



New England Common Assessment Program 2009–2010 Technical Report

August 2010



100 Education Way, Dover, NH 03820 (800) 431-8901

TABLE OF CONTENTS

CHAPTER 1.	OVERVIEW	1
1.1	<i>Purpose of the New England Common Assessment Program</i>	1
1.2	<i>Purpose of this Report</i>	1
1.3	<i>Organization of this Report</i>	2
CHAPTER 2.	DEVELOPMENT AND TEST DESIGN	3
2.1	<i>Grades 5 and 8 Writing Pilot Test Design</i>	3
2.1.1	Test Design of the Grades 5 and 8 Pilot	3
2.1.2	Administration of the Grades 5 and 8 Pilot Test	3
2.1.3	Scoring of the Grades 5 and 8 Pilot Test	4
2.2	<i>Operational Development Process</i>	4
2.2.1	Item Development	4
2.2.2	Grade Level and Grade Span Expectations	4
2.2.3	Internal Item Review	5
2.2.4	External Item Review	5
2.2.5	Bias and Sensitivity Review	6
2.2.6	Item Editing.....	7
2.2.7	Reviewing and Refining.....	7
2.2.8	Operational Test Assembly	7
2.2.9	Editing Drafts of Operational Tests.....	8
2.2.10	Braille and Large-Print Translation.....	8
2.3	<i>Item Types</i>	9
2.4	<i>Operational Test Designs and Blueprints</i>	9
2.4.1	Embedded Equating and Field Test Items	10
2.4.2	Test Booklet Design	10
2.5	<i>Reading Test Designs</i>	10
2.5.2	Reading Blueprint	11
2.6	<i>Mathematics Test Design</i>	13
2.6.2	The Use of Calculators on the NECAP	13
2.6.3	Mathematics Blueprint	13
2.7	<i>Writing Test Design: Grade 11</i>	15
2.7.1	Writing Blueprint: Grade 11.....	15
2.8	<i>Test Sessions</i>	15
CHAPTER 3.	TEST ADMINISTRATION	18
3.1	<i>Responsibility for Administration</i>	18
3.2	<i>Administration Procedures</i>	18
3.3	<i>Participation Requirements and Documentation</i>	18
3.4	<i>Administrator Training</i>	21
3.5	<i>Documentation of Accommodations</i>	21
3.6	<i>Test Security</i>	22
3.7	<i>Test and Administration Irregularities</i>	22
3.8	<i>Test Administration Window</i>	23
3.9	<i>NECAP Service Center</i>	23
CHAPTER 4.	SCORING	24
4.1	<i>Scoring of Standard Test Items</i>	24
4.1.1	Machine-Scored Items.....	24
4.1.2	Hand-Scored Items.....	24
4.1.2.1	Scoring Location and Staff	25
4.1.2.2	Benchmarking Meetings with the NECAP State Specialists	26
4.1.2.3	Reader Recruitment and Qualifications	27
4.1.2.4	Methodology for Scoring Polytomous Items	27
4.1.2.5	Reader Training	28
4.1.2.6	Senior Quality Assurance Coordinator and Senior Reader Training	30
4.1.2.7	Monitoring of Scoring Quality Control and Consistency	30
4.1.2.8	Reports Generated During Scoring.....	34
CHAPTER 5.	CLASSICAL ITEM ANALYSES.....	36
5.1	<i>Classical Difficulty and Discrimination Indices</i>	36
5.2	<i>Differential Item Functioning</i>	38
5.3	<i>Dimensionality Analyses</i>	39
CHAPTER 6.	IRT SCALING AND EQUATING.....	42
6.1	<i>Item Response Theory Scaling</i>	42
6.2	<i>Item Response Theory Analyses</i>	44
6.3	<i>Equating</i>	44

