderate multidimensionality, values of 0.4 to 1.0 indicate moderate to strong multidimensionality, and values greater than 1.0 indicate very strong multidimensionality.

DIMTEST and DETECT were applied to the 2011–12 NECAP. The data for each grade and content area were split into a training sample and a cross-validation sample. Every grade and content area combination had at least 32,000 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 16,000 students. DIMTEST was then applied to every grade and content area combination. DETECT was applied to each dataset for which the DIMTEST null hypothesis was rejected in order to estimate the effect size of the multidimensionality.

The results of the DIMTEST analyses indicated that the null hypothesis was rejected at a significance level of 0.01 for every dataset. Because strict unidimensionality is an idealization that almost never holds exactly for a given dataset, these DIMTEST results were not surprising. Indeed, because of the very large sample sizes of NECAP, DIMTEST would be expected to be sensitive to even quite small violations of unidimensionality. Thus, it was important to use DETECT to estimate the effect size of the violations of local independence found by DIMTEST. Table 5-2 below displays the multidimensional effect size estimates from DETECT.

Table 5-2. 2011–12 NECAP: Multidimensionality Effect Sizes by Grade and Content Area

Content Area	*Grade*	*Multidimensionality Effect Size*	
		2011–12	*2010–11*
Reading	3	0.15	0.13
	4	0.20	0.19
	5	0.22	0.24
	6	0.21	0.23
	7	0.34	0.21
	8	0.30	0.34
	11	0.37	0.29
	Average	0.26	0.23
Mathematics	3	0.12	0.16
	4	0.12	0.13
	5	0.15	0.16
	6	0.15	0.18
	7	0.14	0.14
	8	0.16	0.11
	11	0.11	0.13
	Average	0.14	0.15
Writing	5	0.25	0.24
	8	0.20	0.28
	Average	0.23	0.26

All of the DETECT values for mathematics and writing indicated very weak to weak multidimensionality. For reading, grades 3 through 6 had very weak to weak multidimensionality, whereas grades 7, 8, and 11 had DETECT values indicative of moderate multidimensionality. The average DETECT values for all three content areas indicated weak (reading, writing) or very weak (mathematics) multidimensionality. Also shown in Table 5-2 are the values reported in last year's dimensionality analyses. The individual values for the different grade levels, as well as the averages for the three content areas, in the 2011–12 NECAP are seen to be very similar to those from last year with the exception of grade 7 reading, which had an increase of 0.13 (0.21 to 0.34) and, thus, went from weak to moderate multidimensionality. Looking back over the analyses of the past four years, the previous high for the grade 7 reading DETECT value had been 0.28 in 2007–08, and the average over the past four years had been 0.23.

The way in which DETECT divided the tests into clusters was also investigated to determine whether there were any discernable patterns with respect to the multiple-choice and constructed-response item types. Inspection of the DETECT clusters indicated that multiple-choice–constructed-response separation occurred much more strongly with reading and writing than with mathematics, which is a pattern that has been consistent across all five years of dimensionality analyses for the NECAP fall tests. Specifically, for mathematics, grades 3, 5, and 7 showed some evidence of multiple-choice–constructed-response separation in that they each had one or two clusters dominated by multiple-choice points, one or two clusters dominated by constructed-response points, as well as other clusters evenly mixed between multiple-choice and constructed-response points. Typically for these three mathematics tests, 60% or more of the points were accounted for by the multiple-choice- and constructed-response-dominated clusters. The remaining mathematics tests displayed no substantial separation of multiple-choice and constructed-response items in their DETECT solutions. In reading, on the other hand, grades 4, 5, 6, 7, 8, and 11 all displayed strong multiple-choice–constructed-response separation with one cluster containing all the multiple-choice items with no constructed-response items, and the remaining cluster(s) containing all the constructed-response items. Indeed, for reading grades 7, 8, and 11 (the three with moderate multidimensionality), DETECT indicated a two-cluster solution in every case where one cluster was all multiple-choice and the other was all constructed-response. In reading, only grade 3 displayed no evidence of any multiple-choice–constructed-response separation—this finding was also reported for reading grade 3 for the 2010–11 test. For writing, both grades displayed a strong multiple-choice–constructed-response separation in the same manner as occurred with reading, with one cluster having all the multiple-choice items and the remaining clusters having all the constructed-response items.

Thus, a tendency is suggested for multiple-choice and constructed-response to sometimes measure statistically separable dimensions, especially in regard to the reading and writing tests. This has been consistent across all five years of analyses of the NECAP fall test administrations. However, it is important to emphasize that the degree of violation of unidimensional local independence has not been large in any of the three content areas over the five years of analysis. The degree to which these violations of local independence can be attributed to item type differences tends to be greater for reading and writing than for mathematics.

More investigation by content experts would be required to better understand the violations of local independence that are due to sources other than item type.

In summary of the 2011–12 analyses, the violations of local independence, as evidenced by the DETECT effect sizes, were weak or very weak in all but three cases (reading in grades 7, 8, and 11)—all three of which had effect sizes near the 0.30 border between weak and moderate. Thus, these effects do not seem to warrant any changes in test design or scoring, which are designed to be consistent across a range of grade levels. In addition, the magnitude of the violations of local independence have been consistently low over the years, and the patterns with respect to the multiple-choice and constructed-response items have also been consistent, with reading and writing tending to display more separation than mathematics.

CHAPTER 6. ITEM RESPONSE THEORY (IRT) SCALING AND EQUATING

This chapter describes the procedures used to calibrate, equate, and scale NECAP. 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 *a*-plots and *b*-plots for reasonableness); and evaluation of the scaling results (e.g., parallel processing by the Psychometrics and Research department and Data Analysis department; 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 member Departments of Education for their approval prior to production of student reports.

Table 6-1 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 Results and Equating Results sections below.

Table 6-1. 2011–12 NECAP: Items That Required Intervention During IRT Calibration and Equating

Content Area	*Grade*	*Item Number*	*Reason*	*Action*
Mathematics	3	119763	c parameter	c = 0
		119911	c parameter	c = 0
		201312	c parameter	c = 0
		168752	Delta Analysis	Removed from equating
		168752	IRT Plot Outlier	Removed from equating
	4	145070	c parameter	c = 0
		169080	c parameter	c = 0
		139493	Delta Analysis	Removed from equating
		120224	IRT Plot Outlier	Removed from equating
		120203	Rescore Analysis	Retained for Equating
	5	120757	c parameter	c = 0
		145892	c parameter	c = 0
		139164	IRT Plot Outlier	Removed from equating
		120838	Delta Analysis	Removed from equating
		120838	IRT Plot Outlier	Removed from equating
	6	150829	Delta Analysis	Removed from equating
		150829	IRT Plot Outlier	Removed from equating
	7	120344	c parameter	c = 0
		144742	c parameter	c = 0
		120444	c parameter	c = 0
		120411	c parameter	c = 0

continued

Content Area	Grade	Item Number	Reason	Action
Mathematics	7	125286	c parameter	c = 0
		140025	Delta Analysis	Removed from equating
		140025	IRT Plot Outlier	Removed from equating
		224856	Delta Analysis	Removed from equating
	8	145644	c parameter	c = 0
		120991	c parameter	c = 0
		121040	Delta Analysis	Removed from equating
		121040	IRT Plot Outlier	Removed from equating
	11	169689	c parameter	c = 0
		169763	c parameter	c = 0
		119446	c parameter	c = 0
		140028	IRT Plot Outlier	Removed from equating
		169507	c parameter	c = 0
Reading	3	117648	c parameter	c = 0
		147819	c parameter	c = 0
		148219	c parameter	c = 0
		148178	c parameter	c = 0
		147960	c parameter	c = 0
		117796	Delta Analysis	Removed from equating
		147875	a parameter	a set to initial
		148163	a parameter	a set to initial
		171335	IRT Plot Outlier	Removed from equating
	4	117934	c parameter	c = 0
		117924	c parameter	c = 0
		171630	Delta Analysis	Removed from equating
	5	171916	c parameter	c = 0
		171793	c parameter	c = 0
	6	118255	c parameter	c = 0
		148768	c parameter	c = 0
		148388	c parameter	c = 0
		148399	c parameter	c = 0
		148412	c parameter	c = 0
		148428	c parameter	c = 0
		118333	c parameter	c = 0
		118338	c parameter	c = 0
		118339	c parameter	c = 0
		148508	c parameter	c = 0
		171857	Delta Analysis	Removed from equating
		171857	IRT Plot Outlier	Removed from equating
		172099	Rescore Analysis	Retained for Equating
	7	129211	c parameter	c = 0
		129212	c parameter	c = 0
		129216	c parameter	c = 0
		129217	c parameter	c = 0
		173078	c parameter	c = 0
		173091	c parameter	c = 0
		173094	c parameter	c = 0
		118572	c parameter	c = 0
		118526	c parameter	c = 0
		118573	c parameter	c = 0

continued

Content Area	Grade	Item Number	Reason	Action
Reading	7	147546	IRT Plot Outlier	Removed from equating
		173247	Delta Analysis	Removed from equating
		173268	Rescore Analysis	Retained for Equating
		173260	c parameter	c = 0
	8	172733	c parameter	c = 0
		147322	c parameter	c = 0
		147334	c parameter	c = 0
		147333	c parameter	c = 0
		172379	c parameter	c = 0
		172385	c parameter	c = 0
		172401	c parameter	c = 0
		172743	c parameter	c = 0
		172353	Delta Analysis	Removed from equating
		118626	c parameter	c = 0
		118627	c parameter	c = 0
	11	147433	c parameter	c = 0
		147435	c parameter	c = 0
		147439	c parameter	c = 0
		147463	c parameter	c = 0
		172434	c parameter	c = 0
		144035	c parameter	c = 0
		172833	c parameter	c = 0
		147388	Rescore Analysis	Retained for Equating
		172360	Delta Analysis	Removed from equating
		172360	IRT Plot Outlier	Removed from equating

6.1. ITEM RESPONSE THEORY (IRT)

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

For the 2011–12 NECAP, 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:

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

where
i indexes the items,
j indexes students,
α represents item discrimination,
b represents item difficulty,
c is the pseudoguessing parameter,
ξ_i represents the set of item parameters (α, b, and c) for item i, and
D is a normalizing constant equal to 1.701.

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

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

where
i indexes the items,
j indexes students,
k indexes threshold,
α represents item discrimination,
b represents item difficulty,
d represents threshold, and
D is a normalizing constant equal to 1.701.

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

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

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:

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

where
ξ_i represents the set of item parameters for item i.

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.

$$P_i(1|\theta_j) = \sum_{k}^{m+1} w_{ik} P_{ik}(1|\theta_j)$$

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

6.2. Item Response Theory Results

The tables in Appendix H give the IRT item parameters of all common items on the 2011–12 NECAP tests by grade and content area. In addition, Appendix I shows graphs of the TCCs and TIFs, which are defined below.

TCCs display the expected (average) raw score associated with each θ_j value between –4.0 and 4.0. Mathematically, the TCC is computed by summing the ICCs of all items that contribute to the raw score. Using the notation introduced in Section 6.1, the expected raw score at a given value of θ_j is

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

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

The expected raw score monotonically increases with θ_j, consistent with the notion that students of high ability tend to earn higher raw scores than 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 the test provides at each value of θ_j. Information functions depict test precision across the entire latent trait continuum. There is an inverse relationship between the information of a test and its standard error of measurement (SEM). For long tests, the SEM at a given θ_j is approximately equal to the inverse of the square root of the statistical information at θ_j (Hambleton, Swaminathan, & Rogers, 1991), as follows:

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

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

Table 6-1 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 all cases, items flagged during this step were identified because of the pseudoguessing parameter (*c* parameter) being poorly estimated. Difficulty in estimating the *c* parameter is not at all unusual and is well-documented in the psychometric literature (see, for example, 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 and content area during the IRT analysis can be found in Table 6-2. The number of cycles required fell within acceptable ranges.

Table 6-2. 2011–12 NECAP: Number of Newton Cycles Required for Convergence

Content Area	*Grade*	*Cycles*
Mathematics	3	107
	4	31
	5	58
	6	55
	7	57
	8	56
	11	96
Reading	3	47
	4	56
	5	63
	6	47
	7	45
	8	44
	11	45

6.3. 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 2011–12 administration of NECAP 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 2011–12 NECAP reading 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 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 (that is, 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 2011–12 were placed on the 2010–11 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 2010–11 and 2011–12 NECAP tests should have the same item parameters. After the item parameters for each 2011–12 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 2011–12 TCC for the equating items was as close as possible to that of 2010–11.