6.4	<i>Equating Results</i>	46
6.5	<i>Achievement Standards</i>	49
6.6	<i>Reported Scaled Scores</i>	49
6.6.1	Description of Scale	49
6.6.2	Calculations.....	50
6.6.3	Distributions.....	52
CHAPTER 7.	RELIABILITY	53
7.1	<i>Reliability and Standard Errors of Measurement</i>	54
7.2	<i>Subgroup Reliability</i>	55
7.3	<i>Item Type Reliability</i>	55
7.4	<i>Reporting Categories Reliability</i>	56
7.5	<i>Reliability of Achievement Level Categorization</i>	57
7.5.1	Accuracy and Consistency	57
7.5.2	Calculating Accuracy	57
7.5.3	Calculating Consistency.....	58
7.5.4	Calculating Kappa	58
7.5.5	Results of Accuracy, Consistency, and Kappa Analyses.....	58
CHAPTER 8.	SCORE REPORTING.....	60
8.1	<i>Teaching Year versus Testing Year Reporting</i>	60
8.2	<i>Primary Reporting Deliverables</i>	60
8.3	<i>Student Report</i>	60
8.4	<i>Item Analysis Reports</i>	61
8.5	<i>School and District Results Reports</i>	62
8.6	<i>School and District Summary Reports</i>	65
8.7	<i>School and District Student-Level Data Files</i>	65
8.8	<i>Analysis & Reporting System</i>	66
8.8.1	Interactive Reports	66
8.8.1.1	Item Analysis Report	66
8.8.1.2	Achievement Level Summary	66
8.8.1.3	Released Items Summary Data	67
8.8.1.4	Longitudinal Data Report	67
8.8.2	User Accounts	67
8.9	<i>Decision Rules</i>	68
8.10	<i>Quality Assurance</i>	68
CHAPTER 9.	VALIDITY.....	70
9.1	<i>Questionnaire Data</i>	71
9.2	<i>Validity Studies Agenda</i>	80
9.2.1	External Validity	80
9.2.2	Convergent and Discriminant Validity.....	80
9.2.3	Structural Validity	81
9.2.4	Procedural Validity	82
REFERENCES		84
APPENDICES		86
Appendix A	<i>Committee Membership</i>	
Appendix B	<i>Accommodation Frequencies by Content Area</i>	
Appendix C	<i>Table of Standard Test Accommodations</i>	
Appendix D	<i>Double Blind Interrater Agreement</i>	
Appendix E	<i>Item Level Classical Statistics Results</i>	
Appendix F	<i>Classical Item Statistic Descriptives Summarized by Grade, Content, and Form</i>	
Appendix G	<i>Number of Items Classified into DIF Categories by Subgroup, Test Form, and Item Type</i>	
Appendix H	<i>Common Item DIF Category Counts in the Male vs. Female Comparison</i>	
Appendix I	<i>Item Response Theory Parameters, TCCs and TIFs</i>	
Appendix J	<i>Delta Analyses and Rescore Analysis Results</i>	
Appendix K	<i>Raw to Scaled Score Look-up Tables</i>	
Appendix L	<i>Scaled Score Percentages and Cumulative Percentages</i>	
Appendix M	<i>Detailed Alpha Coefficient Results</i>	
Appendix N	<i>Decision Accuracy and Consistency Results</i>	
Appendix O	<i>Sample Reports</i>	
Appendix P	<i>Analysis and Reporting Decision Rules</i>	
Appendix Q	<i>Student Questionnaire Data</i>	

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 (ME), New Hampshire (NH), Rhode Island (RI), and Vermont (VT) 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). Maine students in grades 3 through 8 participated in NECAP for the first time in 2009. The purposes of the tests are as follows: (1) provide data on student achievement in reading/language arts and mathematics to meet the requirements of NCLB; (2) provide information to support program evaluation and improvement; and (3) provide 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
- 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 2009–10 NECAP. In October of 2009, students in grades 3 through 8 and 11 participated in the administration of the NECAP 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 test 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. DEVELOPMENT AND TEST DESIGN

2.1 Grades 5 and 8 Writing Pilot Test Design

In October of 2009 a pilot test for Grades 5 and 8 writing was conducted to field test all newly developed writing items to be used in the following five years' operational tests.

The test design mirrored the design of the operational writing test in 2005–2008. The assessment framework for writing is based on the *NECAP Grade Level Expectations*, and all items on the NECAP test are designed to measure specific GLEs. The content standards in 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 (constructed-responses only at Grade 8)
- Informational writing (report/procedure for Grades 5 and 8 and persuasive at Grade 8)

The writing prompt and the three constructed-response items each address a different genre. In addition, structures and conventions of language are assessed through multiple-choice items and throughout the student's writing. The pilot test development process mirrored the operational development process described in Section 2.2.