6.4. EQUATING RESULTS

Prior to calculating the Stocking and Lord 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 6-1 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 J presents the results from the delta analysis. This procedure was used to evaluate adequacy of equating items; the discard status presented in the appendix indicates whether the item was flagged as potentially inappropriate for use in equating.

Also presented in Appendix J 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, 0.50 in absolute value.

Finally, *a*-plots and *b*-plots, which show IRT parameters for 2011–12 plotted against the values for 2010–11, are presented in Appendix K. 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 method of equating was used to place the item parameters onto the previous year's scale, as described above. The Stocking and Lord transformation constants are presented in Table 6-3.

Table 6-3. 2011–12 NECAP: Stocking and Lord Transformation Constants

Content Area	*Grade*	*a-slope*	*b-intercept*
Mathematics	3	1.037	0.152
	4	1.109	0.098
	5	1.052	0.137
	6	1.113	0.253
	7	1.049	0.269
	8	0.965	0.288
	11	1.014	0.182
Reading	3	0.922	0.052
	4	1.054	0.233
	5	0.979	0.091
	6	1.096	0.163
	7	1.119	0.098
	8	1.102	0.307
	11	1.106	0.358

The next administration of NECAP (2012–13) will be scaled to the 2011–12 administration using the same equating method described above.

6.5. ACHIEVEMENT STANDARDS

NECAP standards to establish achievement level cut scores in reading and mathematics for grades 3 through 8 were set in January 2006, and in reading, mathematics, and writing for grade 11 in January 2008. Details of the standard setting procedures can be found in the respective standard setting reports, as well as in the technical reports of those years.

Achievement standards for writing grades 5 and 8 were set in December 2010; for complete details of the standard setting, please see the *2010–11 New England Common Assessment Program Standard Setting Report* (Measured Progress, 2011).

The cuts on the theta scale that were established via standard setting and used for reporting in fall 2011 are presented in Table 6-4 below. Also shown in the table are the cutpoints on the reporting score scale (described below). These cutpoints will remain fixed throughout the assessment program unless standards are reset for any reason.

Table 6-4. 2011–12 NECAP: Cutpoints on the Theta Metric and Reporting Scale by Content Area and Grade

Content Area	*Grade*	*Theta*			*Scaled Score*				
		Cut 1	*Cut 2*	*Cut 3*	*Minimum*	*Cut 1*	*Cut 2*	*Cut 3*	*Maximum*
Mathematics	3	-1.03810	-0.26850	0.97040	300	332	340	353	380
	4	-1.15040	-0.37785	0.94930	400	431	440	455	480
	5	-0.92790	-0.28455	1.03130	500	533	540	554	580
	6	-0.87425	-0.22365	1.03425	600	633	640	653	680
	7	-0.70795	-0.07870	1.09945	700	734	740	752	780
	8	-0.64440	-0.02860	1.11775	800	834	840	852	880
	11	-0.11688	0.61895	2.05863	1100	1134	1140	1152	1180
Reading	3	-1.32285	-0.49700	1.03070	300	331	340	357	380
	4	-1.17300	-0.31420	1.14725	400	431	440	456	480
	5	-1.33545	-0.42760	1.04035	500	530	540	556	580
	6	-1.47795	-0.51795	1.12545	600	629	640	659	680
	7	-1.48330	-0.52230	1.20575	700	729	740	760	780
	8	-1.52505	-0.52240	1.13440	800	828	840	859	880
	11	-1.20706	-0.30992	1.00378	1100	1130	1140	1154	1180
Writing	5	-1.28347	-0.00868	1.52440	500	527	540	555	580
	8	-1.34864	-0.10590	1.26816	800	827	840	854	880

The tables in Appendix L show achievement level distributions by content area and grade. Results are shown for each of the last three years for all grades of reading and mathematics and for writing grade 11. For writing grades 5 and 8, because standards were set in December 2010, results are shown only for the 2010–11 and 2011–12 administrations.

6.6. Reported Scaled Scores

Because the θ scale used in IRT calibrations is not readily understood by most stakeholders, reporting scales were developed for NECAP. The reporting scales are simple linear transformations of the underlying θ scale. The reporting scales are developed such that they range from x00 through x80 (where x is grade level). In other words, grade 3 scaled scores ranged from 300 to 380, grade 4 from 400 through 480, and so forth through grade 11, where scores ranged from 1100 through 1180. The lowest scaled score in the Proficient range is fixed at x40 for each grade level. For example, to be classified in the Proficient achievement level or above, a minimum scaled score of 340 was required at grade 3, 440 at grade 4, and so forth.

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 2011–12 NECAP 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 2011–12 NECAP 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 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 and content areas. The raw cut score between Partially Proficient and Proficient could be, say, 35 in mathematics and 33 in reading, yet both of these raw scores would be transformed to scaled scores of x40. It is this uniformity across *scaled scores* that facilitates the understanding of student performance. The psychometric advantage of scaled scores over raw scores comes from their being *linear* transformations of θ. Since the θ scale is used for equating, scaled scores are comparable from one year to the next. Raw scores are not.

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

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

where
m is the slope, and
b is the intercept.

A separate linear transformation is used for each grade and content area combination. For NECAP, the transformation function is determined by fixing the Partially Proficient/Proficient cut score and the bottom of the scale—that is, the x40 and the x00 values (e.g., 440 and 400 for grade 4). The x00 location on the θ 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 θ scale. A raw score of 0 is also assigned a scaled score of x00. The maximum possible raw score is assigned a scaled score of x80 (e.g., 480 in the case of grade 4). Because only two points within the θ scaled score space are fixed, the scaled score cutpoints between Substantially Below Proficient and Partially Proficient and between Proficient and Proficient with Distinction are free to vary across the grade and content area combinations.

Table 6-5 shows the slope and intercept terms used to calculate the scaled scores for each content area and grade. Note that the values in Table 6-5 will not change unless the standards are reset.

Table 6-5. 2011–12 NECAP: Scaled Score Slope and Intercept by Content Area and Grade

Content Area	*Grade*	*Slope*	*Intercept*
Mathematics	3	10.71955	342.87820
	4	11.04316	444.17266
	5	10.76586	543.06342
	6	10.59224	642.36895
	7	10.20070	740.80279
	8	10.07201	840.28806
	11	8.65998	1134.63991
Reading	3	11.41878	345.67514
	4	10.85246	443.40984
	5	11.19695	544.78782
	6	11.48749	645.94994
	7	11.50185	746.00742
	8	11.50219	846.00874
	11	10.83987	1143.35949
Writing	5	10.02173	540.08694
	8	10.27195	841.08780

Appendix M contains raw score to scaled score look-up tables for the 2011–12 NECAP tests. These are the actual tables used to determine student scaled scores, error bands, and achievement levels.

Appendix N contains scaled score distribution graphs for each grade and content area. These distributions were calculated using the sparse data matrix files that were used in the IRT calibrations.

CHAPTER 7. 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 affect 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, 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), that eliminates the problem of the split-half method by comparing individual item variances to total test variance. Cronbach's α was used to assess the reliability of the 2011–12 NECAP:

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

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

7.1. Reliability and Standard Errors of Measurement

Table 7-1 presents descriptive statistics, Cronbach's α coefficient, and raw score standard errors of measurement (SEMs) for each grade and content area. (Statistics are based on common items only.) Note that reliability could not be calculated for grade 11 writing because the test consists of a single writing prompt.

Table 7-1. 2011–12 NECAP: Raw Score Descriptive Statistics, Cronbach's Alpha, and Standard Errors of Measurement (SEM) by Content Area and Grade

Content Area	*Grade*	*Number of Students*	*Raw score*			*Alpha*	*SEM*
			Maximum	*Mean*	*Standard Deviation*		
Mathematics	3	43,414	65	42.68	12.52	0.93	3.33
	4	43,806	65	42.36	13.49	0.93	3.46
	5	44,289	66	36.52	13.72	0.92	3.86
	6	44,130	66	34.48	14.62	0.92	4.05
	7	44,630	66	32.41	14.02	0.92	3.85
	8	46,432	66	34.12	14.56	0.93	3.89
	11	31,711	64	23.89	13.61	0.94	3.46
Reading	3	43,360	52	34.25	8.75	0.89	2.88
	4	43,741	52	34.16	9.12	0.89	3.05
	5	44,243	52	31.06	8.68	0.89	2.89
	6	44,117	52	33.04	8.31	0.88	2.83
	7	44,522	52	32.56	9.25	0.89	3.06
	8	46,376	52	32.71	9.21	0.89	3.09
	11	31,927	52	33.46	9.87	0.90	3.08
Writing	5	44,008	34	18.94	4.59	0.77	2.22
	8	46,171	34	20.83	5.58	0.77	2.65
	11	31,859	12	6.26	2.07		

Because different grades and content areas 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 and/or content area.

7.2. 2011–12 SUBGROUP RELIABILITY

The reliability coefficients discussed in the previous section were based on the overall population of students who took the 2011–12 NECAP test. Appendix O presents reliabilities for various subgroups of interest. Subgroup Cronbach's α's were calculated using the formula defined above based only on the members of the subgroup in question in the computations; values are only calculated 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 and content areas preclude making valid inferences about the quality of a test based on statistical comparisons with other tests. Second, reliabilities are dependent not only on the measurement properties of a test but on the statistical distribution of the studied subgroup. For example, it can be readily seen in Appendix O 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.

7.3. REPORTING SUBCATEGORY RELIABILITY

Of even more interest are reliabilities for the reporting subcategories within NECAP content areas, described in Chapter 2. Cronbach's α coefficients for subcategories were calculated via the same formula defined previously using just the items of a given subcategory in the computations. Results are presented in Appendix O. Once again as expected, because they are based on a subset of items rather than the full test, computed subcategory reliabilities were lower (sometimes substantially so) than were overall test reliabilities, and interpretations should take this into account. The subcategory reliabilities were lower than those based on the total test and approximately to the degree one would expect based on classical test theory. Qualitative differences between grades and content areas once again preclude valid inferences about the quality of the full test based on statistical comparisons among subcategories.

7.4. INTERRATER CONSISTENCY

Chapter 4 of this report describes in detail the processes that were implemented to monitor the quality of the hand-scoring of student responses for constructed-response items. One of these processes was double-blind scoring: approximately 2% of student responses were randomly selected and scored independently by two different scorers, except for the writing prompts in grades 5, 8, and 11, which were 100% double scored. 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 NECAP. A summary of the interrater consistency results are presented in Table 7-2 below. Results in the table are collapsed across the

hand-scored items by grade and content area. 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 percentage of responses that required a third score. This same information is provided at the item level in Appendix P.

Table 7-2. 2011–12 NECAP: Summary of Interrater Consistency Statistics Collapsed Across Items by Content Area and Grade

Content Area	*Grade*	*Number of Score Categories*	*Number of Included Scores*	*Percent Exact*	*Percent Adjacent*	*Correlation*	*Percent of Third Scores*
Mathematics	3	2	8,646	97.68	2.32	0.94	0.00
		3	8,671	94.52	5.35	0.96	0.13
	4	2	8,790	96.67	3.33	0.93	0.00
		3	8,847	95.94	3.82	0.96	0.24
	5	2	5,332	98.14	1.86	0.96	0.00
		3	5,291	90.38	8.96	0.91	0.66
		5	3,622	88.18	10.96	0.96	0.86
	6	2	5,291	97.85	2.15	0.95	0.00
		3	5,358	88.67	10.49	0.90	0.84
		5	3,577	89.24	9.87	0.97	0.89
	7	2	5,314	97.48	2.52	0.95	0.00
		3	5,285	90.61	9.04	0.93	0.34
		5	3,597	84.65	13.12	0.93	2.20
	8	2	5,551	95.30	4.70	0.91	0.00
		3	5,546	91.74	7.86	0.94	0.38
		5	3,791	84.12	13.95	0.94	1.90
	11	2	7,254	98.70	1.30	0.97	0.00
		3	3,481	95.35	4.40	0.95	0.26
		5	2,473	89.45	9.38	0.94	1.13
Reading	3	5	5,318	79.84	18.90	0.93	1.24
	4	5	5,276	76.50	21.95	0.89	1.48
	5	5	5,457	64.39	33.10	0.76	2.40
	6	5	5,524	63.83	33.94	0.74	2.19
	7	5	5,572	63.98	33.54	0.80	2.44
	8	5	5,726	61.33	35.78	0.79	2.72
	11	5	3,916	61.87	35.83	0.79	2.20
Writing	5	5	2,713	65.09	33.62	0.75	1.29
		7	43,450	61.83	36.29	0.67	1.87
	8	5	2,813	65.41	32.99	0.77	1.60
		7	45,097	57.30	38.38	0.73	4.26
	11	7	31,150	59.45	38.32	0.79	2.09

7.5. 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, 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, and results are given.