2.1.1 Test Design of the Grades 5 and 8 Pilot

The pilot test forms were constructed to mirror the intended operational test design. The only difference was that all item positions on the pilot test forms were populated with field test items. Eight forms were field tested.

Each of the eight forms of the NECAP writing pilot in grades 5 and 8 for 2010 included ten multiple-choice items, three constructed-response items and one writing prompt. Each multiple-choice item was worth one point, each constructed-response item four points, and the writing prompt 12 points.

2.1.2 Administration of the Grades 5 and 8 Pilot Test

All schools and all students in grades 5 and 8 participated in the pilot test. The test administration procedures for the pilot test mirrored the procedures for the operational test to ensure an even distribution of forms among all schools and all students.

2.1.3 Scoring of the Grades 5 and 8 Pilot Test

All student responses to multiple-choice questions were scanned and analyzed to produce item statistics. All short-answer, constructed-response, and writing prompt items were benchmarked and scored on a sample of roughly 1,200 students.

Because the pilot test was conducted to emulate the subsequent operational test as much as possible, readers are referred to other chapters of this report for more specific details.

2.2 Operational Development Process

2.2.1 Item Development

Curriculum and assessment content specialists at Measured Progress begin the item development process by selecting passages and graphics and developing items and scoring guides according to guidance and specifications provided by the NECAP states. During this process, the curriculum and assessment specialists do the following:

- Work in close collaboration with the states from conceptualization to production of the final camera ready copies
- Review the Grade-Level Expectations (GLEs), Grade-Span Expectations (GSEs), and test specifications to ensure that the items developed truly measure student learning and meet each state’s goals for accountability
- Write and edit items that adhere to the NECAP test specifications
- Lead Item Review Committees
- Participate in benchmarking to ensure NECAP scoring reflects the true intent of the items
- Select items to create tests that are educationally significant, as well as valid and reliable for the purposes of reporting data

2.2.2 Grade Level and Grade Span Expectations

NECAP test items are directly linked to the *content standards* and *performance indicators* described in the GLEs/GSEs. The content standards for each grade are grouped into content clusters for the purpose of reporting results; the performance indicators are used by the content specialists to help guide the development of test questions. An item may address one, several, or all of the performance indicators.

2.2.3 Internal Item Review

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

- Review of the formatted item, open-response scoring guide, and any reading selections and graphics
- Evaluation of item “integrity,” content, and structure; appropriateness to designated content area; item 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 that 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, Depth of Knowledge (DOK) and identified contract)?
- Is the item appropriate for the designated grade level?

2.2.4 External Item Review

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 committee’s role is to review test items for the NECAP, provide feedback, and make recommendations about which items should be selected for program use. The 2009–10 NECAP IRCs for each content area in grade levels 3 through 8 and 11 met in the spring of 2009. Committee members reviewed the entire set of embedded field test items proposed for the 2009–10 operational test and made recommendations about selecting, revising, or eliminating specific items from the item pool. Members reviewed each item against the following criteria:

- Grade-Level/Grade-Span Expectation Alignment
 - Is the test item aligned to the appropriate GLE/GSE?

- If not, which GLE/GSE or grade level is more appropriate?
- 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 Knowledge¹
 - Are the items coded to the appropriate Depth of Knowledge?
 - If consensus cannot be reached, is there clarity around why the item might be on the borderline of two levels?
- Language
 - Is the item language clear?
 - Is the item language accurate (syntax, grammar, conventions)?
- 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?

2.2.5 Bias and Sensitivity Review

Bias review is an essential part of the development process. During the bias review process, NECAP passages and items were reviewed by a committee of teachers, English Language Learner (ELL) specialists, special education teachers, and other educators and members of major constituency groups who represent the interests of legally protected and/or educationally disadvantaged groups. (A list of bias and sensitivity review committee 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.

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.

2.2.6 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.

2.2.7 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.

2.2.8 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 2009–10 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 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.

2.2.9 Editing Drafts of Operational Tests

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).

2.2.10 Braille and Large-Print Translation

Common items for grades 3 through 8 and 11 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 adapted into a large-print version.