Accuracy refers to the extent to which decisions based on test scores match decisions that would have been made if the scores did not contain any measurement error. Accuracy must be estimated, because errorless test scores do not exist. Consistency measures the extent to which classification decisions based on test scores match the decisions based on scores from a second, parallel form of the same test. Consistency can be evaluated directly from actual responses to test items if two complete and parallel forms of the test are given to the same group of students. In operational test programs, however, such a design is usually impractical. Instead, techniques have been developed to estimate both the accuracy and consistency of classification decisions based on a single administration of a test. The Livingston and Lewis (1995) technique was used for the 2011–12 NECAP because it is easily adaptable to all types of testing formats, including mixed format tests.

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

For the 2011–12 NECAP, after various technical adjustments (described in Livingston & Lewis, 1995), a four-by-four contingency table of accuracy was created for each grade and content area, 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 content area 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:

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

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;

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.

7.5.1. Accuracy and Consistency Results

The accuracy and consistency analyses described above are provided in Table Q-1 of Appendix Q. 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 0.84 for Substantially Below Proficient for grade 3 mathematics. This figure indicates that among the students whose true scores placed them in this classification, 84% would be expected to be in this classification when categorized according to their observed scores. Similarly, a consistency value of 0.77 indicates that 77% 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 NCLB 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 the 2011–12 NECAP, Table Q-2 in Appendix Q 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 cutpoint and whose true scores were below the cutpoint. A false negative is the proportion of students whose observed scores were below the cutpoint and whose true scores were above the cutpoint.)

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, accuracy and consistency statistics calculated based on small groups can be expected to be lower than those calculated based on larger groups. For this reason, the values presented in Appendix Q should be interpreted with caution. In addition, it is important to remember that it is inappropriate to compare accuracy and consistency statistics between grades and content areas.

CHAPTER 8. SCORE REPORTING

8.1. TEACHING YEAR VERSUS TESTING YEAR REPORTING

The data used for the NECAP reports are the results of the fall 2011 NECAP test administration. It is important to note that the NECAP tests are based on the grade level expectations (GLEs) from the previous year. For example, the grade 7 NECAP test administered in the fall of seventh grade is based on the grade 6 GLEs. Because many students receive instruction at a different school from where they were tested, the state Departments of Education determined that access to results information would be valuable to both the school where the student was tested and the school where the student received instruction. To achieve this goal, separate Item Analysis, School and District Results, and School and District Summary Reports were created for the "testing" school and the "teaching" school. Every student who participated in the NECAP test was represented in testing reports, and most students were represented in teaching reports. In some cases (e.g., a student who recently moved to the state), it is not possible to provide information for a student in a teaching report.

8.2. PRIMARY REPORTING DELIVERABLES

The following reporting deliverables were produced for the 2011–12 NECAP:

- Student Report
- School and District Results Report
- School and District Summary Report
- School and District Student-Level Data File
- Analysis & Reporting System

With the exception of the Student Report, these reports and data files were available for schools and districts to view or download via the NECAP Analysis & Reporting System, a password-secure Web site hosted by Measured Progress. Each of these reporting deliverables is described in the following sections. Sample reports are provided in Appendix R.

Support is provided by the state Departments of Education and Measured Progress to stakeholders who use the various reporting deliverables by hosting report interpretation workshops and by providing the *Guide to Using the 2011 NECAP Reports*. These resources help foster proper use and interpretation of NECAP results.

The *Guide* includes a table that shows the number of scaled score points that would indicate a statistically significant difference between two equally sized groups of students. The calculations are

performed by computing the standard error of the difference in means ($\sigma_{\overline{x}_1-\overline{x}_2}$) for different values of n, based on the observed scaled score standard deviations for each grade and content area. The formula for the variance error of the difference in means is:

$$\sigma^2_{\overline{x}_1-x_2} = \sigma^2_2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$$

By assuming $n_1 = n_2 = n$ and $\sigma_1^2 = \sigma_2^2 = \sigma_w^2$, this equation simplifies to:

$$\sigma^2_{\overline{x}_1-x_2} = \sigma_w^2\left(\frac{1}{n}+\frac{1}{n}\right) = \frac{2\sigma^2}{n}, \text{ and}$$

$$\sigma_{\overline{x}_1-x_2} = \sqrt{2\sigma^2/n}$$

Therefore, if the difference in scaled scores of two equally sized groups is greater than or equal to $\sigma_{\overline{x}_1-x_2}$, you can be 67% certain that there is a true difference in performance between the two groups. Differences between two unequally sized groups can be interpreted, conservatively, by using the value associated with the size of the smaller group.

The *Guide* also includes a second table that shows corresponding values based on percentages of students, to help interpret differences in percentages of students in performance level categories. The calculations for this table are based on the variance error of a proportion:

$$s_p^2 = \frac{S^2}{n} = \frac{p(1-p)}{n}, \text{ and}$$

$$s_p = \sqrt{p(1-p)/n}$$

Together, these two tables in the *Guide to Using the 2011 NECAP Reports* help teachers, schools, and districts interpret differences in scores between two groups of students and, in this way, support appropriate interpretation of NECAP scores.

8.3. STUDENT REPORTS

The *NECAP Student Report* is a single-page, double-sided report printed on 8.5”-by-14” paper. The front of the report includes informational text about the design and uses of the assessment. The front of the report also contains text describing the three corresponding sections on the reverse side of the report and the achievement level descriptions. The reverse side of the *Student Report* provides a complete picture of an

individual student's performance on the NECAP test, divided into three sections. The first section provides the student's overall performance for each content area. The student's achievement levels are provided, and scaled scores are presented numerically as well as in a graphic that depicts the scaled score with the standard error of measurement bar constructed about it, set within the full range of possible scaled scores demarcated into the four achievement levels.

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

The third section shows the student's raw score performance in content area reporting categories relative to possible points; gives the average points earned for the school, district, and state; and gives the average points earned by students at the Proficient level on the overall content area test. For reading, with the exception of Word ID/Vocabulary items, items are reported by Type of Text (Literary, Informational) and Level of Comprehension (Initial Understanding, Analysis and Interpretation). For mathematics, the reporting subcategories are Numbers and Operations; Geometry and Measurement; Functions and Algebra; and Data, Statistics, and Probability. Grade 5 and 8 writing report Multiple Choice, Short Responses, and Extended Response as categories. Grade 11 writing only reports Extended Response as a category.

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.

8.4. ITEM ANALYSIS REPORTS

The *NECAP Item Analysis Report* provides a roster of all students in a school and provides their performance on the common items that are released to the public, one report per content area. For all grades and content areas, the student names and identification numbers are listed as row headers down the left side of the report. For grades 3 through 8 and 11 in reading and mathematics, the items are listed as column headers in the same order they appeared in the released item documents (not the position in which they appeared on the test).

For each item, seven pieces of information are shown: the released item number, the content strand for the item, the GLE/GSE code for the item, the depth of knowledge (DOK) code for the item, the item type, the correct response key for multiple-choice items, and the total possible points.

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 points possible. 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 why the student did not test. Descriptions of these codes can be found on the legend, after the last page of data on the report. It is important to note that not all items used to compute student scores are included in this report, only released items. At the bottom of the report, the average percent 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.

For grade 11 writing, the top portion of the *NECAP Item Analysis Report* consists of a single row of item information containing the content strand, GSE codes, DOK code, item type/writing prompt, and total possible points. The student names and identification numbers are listed as row headers down the left side of the report. The Total Test Results section to the right includes Total Points Earned and Achievement Level for each student. At the bottom, the average points earned on the writing prompt are provided for the school, district, and state.

The *NECAP Item Analysis Report* 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.

8.5. School and District Results Reports

The *NECAP School Results Report* and the *NECAP District Results Report* consist of three parts: the grade level summary report, the results for the content areas, and the disaggregated content area results.

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 October 1, 2011. 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 Student 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 Student 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 test, actual scores were reported for the content areas in which they participated. These reporting decisions were made to support the requirement that all students participate in the NECAP.

Data are provided for the following groups of students who may not have completed the entire battery of NECAP tests:

- **Alternate Assessment**: Students in this category completed an alternate test for the 2010–11 school year.
- **First-Year LEP**: Students in this category are defined as being new to the United States after October 1, 2010, and were not required to take the NECAP tests in reading and writing. Students in this category were expected to take the mathematics portion of the NECAP.
- **Withdrew after October 1**: Students withdrawing from a school after October 1, 2011, may have taken some sessions of the NECAP tests prior to their withdrawal from the school.
- **Enrolled after October 1**: Students enrolling in a school after October 1, 2011, may not have had adequate time to participate fully in all sessions of NECAP testing.
- **Special Consideration**: Schools received state approval for special consideration for an exemption on all or part of the NECAP tests for any student whose circumstances are not described by the previous categories but for whom the school determined that taking the NECAP tests would not be possible.
- **Other**: Occasionally students will not have completed the NECAP tests for reasons other than those listed above. These "other" categories were considered not state approved.

The results section in the bottom half of the page shows the number and percentage of students performing at each achievement level in each of the content areas across the school, district, and state. In addition, a mean scaled score is provided for each content area across school, district, and state levels except for grade 11 writing where the mean raw score is provided across the school, district, and state. School information is blank for the district version of this report.

For reading and mathematics, the content area results pages provide information on performance in specific content categories of the tested content areas (for example, geometry and measurement within mathematics). For writing in grades 5 and 8, information is provided by item type (multiple-choice, short-response, and extended-response). The purpose of these sections is to help schools determine the extent to which their curricula are effective in helping students to achieve the particular standards and benchmarks contained in the GLEs and GSEs. The content area results pages provide data for the 2009–10, 2010–11, and 2011–12 individual test administrations as well as cumulative data for the three years in reading and mathematics. For writing grades 5 and 8, data are only provided for the 2010–11 and the 2011–12 test

administrations as well as cumulative data for the two years. Data do not exist for the 2009–10 test administration for writing in grades 5 and 8 because the test was a pilot and results were not produced.

Information about each content area (reading and mathematics for all grades and writing for grades 5 and 8) for school, district, and state levels include:

- the total number of students enrolled, not tested (state-approved reason), not tested (other 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 content area reporting category for reading and mathematics in all grades and about each item type for writing in grades 5 and 8 includes the following:

- The total possible points for that reporting category or item type. In order to provide as much information as possible for each category, the total number of points includes both the common items used to calculate scores and additional items in each category used for equating the test from year to year.
- A graphic display of the percentage of total possible points for each subcontent area/reporting category for the school, district, and state. In this graphic display, there are symbols representing school, district, and state performances. In addition, there is a line representing the standard error of measurement. 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 first content area results page for the grade 11 writing test provides data for the 2009–10, 2010–11, and 2011–12 individual test administrations as well as cumulative data for the three years. Information provided for the school, district, and state includes:

- the total number of students enrolled, not tested (state-approved reason), not tested (other 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 raw score.

The bottom half of this page includes a table that lists the type of writing (genre) for the common prompt (i.e., the prompt for which the results on the top half of the page are about) for each of the last three test administrations. The type of writing and a description of that type is included for each year.

The second content area results page lists the types of writing that are assessed in the grade 11 writing test. The list is comprised of the types of writing used to make both a common prompt (one that is administered to all students) and matrix prompts (ones that vary across the eight different forms of the test). The first column on this page provides the name and a description of each type of writing. The second column provides separate rows for the current year (2011–12) and the previous year that each type of writing was assessed. The symbol "(C)" indicates the type of writing that was common in that year's test. The number of students tested and the mean raw score are provided for the school, district, and state levels. A graphic display is also provided for each year and type of writing that shows the average score attained on the 0-to-12 scale for the school, district, and state levels. The scale of 0 to 12 on the graphic display represents the possible score range for the writing prompt. The two scores assigned to the student's response from the 6-point rubric are added; the result is a score from 0 to 12. A score of 7 depicted on the scale represents the score needed to be proficient.

Finally, the third page of the grade 11 writing content area results contains a table that presents information on the distribution of scores across the 0-to-12 range. The first column of the table lists the possible scores from 12 down to 0. The next two columns ("Score 1" and "Score 2") represent the two independent scores assigned to a student's response to the common writing prompt. The student's total score on the common writing prompt is the sum of these two scores. The next four columns list the total number of students, N, and the percentage, %, of students for each score on the 0-to-12 scale for the school and district. The last column provides the percentage, %, of students for each score on the 0-to-12 scale for the state. The 6-point scoring rubric that is used to score student responses to the common writing prompt is also included on this page of the report.

The disaggregated content area results pages (all grades and content areas) present the relationship between performance and student reporting categories (see list below) in each content area across school, district, and state levels. Each content area page shows the number of students categorized as enrolled, not tested (state-approved reason), not tested (other reason), and tested. The tables also provide the number and percentage of students within each of the four achievement levels and the mean scaled score (or mean raw score for grade 11 writing) by each reporting category.

The list of student reporting categories is as follows:

- All Students
- Gender
- Race/Ethnicity
- LEP Status
- IEP
- SES (socioeconomic status)
- Migrant

- Title I
- 504 Plan

The data for achievement levels and mean scaled score (or mean raw score for grade 11 writing) are based on the number shown in the tested column. The data for the reporting categories were provided by information coded on the students' answer booklets by teachers and/or data linked to the student label. 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 no data were reported for the 504 Plan in any of the content areas for New Hampshire and Vermont. Additionally, no data were reported for Title I in any of the content areas for Vermont.

8.6. School and District Summary Reports

The *NECAP School Summary Report* and the *NECAP District Summary Report* provide details, broken down by content area, on student performance by grade level tested in the school. The purpose of the summary is to help schools determine the extent to which their students achieve the particular standards and benchmarks contained in the GLEs and GSEs.

Information about each content area and grade level for school, district, and state includes:

- the total number of students enrolled, not tested (state-approved reason), not tested (other 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 (mean raw score for grade 11 writing).

The data reported, the report format, and the guidelines for using the reported data are identical for both the school and district reports. The only difference between the reports is that the *NECAP District Summary Report* includes no individual school data. Separate school reports and district reports were produced for each grade level tested.

8.7. School and District Student-Level Data Files

In addition to the reports described above, districts and, for the first time this year, schools received access to and were able to download student-level data files from the Analysis & Reporting System for each grade of students tested within their district or school. Student-level data files were produced for both "teaching year" and "testing year."

The student-level data files list students alphabetically within each school and contain all of the demographic information that was provided by the state for each student. Student records contain the scaled score, achievement level, and subscores earned by the student for each content area tested. In addition, the student records contain each student's actual performance on each of the released items for each content area tested as well as the student's responses to the student questionnaire.

The data collected from the optional reports field, if it was coded by schools on page 2 of the Student Answer Booklets, are also available for each student in the student-level data file. The optional reports field was provided to allow schools the option of grouping individual students into additional categories (e.g., by class or by previous year's teacher). This allows schools to make comparisons between subgroups that are not already listed on the disaggregated results pages of the school and district results reports.

The file layout of the student-level data files that lists all of the field names, variable information, and valid values for each field was also available to districts and schools on the Analysis & Reporting System.

8.8. ANALYSIS & REPORTING SYSTEM

NECAP results for the 2011–12 test administration were accessible online via the Analysis & Reporting System. In addition to accessing and downloading reports and student-level data files in the same manner as in previous years, this new system includes interactive capabilities that allow school and district users to sort and filter item and subgroup data to create custom reports.

8.8.1. Interactive Reports

There are four interactive reports that were available from the Analysis & Reporting System: "Item Analysis Report," "Achievement Level Summary," "Released Items Summary Data," and "Longitudinal Data ." Each of these interactive reports is described in the following sections. To access these four interactive reports, the user needed to click the "Interactive" tab on the home page of the system and select the report desired from the drop-down menu. Next, the user had to apply 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 for which to analyze their results. Advanced filtering options included gender, ethnicity, LEP, IEP, and SES. Users also needed to select either the "Teaching" or "Testing" cohort of students using the "Filter by Group" drop-down menu. All interactive reports, with the exception of the Longitudinal Data Report, allowed the user to provide a custom title for the report.

8.8.1.1. Item Analysis Report

The Item Analysis Report provides individual student performance data on the released items and total test results for a selected grade and content area. . Please note that when advanced filtering criteria are applied by the user, the School and District Percent Correct and 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.

8.8.1.2. Achievement Level Summary

The Achievement Level Summary provides a visual display of the percentages of students in each achievement level for a selected grade and content area. The four achievement levels (Proficient with Distinction, Proficient, Partially Proficient, and Substantially Below Proficient) 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 or JPG file.

8.8.1.3. Released Items Summary Data

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 and content area. The report is divided into two sections by item type (multiple-choice and open-response). For multiple-choice items, the content strand and GE code linked to the item are included as well as the total number and percentage of students who answered the item correctly and the number of students who chose each incorrect option or provided an invalid response. 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 open-response items, the content strand and GE code linked to the item are included as well as the point value and average score for the item. Users are also able to view the actual released items within the 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.

8.8.1.4. Longitudinal Data

The Longitudinal Data report is a confidential student-level report that provides individual student performance data for multiple test administrations. Fall 2011 NECAP scores and achievement levels are provided for each tested student in reading, mathematics, and writing. In addition, fall NECAP 2008, 2009, and 2010 reading, mathematics, and writing scores and achievement levels as well as spring NECAP science scores and achievement levels are also included for students in New Hampshire, Rhode Island, and Vermont. Maine students in grades 3 through 8 will show fall 2009, 2010, and 2011 NECAP scores and achievement levels in reading and mathematics, since this is only the third test administration for Maine since joining NECAP. 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.

8.8.2. User Accounts

In the Analysis & Reporting System, principals have the ability to create unique user accounts by assigning specific usernames and passwords to educators, such as teachers, curriculum coordinators, or special education coordinators, in their school. Once the accounts have been created, individual students may be assigned to each user account. After users have received their usernames and passwords, they are able to log in to their accounts and access the interactive reports, which will be populated only with the subgroup of students assigned to them.

Information about the interactive reports and setting up user accounts is available in the *Analysis & Reporting System User Manual* that is available for download on the Analysis & Reporting System.

8.9. DECISION RULES

To ensure that reported results for the 2011–12 NECAP 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 NECAP test data and in reporting the test results. Moreover, these rules are the main reference for quality assurance checks.

The decision rules document used for reporting results of the October 2011 administration of the NECAP is found in Appendix S.

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.

8.9.1.1. 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 NECAP implement quality control checks of their respective computer programs and intermediate products. Moreover, when data are handed off to different functions within the Data and Reporting and Psychometrics departments, the sending function verifies that the data are accurate before handoff. Additionally, when a function receives a data set, the first step is to verify the data for accuracy.

Another type of quality assurance measure is parallel processing. Students' scaled scores for each content area are assigned by a psychometrician through a process of equating and scaling. The scaled scores are also computed by a data analyst to verify that scaled scores and corresponding achievement levels are

assigned accurately. Respective scaled scores and assigned achievement levels are compared across all students for 100% agreement. Different exclusions that determine whether each student received scaled scores and/or was 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 content area and grade combination, the exclusions assigned by each data analyst are compared across all students. Only when 100% agreement is achieved can the rest of data analysis be completed.

The third aspect of quality control involves the procedures implemented by the quality assurance group to check the 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 reports. The selection of sample schools and districts for this purpose is very specific and can affect the success of the quality control efforts. There are two sets of samples selected that may not be mutually exclusive.

The first set includes those that satisfy the following criteria:

- one-school district
- two-school district
- multi-school district

The second set of samples includes districts or schools that have unique reporting situations as indicated by decision rules. This second set is necessary to ensure 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 with schools that are too small to receive a school report
- school with excluded (not tested) students
- school with home-schooled 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 sign-off.

CHAPTER 9. VALIDITY

Because interpretations of test scores, and not a test itself, are evaluated for validity, the purpose of the *2011–12 NECAP Technical Report* is to describe several technical aspects of the NECAP tests in support of score interpretations (AERA, 1999). 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., 1999) 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 aspect of validity 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 and content area. 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 this lens provided by the standards, evidence based on test content was extensively described in Chapters 2 and 3. Item alignment with NECAP 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 questions are aligned by educators from the member states to specific NECAP content standards, and undergo several rounds of review for content fidelity and appropriateness. Items are presented to students in multiple formats (constructed-response, short-answer, 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* and *Test Administrator Manuals*.

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

Evidence based on internal structure is presented in great detail in the discussions of item analyses, reliability, and scaling and equating in Chapters 5 through 7. 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 analyses, dimensionality analyses, reliability, standard errors of measurement, and item response theory parameters and procedures. Each test is equated to the same grade and content area

test from the prior year in order 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 6 and the reporting information in Chapter 8, as well as in the *Guide to Using the 2010 NECAP Reports*, which is a separate document. Each of these chapters speaks to the efforts undertaken to promote accurate and clear information provided to the public regarding test scores. Scaled scores offer the advantage of simplifying the reporting of results across content areas, grade levels, and subsequent years. Achievement levels provide users with reference points for mastery at each grade and content area, 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 affect 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 results to other variables, including the extent to which scores from NECAP 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.

9.1. QUESTIONNAIRE DATA

External validity of the NECAP assessment is conveyed by the relationship of test scores and situational variables such as time-spent patterns and attitude toward content matter. These situational variables were all based on student questionnaire data collected during the administration of the NECAP 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.

9.1.1. Difficulty of Assessment

Examinees in all grades and content areas were asked how difficult the test was relative to their regular schoolwork. In the sections below, results are presented for selected grade levels for each content area.

9.1.1.1. Difficulty: Reading

Figures 9-1 and 9-2 below show that students in grades 8 and 11 who thought the test was easier than their regular reading schoolwork did better overall than those who thought it was more difficult.

Question: How difficult was the reading test?

Figure 9-1. 2011–12 NECAP: Reading Grade 8 Questionnaire Responses—Difficulty

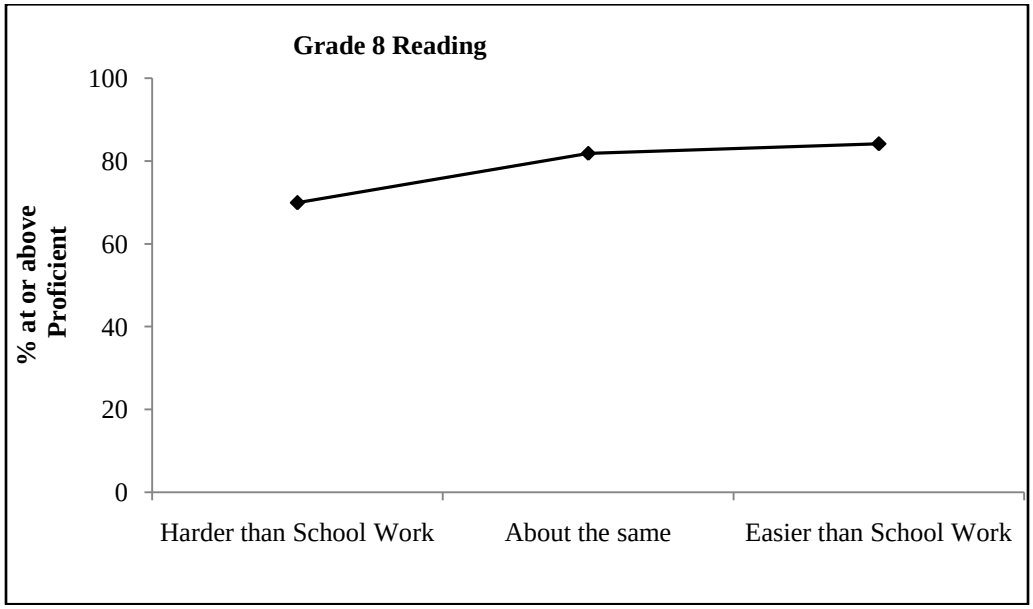
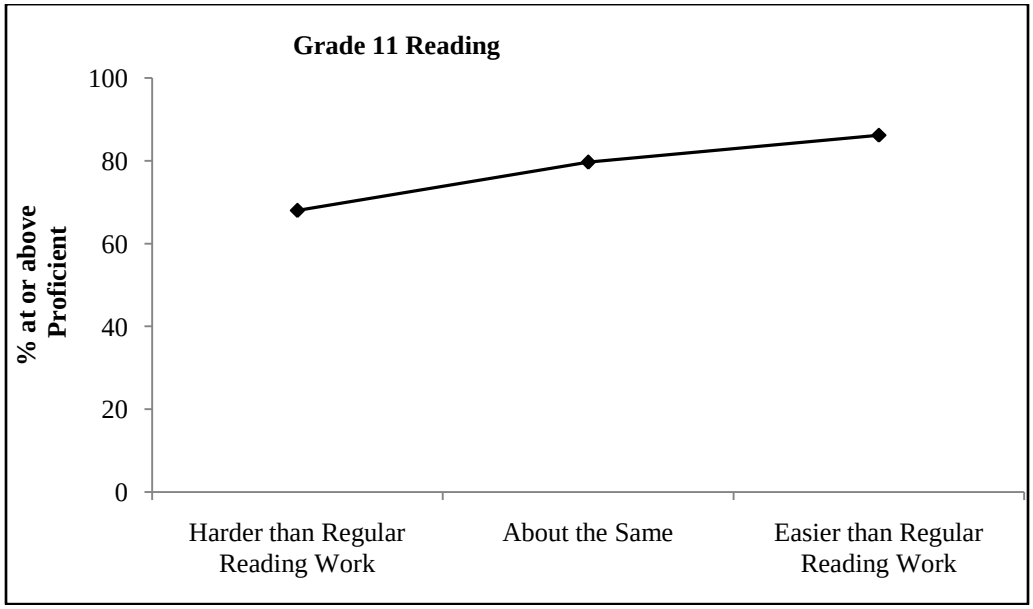


Figure 9-2. 2011–12 NECAP: Reading Grade 11 Questionnaire Responses—Difficulty



9.1.1.2. Difficulty: Mathematics

Figures 9-3 and 9-4 below show a very similar pattern to that for reading: students in grades 8 and 11 who thought the test was easier than their regular mathematics schoolwork did better overall than those who thought it was more difficult.