2.3 *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, to provide breadth of coverage of the GLEs/GSEs. Because they require approximately one minute for most students to answer, these items make efficient use of limited testing time and allow coverage of a wide range of knowledge and skills, including, for example, word identification and vocabulary skills.

Short-answer items were administered in grades 3 through 8 and 11 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. Short-answer items require 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. Constructed-response items require approximately five to ten minutes for most students to complete. These items were administered in grades 3 through 8 and 11 in reading, and in grades 5 through 8 and 11 in mathematics.

Writing prompts A single common writing prompt and one additional matrix writing prompt per form were administered in grade 11. 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 2009–10. The released NECAP items are posted on a Web site hosted by Measured Progress and on the Department of Education Web sites. 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 assessment.

2.4 *Operational Test Designs and Blueprints*

Since the beginning of the program, the goal of NECAP has been to measure what students know and are able to do by using a variety of test item types. The program was structured to use both common and matrix-sampled items. (Common items are those taken by all students at a given grade level. Matrix-sampled items comprise a pool that is divided among the multiple forms of the test at each grade level. Their purpose is described in section 2.4.1.) This design provides reliable and valid results at the student level, and breadth of coverage of a content area at the school results level while minimizing testing time. (Note: Only common items count toward students' scaled scores.)

2.4.1 Embedded Equating and Field Test Items

To ensure that NECAP scores obtained from different test forms and different years are equivalent to each other, a set of equating items is matrixed across forms of the reading and mathematics tests. Chapter 5 presents more detail on the equating process. (Note: Equating items are not counted toward students' scaled scores.)

NECAP also includes embedded field test items in all content areas except grades 5 and 8 writing. Because the field test items are taken by many students, the sample is sufficient to produce reliable data from which to inform the process of selecting items for future tests. Embedding field test items achieves two other objectives. First, it creates a pool of replacement items in reading and mathematics that are needed because of the release of common items each year. Second, embedding field test items into the operational test ensures that students take the items under operational conditions. (Note: As with the matrixed equating items, field test items are not counted toward students' scaled scores.)

2.4.2 Test Booklet Design

To accommodate the embedded equating and field test items in the 2009–10 NECAP, there were nine unique test forms in grades 3 through 8 and eight unique forms in grade 11. In all reading and mathematics test sessions, the equating and field test items were distributed among the common items in a way that was not evident to test takers. The grade 11 writing design called for each student to respond to two writing prompts. The first writing prompt was common for all students and the second writing prompt was either a matrix prompt or a field test prompt, depending on the particular test form.

2.5 Reading Test Designs

Table 2-1 summarizes the number and types of items that were used in the 2009–10 NECAP reading test 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 2-1. 2009–10 NECAP: Item Type and Number of Items—Reading Grades 3–8

	<i>Long passages</i>	<i>Short passages</i>	<i>Stand-alone MC</i>	<i>Total MC</i>	<i>Total CR</i>
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	0	3	2	14	3
Total per Student					
Forms 1–7	3	3	6	42	9
Forms 8–9	2	5	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 2-2 summarizes the numbers and types of items that were used in the 2009–10 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 2-2. 2009–10 NECAP: Item Type and Number of Items—Reading Grade 11

	<i>Long passages</i>	<i>Short passages</i>	<i>Stand-alone MC</i>	<i>Total MC</i>	<i>Total CR</i>
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

2.5.2 Reading Blueprint

As indicated earlier, the test framework for reading in grades 3 through 8 was based on the NECAP GLEs, and all items on the NECAP test were designed to measure a specific GLE. The test framework for reading in grade 11 was based on the NECAP GSEs, and all items on the NECAP test were designed to measure a specific GSE. 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; drama; 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.

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. The distribution of emphasis for reading is shown in Table 2-3.

Table 2-3. 2009–10 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 2-4 shows the content category 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 2-4. 2009–10 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	20	18	10	9	10	11	10
Type of Text							
Literary	15	18	21	23	21	20	21
Informational	17	16	21	20	21	21	21
Level of Comprehension							
Initial Understanding	21	19	18	21	19	19	15
Analysis and Interpretation	11	15	24	22	23	22	27
Total	52	52	52	52	52	52	52

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).

Table 2-5 lists the percentage of total score points assigned to each DOK level in reading.