Question: How difficult was the mathematics test?

Figure 9-3. 2011–12 NECAP: Mathematics Grade 8 Questionnaire Responses—Difficulty

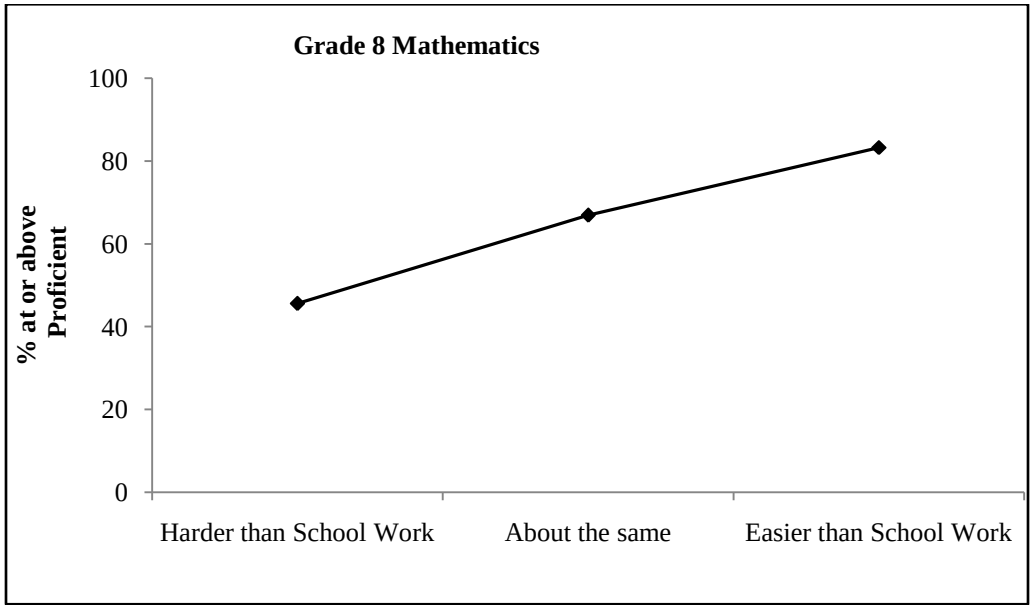
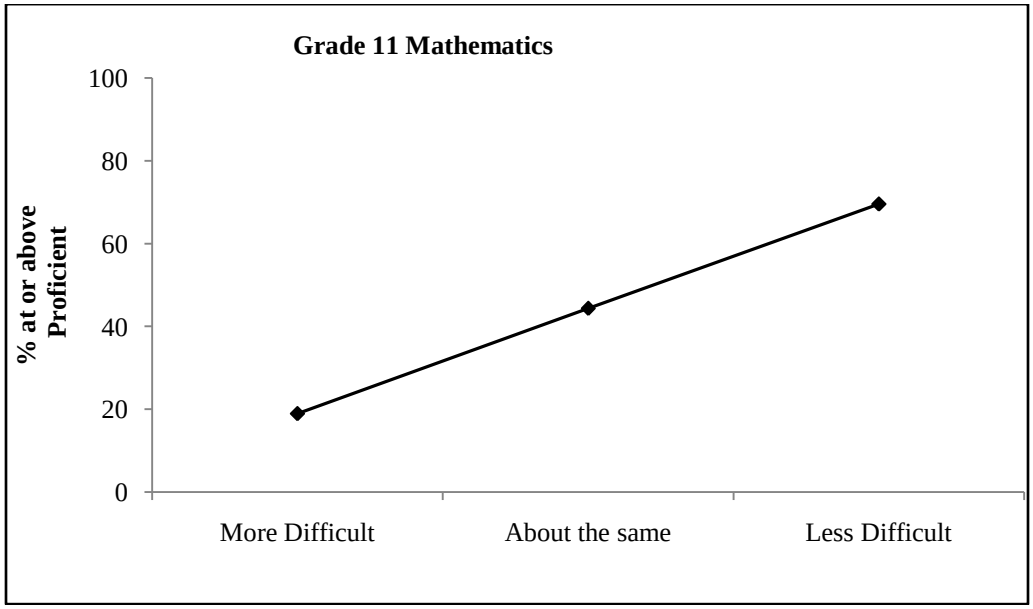


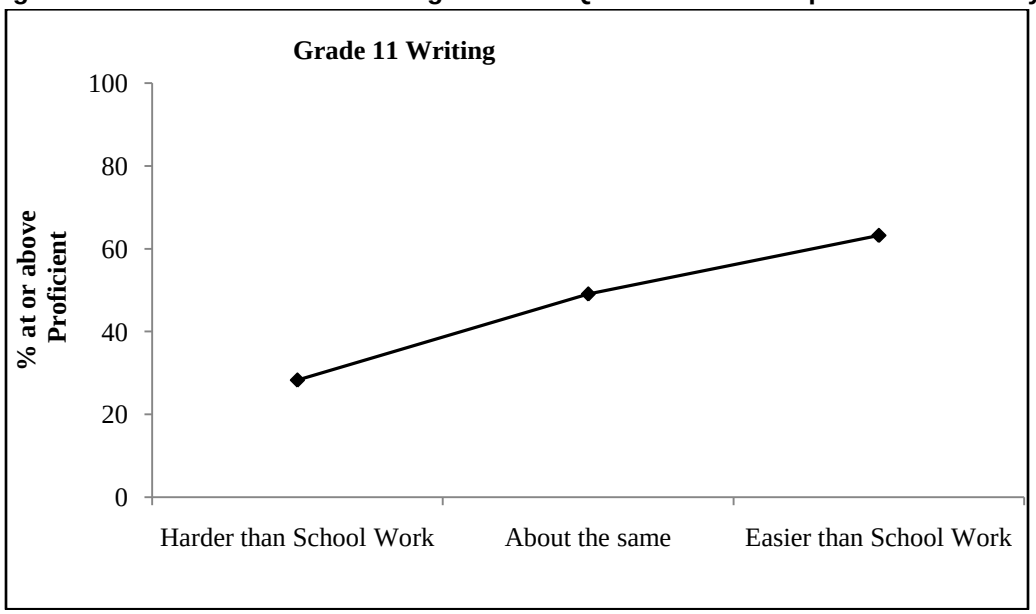
Figure 9-4. 2011–12 NECAP: Mathematics Grade 11 Questionnaire Responses—Difficulty



9.1.1.3. Difficulty: Writing

For writing, as shown in Figure 9-5 below, there was a pronounced relationship between perception of the difficulty of the test and student performance at grade 11.

Figure 9-5. 2011–12 NECAP: Writing Grade 11 Questionnaire Responses—Difficulty



9.1.2. Content

Across grades, examinees were asked about the frequency with which they engage in academic activities (specific to content area) that are expected to be related to test performance. In the sections below, results are presented for selected grade levels for each content area.

9.1.2.1. Content: Reading

Examinees in reading were asked how often they are asked to write at least one paragraph for reading/English language arts (grades 3 through 8) or reading (grade 11) class. Figures 9-6 through 9-9 show that students who indicated they write at least one paragraph a few times a week perform better than any of the other groups.