Table 2-5. 2009–10 NECAP: Depth of Knowledge in Terms of Targeted Percentage of Test by Grade—Reading Grades 3–8 and 11

DOK	Grade						
	3	4	5	6	7	8	11
Level 1	60%	52%	23%	21%	15%	8%	21%
Level 2	40%	48%	69%	77%	69%	69%	40%
Level 3	0%	0%	8%	2%	15%	23%	38%
Total	100%						

2.6 Mathematics Test Design

Table 2-6 summarizes the numbers and types of items that were used in the 2009–10 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 either one or two points, and each constructed-response item four points. Score points within a grade level were evenly divided, so that multiple-choice items represented approximately 50% of possible score points, and short-answer and constructed-response items together represented approximately 50% of score points.

Table 2-6. 2009–10 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; six contained unique matrix-equating items while Forms 7-9 contained the same matrix-equating items as Forms 1-3.

For grade 11, total of eight forms; six contained unique matrix-equating items while Forms 7-8 contained the same matrix-equating items as Forms 1-2.

2.6.2 The Use of Calculators on the NECAP

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 test 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. (Test sessions are discussed in greater detail at the end of this chapter.)

2.6.3 Mathematics Blueprint

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

- 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. The distribution of emphasis for mathematics reporting subcategories is shown in Table 2-7.

**Table 2-7. 2009–10 NECAP: Distribution of Emphasis
In Terms of Target Percentage of Test by Grade—Mathematics Grades 3–8 and 11**

<i>Subcategory</i>	<i>Grade</i>						
	2 (3)	3 (4)	4 (5)	5 (6)	6 (7)	7 (8)	9–10 (1)1
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 2-8 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 2-8. 2009–10 NECAP: Reporting Subcategories and
Possible Raw Score Points by Grade—Mathematics Grades 3–8 and 11**

<i>Subcategory</i>	<i>Grade tested</i>						
	3	4	5	6	7	8	11
Numbers and Operations	35	32	30	26	20	13	9
Geometry and Measurement	10	13	13	17	16	17	19
Functions and Algebra	10	10	13	13	20	26	26
Data, Statistics, and Probability	10	10	10	10	10	10	10
Total	65	65	66	66	66	66	64

Table 2-9 on the next page lists the percentage of total score points assigned to each level of DOK in mathematics.

Table 2-9. 2009–10 NECAP: Depth of Knowledge in Terms of Targeted Percentage of Test by Grade—Mathematics Grades 3–8 and 11

<i>DOK</i>	<i>Grade</i>						
	3	4	5	6	7	8	11
Level 1	23%	22%	35%	26%	27%	29%	27%
Level 2	68%	71%	65%	64%	67%	62%	70%
Level 3	9%	8%	0%	11%	6%	9%	3%
Total	100%	100%	100%	100%	100%	100%	100%

2.7 Writing Test Design: Grade 11

For the 2009–10 NECAP writing test for grade 11, there were a total of 8 forms: five equating forms and three field test forms. Therefore, each student responded to two different writing prompts, one common and either one matrix-equating or one field test prompt. The common prompt was worth 12 points.

2.7.1 Writing Blueprint: Grade 11

The test framework for grade 11 writing was based on the NECAP GSEs, and all items on the NECAP test were designed to measure a specific GSE. The content standards for grade 11 writing identify six genres that are grouped into three major strands:

- Writing in response to text (literary and informational)
- Informational writing (report, procedure, and persuasive essay)
- Expressive writing (reflective essay)

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.

2.8 Test Sessions

The NECAP tests were administered October 1–22, 2009 to grades 3 through 8 and 11. During the testing window, schools were able to schedule testing sessions at any time, but were instructed to follow the sequence in the scheduling guidelines as detailed in the test administration manual. It was also mandatory that

all testing classrooms at a grade level within a school be on the same schedule. A third week during the testing window was reserved for makeup testing of students who were absent during initial test sessions.

The timing and scheduling guidelines for the NECAP tests were based on estimates of the time it would take an average student to respond to each type of item on the test:

- multiple-choice—1 minute
- short answer (1 point)—1 minute
- short answer (2 point)—2 minutes
- constructed response—10 minutes
- long writing prompt—45 minutes

For the reading sessions, the scheduling guidelines estimate that reading the stimulus material (passage) will take approximately 10 minutes. Table 2-10 shows the distribution of items across the test sessions for each content area and grade level.

Table 2-10. 2009–10 NECAP: Number of Items by Item Type by Session

<i>Content area</i>	<i>Grade</i>	<i>Item type</i>	<i>Session 1</i>	<i>Session 2</i>	<i>Session 3</i>
Reading	3–8	MC	14	14	14
		CR	3	3	3
	11	MC	22	20	—
		CR	4	5	—
Mathematics	3–4	MC	15	15	15
		SA1	4	3	6
		SA2	4	5	4
	5–8	MC	14	14	13
		SA1	3	3	3
		SA2	3	3	3
		CR	2	2	2
	11	MC	16	16	—
		SA1	6	6	—
		SA2	6	6	—
CR		3	3	—	
Writing	11	MC	0	0	—
		CR	0	0	—
		SA1	0	0	—
		WP	1	1	—

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

Although the scheduling guidelines are based on the assumption that most students will complete the test within the estimated time, each test session allowed additional time for students who may have needed it. Up to 100% additional time was allocated for each session (i.e., a 45-minute session could be extended by an additional 45 minutes).

If classroom space was not available for students who required additional time to complete the tests, schools were allowed to consider using another space for this purpose, such as a guidance office. If additional

areas were not available, it was recommended that each classroom used for test administration be scheduled for the maximum amount of time. Detailed instructions regarding test administration and scheduling were provided in the test coordinators' and administrators' manuals.

Chapter 3. TEST ADMINISTRATION

3.1 *Responsibility for Administration*

The 2009 *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, 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 test. 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 Department 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 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 2008–09 school year and therefore were not required to participate in the NECAP in 2009–10. 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 booklets 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, 2008 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 2009; however, these students were required to take the mathematics test in 2009.
- Student withdrew from school after October 1, 2009
 - B. If a student withdrew after October 1, 2009 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, 2009
 - C. If a student enrolled after October 1, 2009 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, 2009 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.”

Tables 3-1, 3-2, and 3-3 list the participation rates of the three states combined in reading, mathematics, and writing.

Table 3-1. 2009–10 NECAP: Participation Rates—Mathematics

<i>Category</i>	<i>Description</i>	<i>Enrollment</i>	<i>Not tested State-approved</i>	<i>Not tested/ other</i>	<i>Number tested</i>	<i>% tested</i>
All	All Students	311,629	3,743	2,404	305,482	98
Gender	Male	161,046	2,344	1,442	157,260	98
	Female	150,541	1,399	960	148,182	98
	Not Reported	42	0	2	40	95
Ethnicity	Am. Indian or Alaskan Nat.	1,798	39	35	1,724	96
	Asian	7,271	70	65	7,136	98
	Black or African American	12,469	211	146	12,112	97
	Hispanic or Latino	19,279	239	236	18,804	98
	Nat. Hawaiian or Pacific Is.	101	4	1	96	95
	White (Non-Hispanic)	268,770	3,149	1,884	263,737	98
	Not Reported	1,941	31	37	1,873	96
LEP	Current	8,514	72	72	8,370	98
	Monitoring Year 1	1,287	7	2	1,278	99
	Monitoring Year 2	1,103	7	5	1,091	99
	Other	300,725	3,657	2,325	294,743	98
IEP	IEP	50,804	3,089	1,077	46,638	92
	Other	260,825	654	1,327	258,844	99

continued

<i>Category</i>	<i>Description</i>	<i>Enrollment</i>	<i>Not tested State-approved</i>	<i>Not tested/ other</i>	<i>Number tested</i>	<i>% tested</i>
SES	SES	107,615	1,733	1,025	104,857	97
	Other	204,014	2,010	1,379	200,625	98
Migrant	Migrant	83	1	0	82	99
	Other	311,546	3,742	2,404	305,400	98
Title 1	Title 1	43,293	441	289	42,563	98
	Other	268,336	3,302	2,115	262,919	98
Plan 504	Plan 504	2,667	15	13	2,639	99
	Other	308,962	3,728	2,391	302,843	98