Figure 9-6. 2011–12 NECAP: Grade 3 Reading Questionnaire Responses—Content

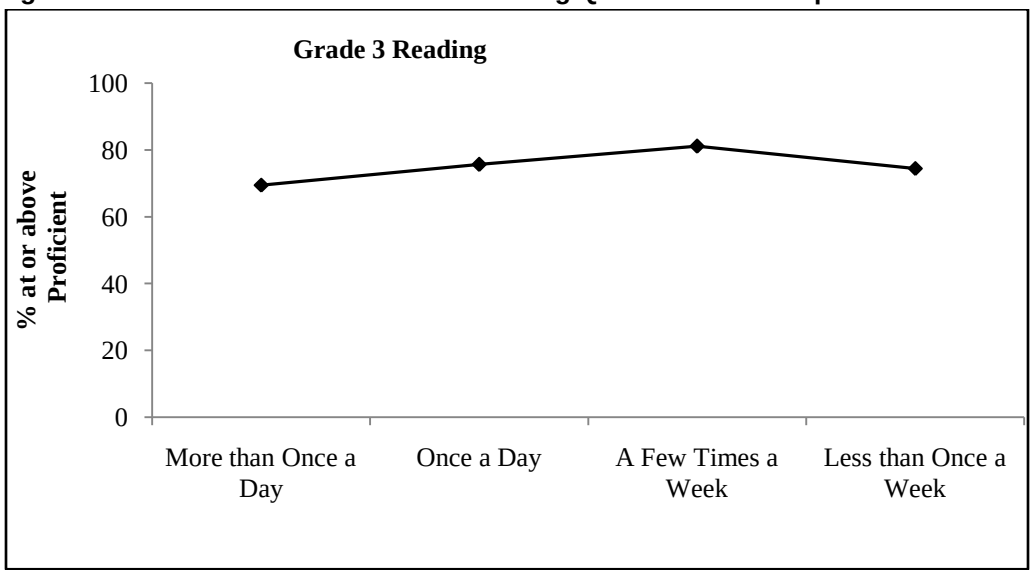


Figure 9-7. 2011–12 NECAP: Grade 4 Reading Questionnaire Responses—Content

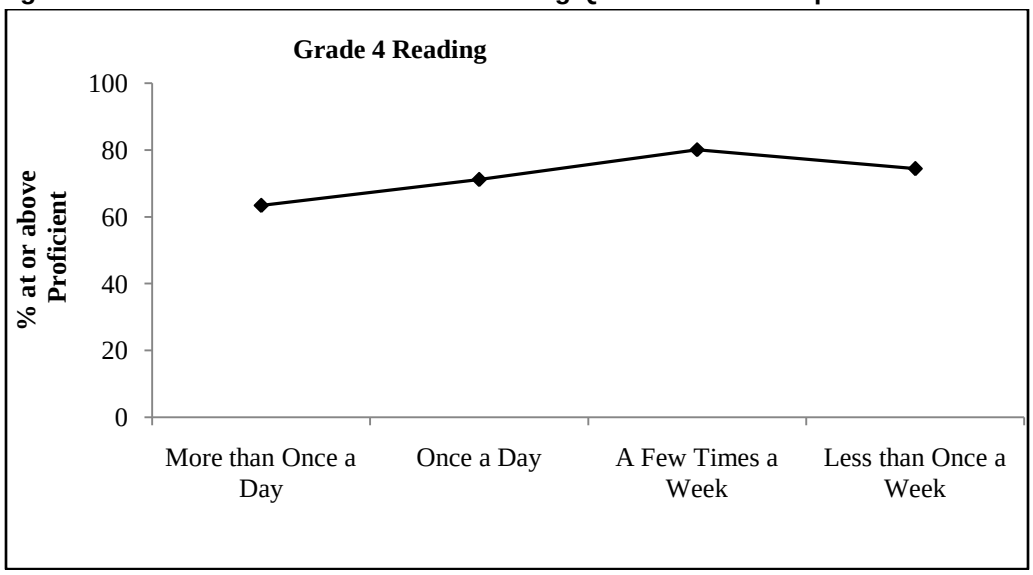


Figure 9-8. 2011–12 NECAP: Grade 7 Reading Questionnaire Responses—Content

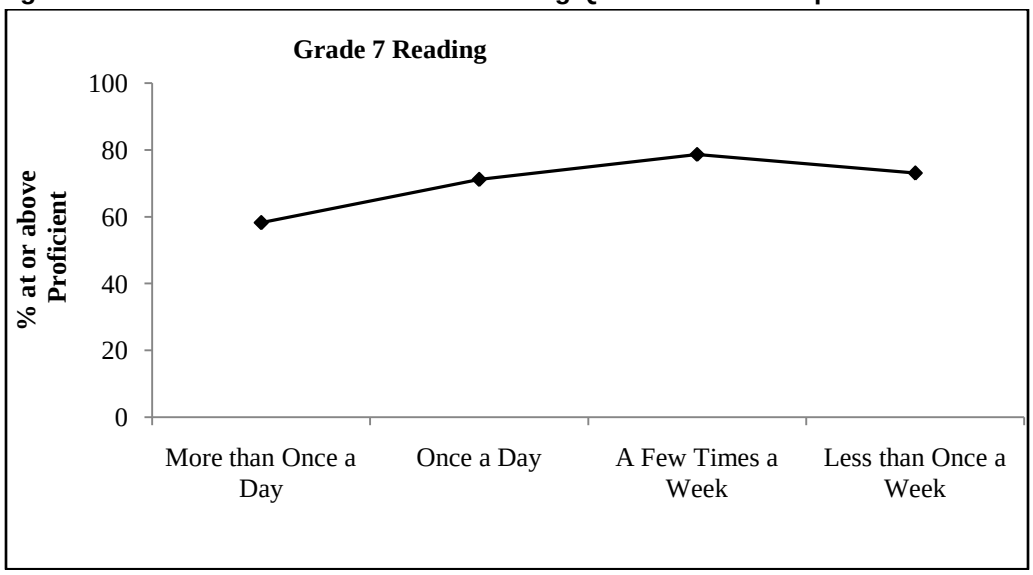
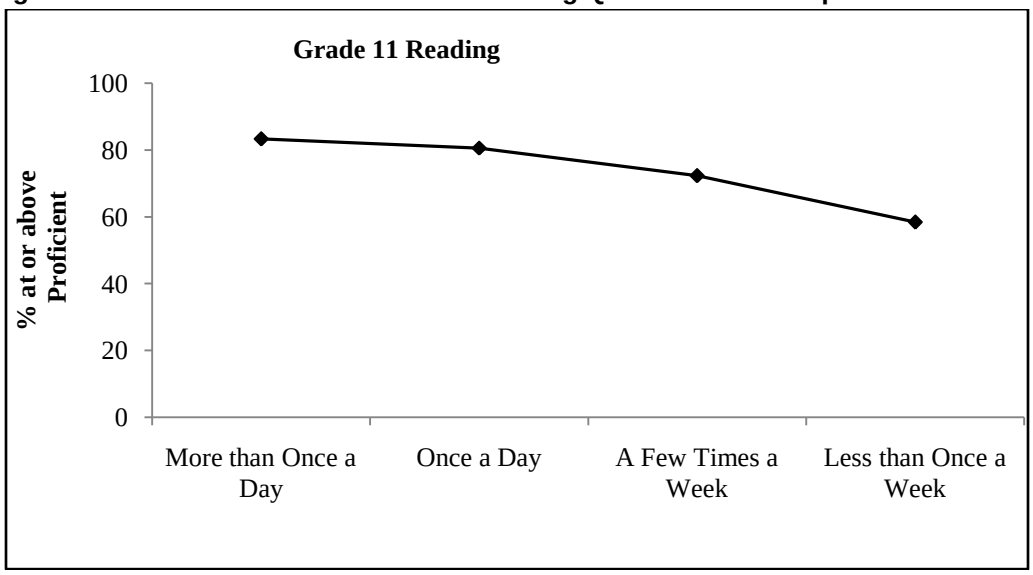


Figure 9-9. 2011–12 NECAP: Grade 11 Reading Questionnaire Responses—Content



9.1.2.2. Content: Mathematics

For mathematics, examinees in grades 7 and 8 were asked whether they were currently enrolled in an Algebra I or higher mathematics class. In grade 11, examinees were asked which mathematics course they last completed (e.g., Geometry). Figures 9-10 through 9-12 seem to suggest that students with more exposure to mathematics coursework tend to perform better than students who have been exposed to fewer mathematics courses.

Figure 9-10. 2011–12 NECAP: Grade 7 Mathematics Questionnaire Responses—Content

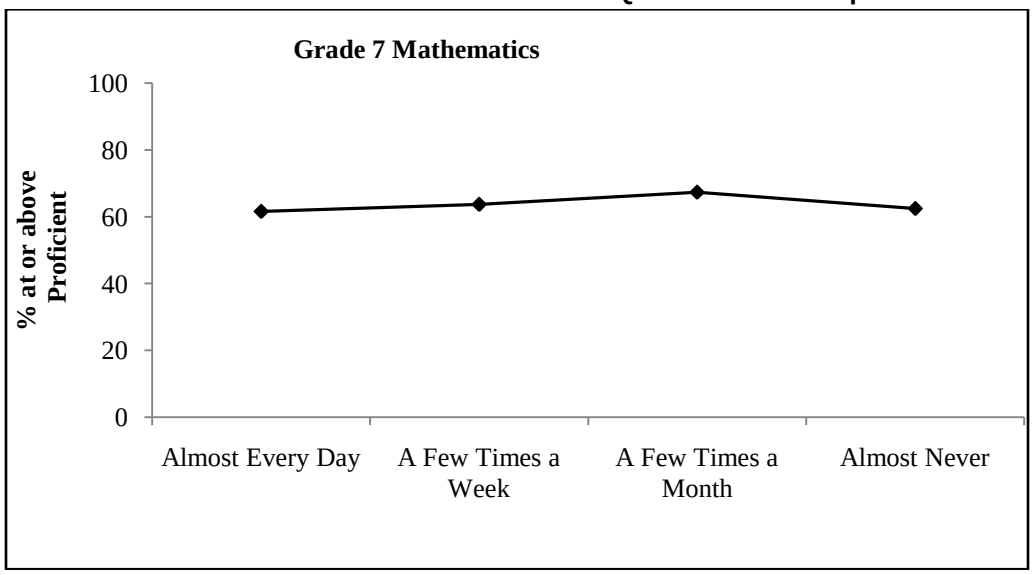


Figure 9-11. 2011–12 NECAP: Grade 8 Mathematics Questionnaire Responses—Content

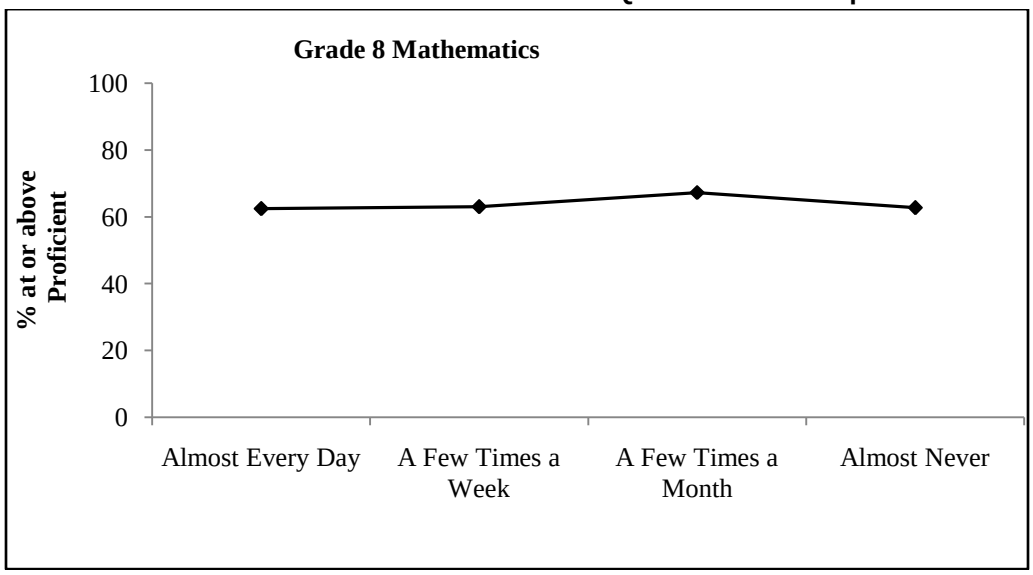
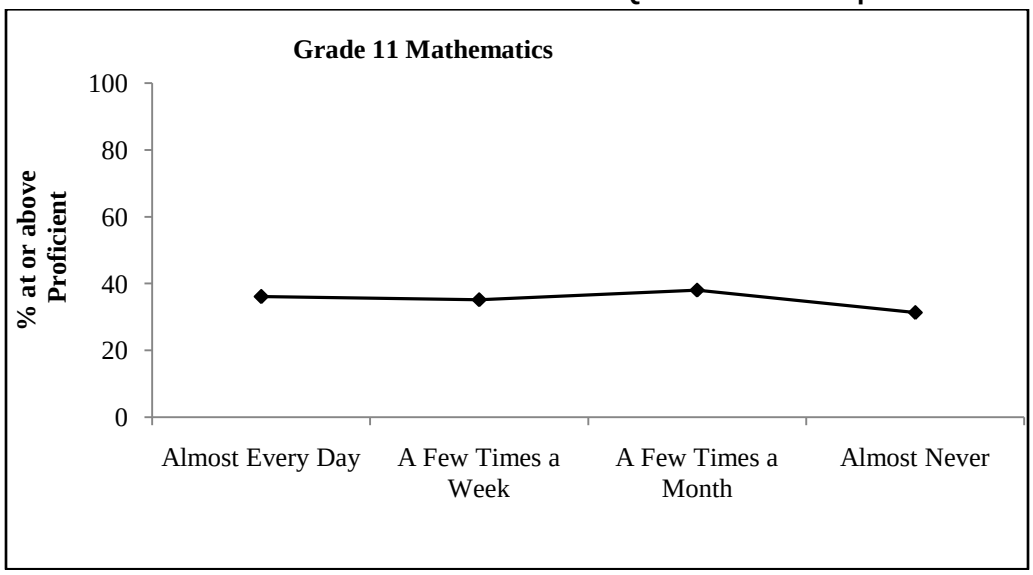


Figure 9-12. 2011–12 NECAP: Grade 11 Mathematics Questionnaire Responses—Content



9.1.2.3. Content: Writing

Examinees in writing were asked how often they write more than one draft. Figures 9-13 through 9-15 show that students who indicated they write multiple drafts more frequently did better than students who write multiple drafts less frequently, although the differences at grade 5 were slight.

Figure 9-13. 2011–12 NECAP: Grade 5 Writing Questionnaire Responses—Content

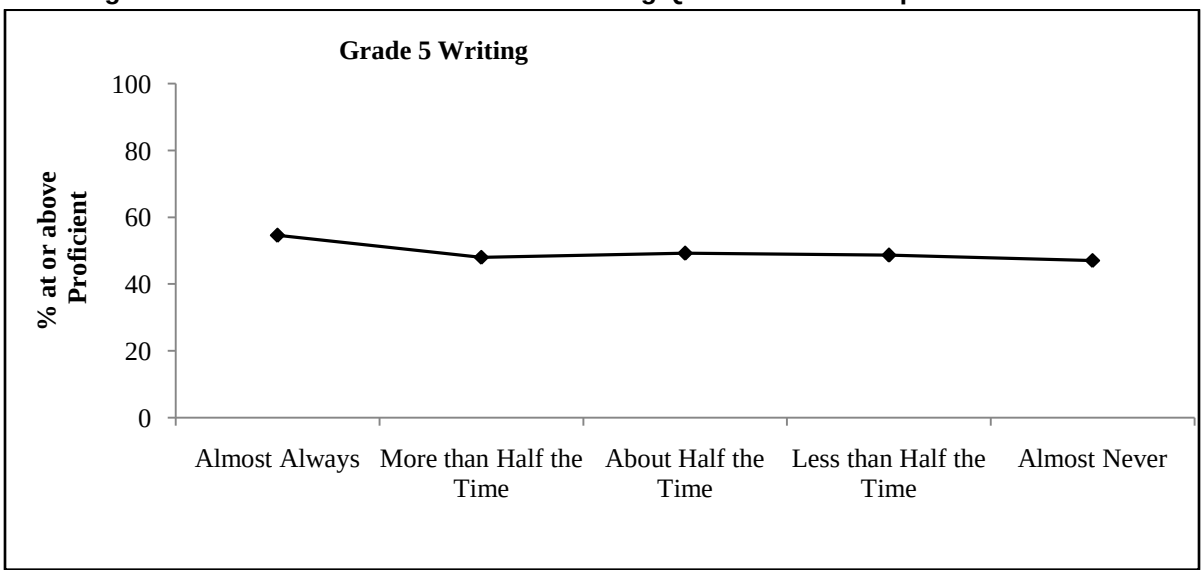


Figure 9-14. 2011–12 NECAP: Grade 8 Writing Questionnaire Responses—Content

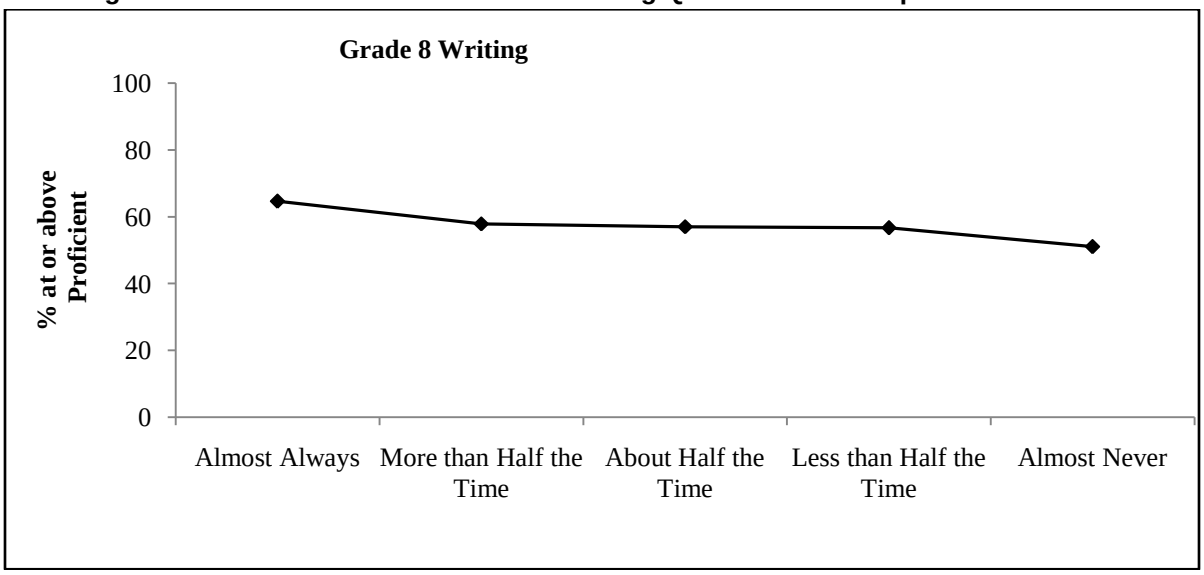
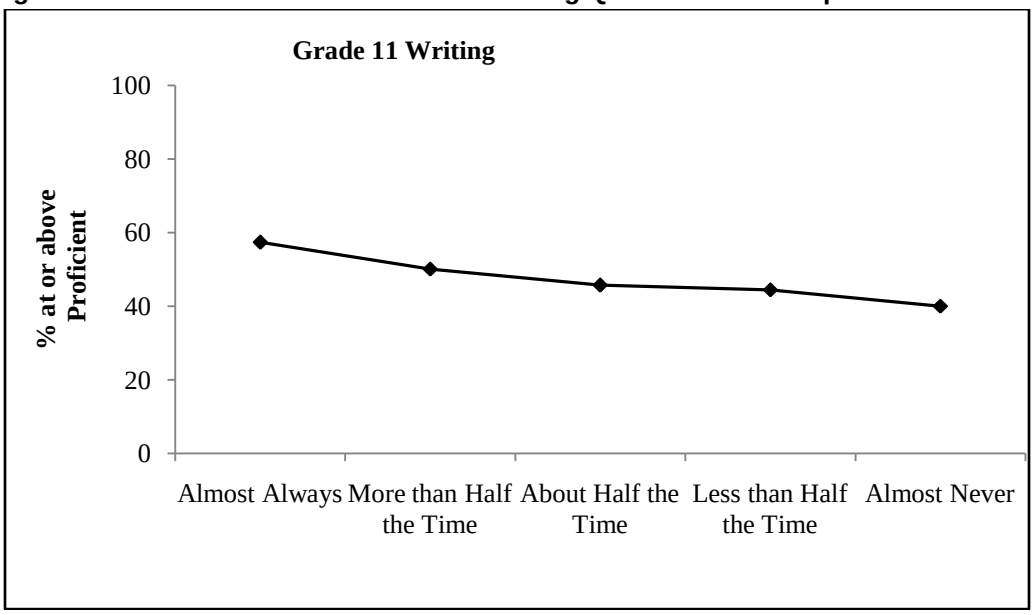


Figure 9-15. 2011–12 NECAP: Grade 11 Writing Questionnaire Responses—Content



9.1.3. Homework

Examinees in all grades in reading and mathematics were asked how often they have homework. In the sections below, results are presented for selected grade levels for each content area.

9.1.3.1. Homework: Reading

Figures 9-16 through 9-18 below show that students in grades 4, 7, and 11 who indicated they had reading homework more frequently performed better than students who had less homework. The relationship is more pronounced in the higher grades.

Figure 9-16. 2011–12 NECAP: Grade 4 Reading Questionnaire Responses—Homework

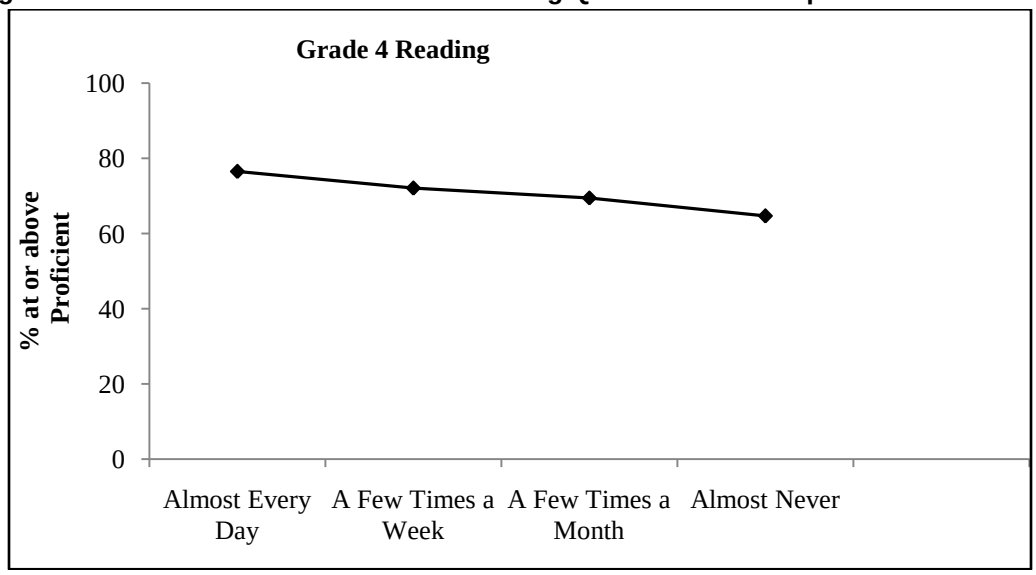


Figure 9-17. 2011–12 NECAP: Grade 7 Reading Questionnaire Responses—Homework

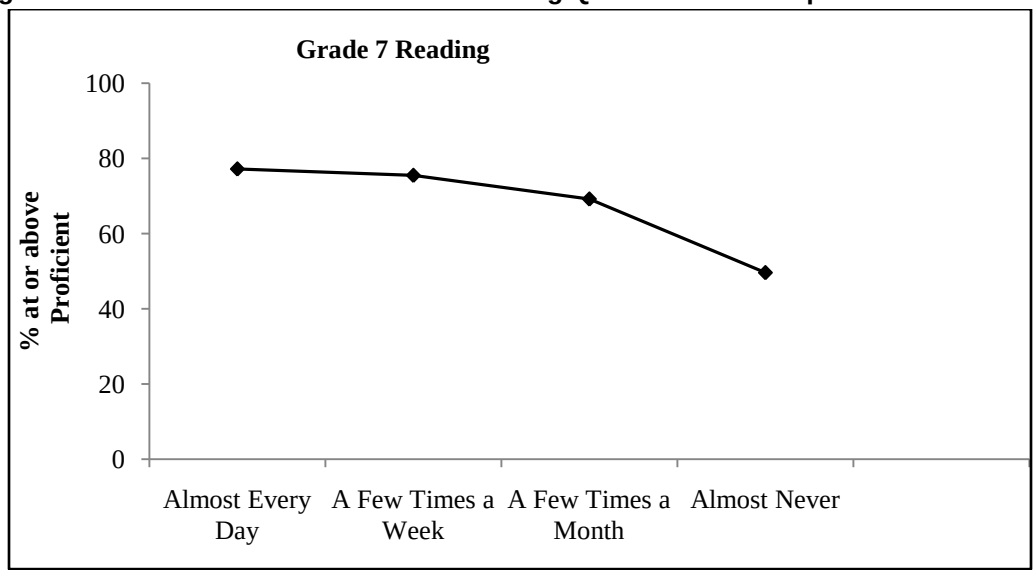
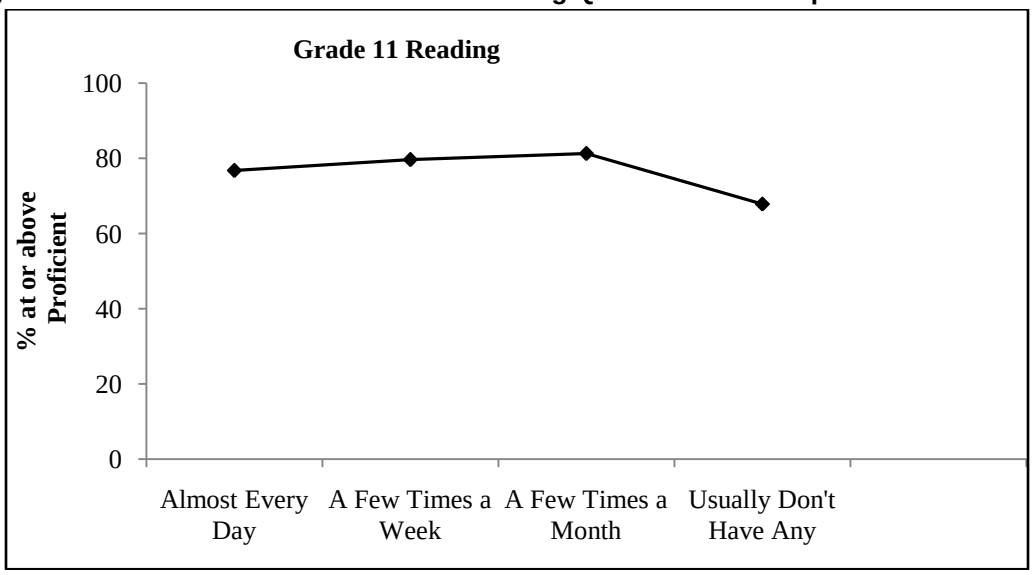


Figure 9-18. 2011–12 NECAP: Grade 11 Reading Questionnaire Responses—Homework



9.1.3.2. Homework: Mathematics

Figures 9-19 through 9-22 below show results that are very similar to those for reading: students in grades 4, 5, 8, and 11 who indicated that they had mathematics homework more frequently performed better than students who had less homework. Again, the pattern is more pronounced in the higher grades.