Table 3-2. 2009–10 NECAP: Participation Rates—Reading

<i>Category</i>	<i>Description</i>	<i>Enrollment</i>	<i>Not tested state-approved</i>	<i>Not tested/ other</i>	<i>Number Tested</i>	<i>% Tested</i>
All	All Students	311,629	4,369	2,323	304,937	98
Gender	Male	161,046	2,686	1,375	156,985	97
	Female	150,541	1,683	943	147,915	98
	Not Reported	42	0	5	37	88
Ethnicity	Am. Indian or Alaskan Nat.	1,798	42	32	1,724	96
	Asian	7,271	219	107	6,945	96
	Black or African American	12,469	353	166	11,950	96
	Hispanic or Latino	19,279	557	256	18,466	96
	Nat. Hawaiian or Pacific Is.	101	8	1	92	91
	White (Non-Hispanic)	268,770	3,152	1,729	263,889	98
	Not Reported	1,941	38	32	1,871	96
LEP	Current	8,514	724	174	7,616	89
	Monitoring Year 1	1,287	8	2	1,277	99
	Monitoring Year 2	1,103	7	5	1,091	99
	Other	300,725	3,630	2,142	294,953	98
IEP	IEP	50,804	3,140	966	46,698	92
	Other	260,825	1,229	1,357	258,239	99
SES	SES	107,615	2,140	979	104,496	97
	Other	204,014	2,229	1,344	200,441	98
Migrant	Migrant	83	1	0	82	99
	Other	311,546	4,368	2,323	304,855	98
Title 1	Title 1	45,318	773	287	44,258	98
	Other	266,311	3,596	2,036	260,679	98
Plan 504	Plan 504	2,667	16	9	2,642	99
	Other	308,962	4,353	2,314	302,295	98

Table 3-3. 2009–10 NECAP: Participation Rates—Writing

<i>Category</i>	<i>Description</i>	<i>Enrollment</i>	<i>Not tested state-approved</i>	<i>Not tested/ other</i>	<i>Number Tested</i>	<i>% Tested</i>
All	All Students	34,024	430	864	32,730	96
Gender	Male	17,294	250	528	16,516	96
	Female	16,722	180	334	16,208	97
	Not Reported	8	0	2	6	75
Ethnicity	Am. Indian or Alaskan Nat.	152	4	6	142	93
	Asian	733	17	25	691	94
	Black or African American	1,449	37	67	1,345	93
	Hispanic or Latino	2,489	63	109	2,317	93

continued

<i>Category</i>	<i>Description</i>	<i>Enrollment</i>	<i>Not tested state-approved</i>	<i>Not tested/ other</i>	<i>Number Tested</i>	<i>% Tested</i>
Ethnicity	Nat. Hawaiian or Pacific Is.	17	1	1	15	88
	White (Non-Hispanic)	28,959	304	641	28,014	97
	Not Reported	225	4	15	206	92
LEP	Current	519	65	23	431	83
	Monitoring Year 1	121	0	1	120	99
	Monitoring Year 2	77	0	4	73	95
	Other	33,307	365	836	32,106	96
IEP	IEP	5,392	274	337	4,781	89
	Other	28,632	156	527	27,949	98
SES	SES	8,308	153	315	7,840	94
	Other	25,716	277	549	24,890	97
Migrant	Migrant	1	0	0	1	100
	Other	34,023	430	864	32,729	96
Title 1	Title 1	2,783	64	102	2,617	94
	Other	31,241	366	762	30,113	96
Plan 504	Plan 504	234	2	5	227	97
	Other	33,790	428	859	32,503	96

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 and to provide training on the policies and procedures regarding administration of the tests. These workshops were geared toward new or inexperienced NECAP test coordinators. Two audio PowerPoint CDs were also produced and sent to every school. One CD was for training experienced test coordinators and highlighted new procedures for 2009 and emphasized important policies. The second CD was for test coordinators to use when training test administrators.

3.5 Documentation of Accommodations

The *Principal/Test Coordinator Manual* and *Test Administrator Manuals* 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 list of allowable accommodations was revised in August of 2009. The NECAP states worked together to change the coding system, revise existing accommodations, and add or delete certain accommodations. The new 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 B shows the accommodation frequencies by content area for the October 2009 NECAP test administration. The accommodation codes are defined in the Table of Standard Test Accommodations, which can be found in Appendix C.