Figure 9-19. 2011–12 NECAP: Grade 4 Mathematics Questionnaire Responses—Homework

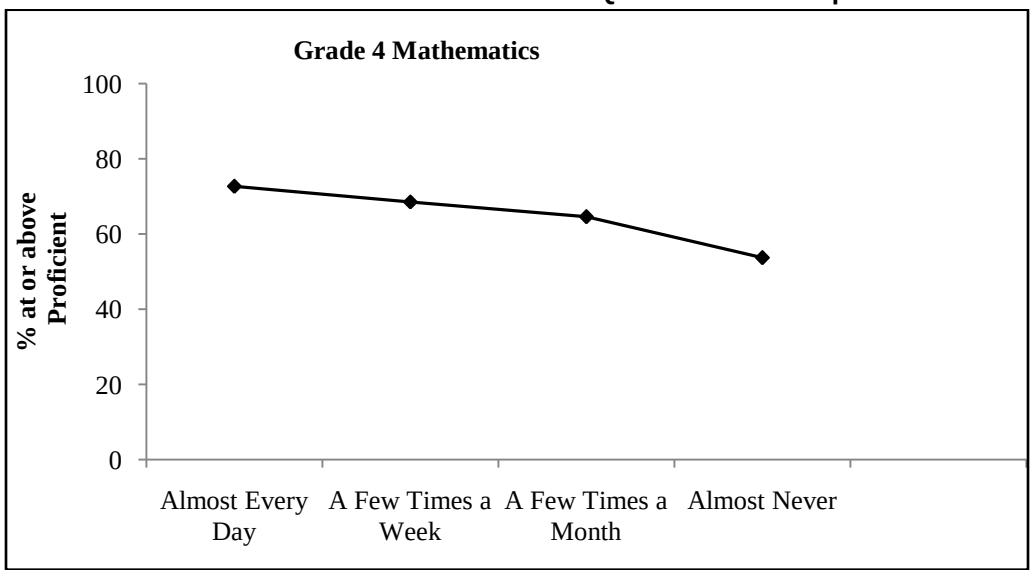


Figure 9-20. 2011–12 NECAP: Grade 5 Mathematics Questionnaire Responses—Homework

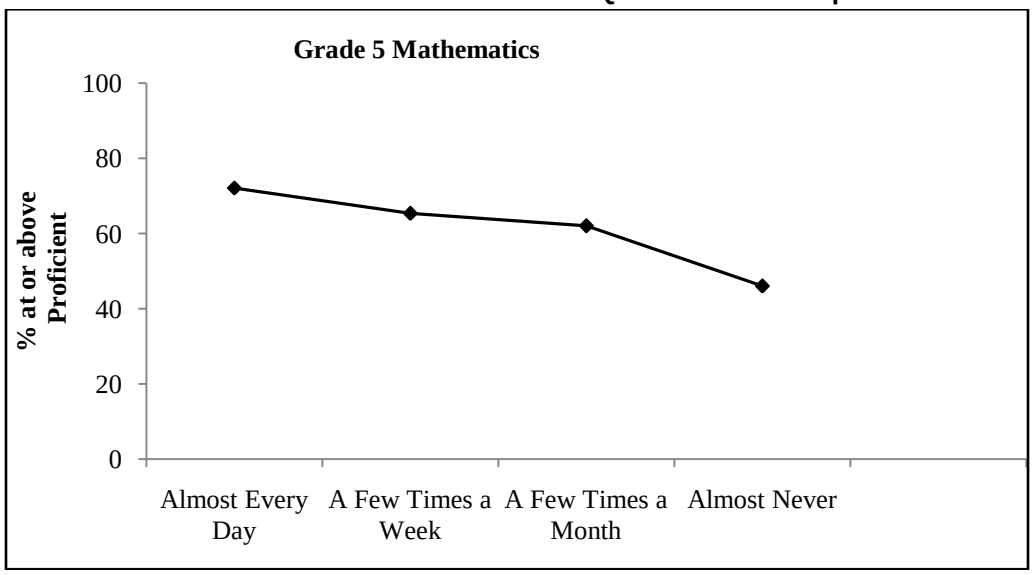


Figure 9-21. 2011–12 NECAP: Grade 8 Mathematics Questionnaire Responses—Homework

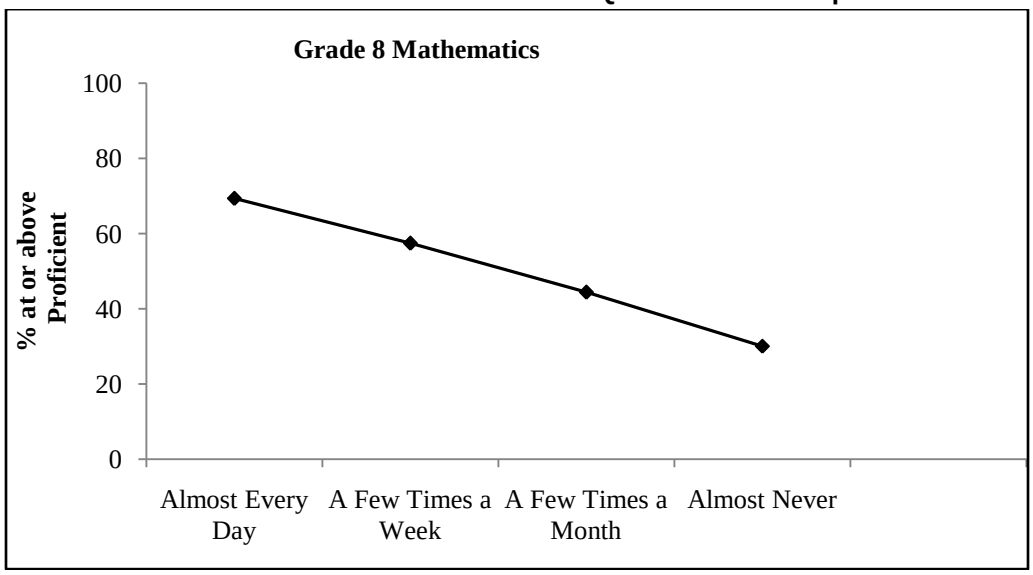
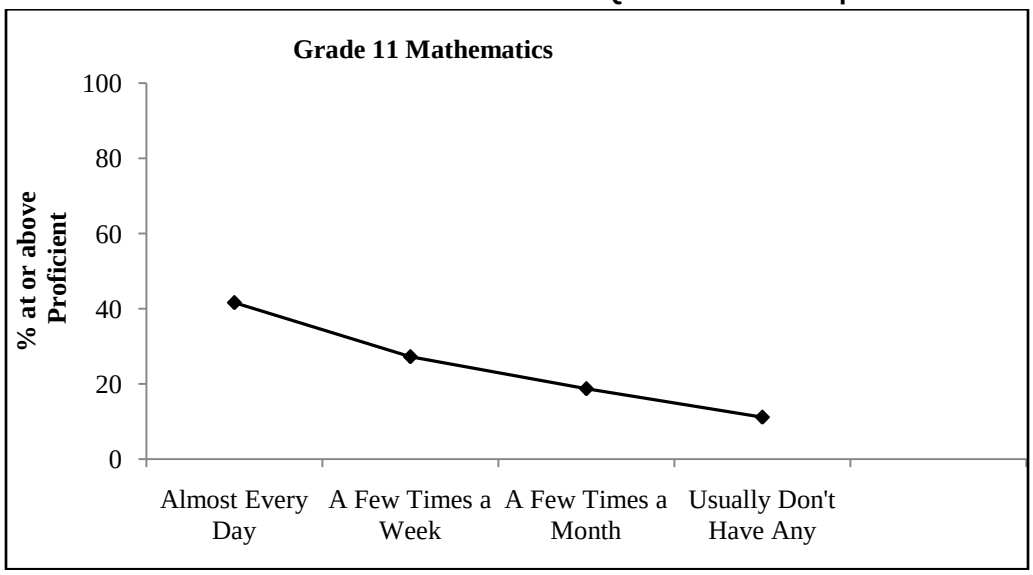


Figure 9-22. 2011–12 NECAP: Grade 11 Mathematics Questionnaire Responses—Homework



9.1.4. Performance in Courses

Students in grade 11 for both reading and mathematics were asked what their most recent course grade was. Figures 9-23 and 9-24 indicate that, for both reading and mathematics, there was a strong positive relationship between the most recent course grade and NECAP scores in that subject area.

Figure 9-23. 2011–12 NECAP: Grade 11 Questionnaire Responses—Grade in Reading

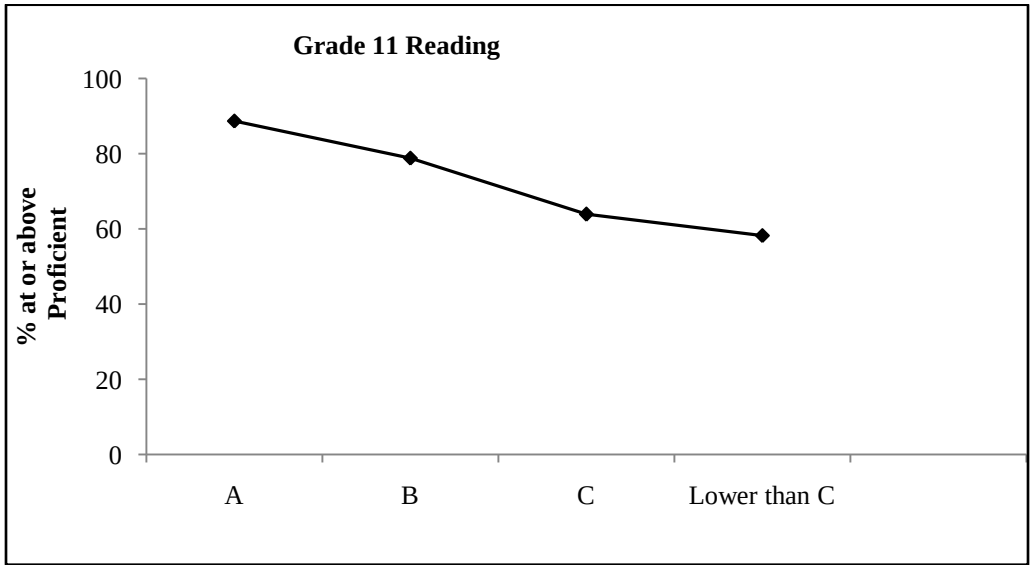
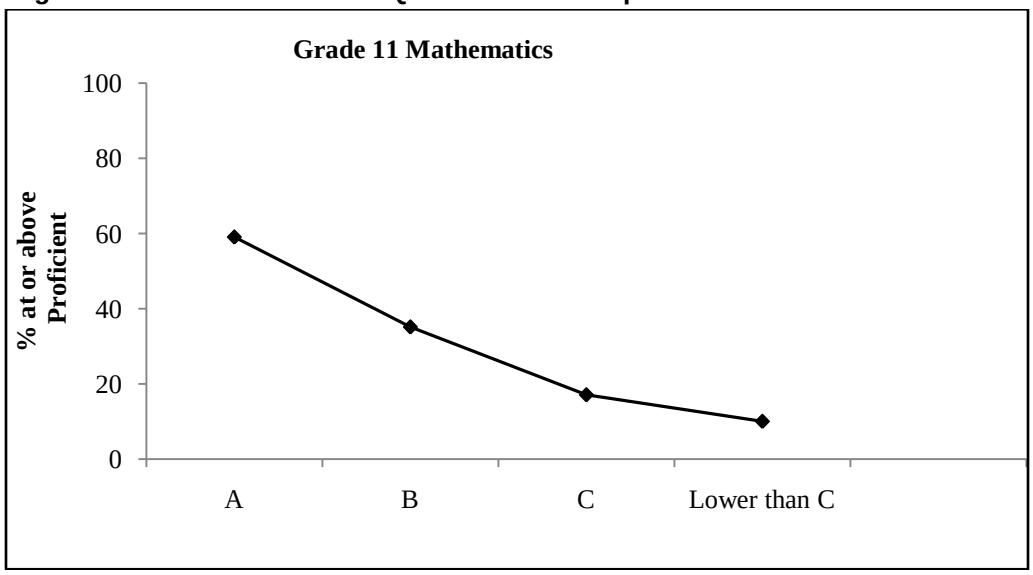


Figure 9-24. 2011–12 NECAP: Questionnaire Responses—Grade in Mathematics



The evidence presented in this report supports inferences made about student achievement on the content represented in the NECAP standards. As such, the evidence provided also supports the use of NECAP 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, Calif.: Wadsworth, Inc.

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

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

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

Chicago Manual of Style (15th ed., 2003). Chicago: University of Chicago Press.

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

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

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

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

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

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

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

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

Muraki, E., & Bock, R.D. (2003). PARSCALE 4.1. Lincolnwood, Ill.: 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).

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