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 the *Test Administrator Manual* explain in detail all test security measures and test administration procedures. School personnel were informed that any concerns about breaches in test security were to be reported to the school’s test coordinator and/or principal immediately. The test coordinator and/or principal 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 at test administration workshops that were conducted in all three states. The three 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 that 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

Prior to test administration, but after shipments were sent to schools, a packing issue was discovered with a number of grade 3 materials. A number of schools receiving a grade 3 test shipment were sent grade 4 test administrator manuals. It was determined that one box of grade 4 manuals was inadvertently set on the grade 3 packing line and approximately 400 manuals for the incorrect grade were sent. All affected schools

called the Service Center and were immediately sent the appropriate number of grade 3 manuals prior to testing.

3.8 Test Administration Window

The test administration window was October 1–22, 2009.

3.9 NECAP Service Center

To provide additional support to schools before, during, and after testing, Measured Progress established the NECAP Service Center. The support of the 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 which includes notes regarding the issue and resolution of each call.

The Service Center was staffed year-round by a varying number of representatives depending upon need and call volume and was open to receive calls from 8:00 AM to 4:00 PM Monday through Friday. Extra representatives were available beginning two weeks before the start of testing and ending two weeks after testing.

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 backwards and identification of critical data that are missing (e.g., a student ID number), test forms that are out of range or missing, and page or document sequence errors. When a problem is detected, the scanner stops and displays an error message directing the operator to investigate and to correct the situation.

4.1.2 **Hand-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 onsite. (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 2009–10 NECAP test item responses by grade and content area were scored.

Table 4-1. 2009–10 NECAP: Operational Scoring Locations by Content and Grade

<i>Content area</i>	<i>Grade</i>	<i>Louisville, KY</i>	<i>Dover, NH</i>	<i>Troy, NY</i>	<i>Longmont, CO</i>
Mathematics	3	X			
	4	X			
	5				X
	6				X
	7				X
	8				X
Reading	11				X
	3			X	
	4			X	
	5			X	
	6	X			
	7	X			
Writing	8	X			
	11	X			
	5				
	8				
	11				X

* NECAP Writing Grades 5 and 8 consisted entirely of field test items which were scored in Dover, NH—no operational writing items were administered or scored for these two grades.

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 2009–10 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. Chief Readers are employees of Measured Progress.
- Numerous Quality Assurance Coordinators (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 (SRs) 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 SRs to Readers was approximately 1:11.
- Readers at scoring sites scored operational and field test NECAP 2009–10 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 2009–10 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 about Readers (e.g., gender, race, educational background) 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., resume and/or transcripts) of their qualifications.

Table 4-2 summarizes the qualifications of the 2009–10 NECAP scoring leadership and Readers.

Table 4-2. 2009–10 NECAP: Qualifications of Scoring Leadership and Readers—Fall Administration

<i>Scoring responsibility</i>	<i>Educational credentials</i>				<i>Total</i>
	<i>Doctorate</i>	<i>Master's</i>	<i>Bachelor's</i>	<i>Other</i>	
Scoring Leadership	2.9%	34.5%	57.6%	5.0%	100.0%
Readers	4.4%	27.2%	55.5%	12.9%	100.0%

Scoring Leadership = Chief Readers, QACs, and SRs

*3 QAC/SRs had an Associate's degree and 4 at least 48+ college credits

**77 Readers had an Associate's degree and 64 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. 2009–10 NECAP: Possible Score Points for Polytomous Item Types

<i>Polytomous item type</i>	<i>Possible score point range</i>
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 document (electronic copy or hard copy, as needed) to identify the correct location (in the answer document) or to more closely examine the response and then assign a score.

Scoring Procedures

Scoring procedures for polytomous items included both single scoring and double 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 4.1.2.7 and Appendix D).

4.1.2.5 Reader Training

Reader training began with an introduction of the onsite scoring staff and providing an overview of the NECAP program’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 reviewing or scoring 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 first Qualifying Sets after the qualification round), even if all or part of the set was also presented online via computer.

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 test development staff. Each contains one client-approved sample response per score point considered to be a midrange exemplar. The set includes a second sample response if there is 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 group discussion of one or two responses, directing Reader attention to difficult scoring issues (e.g., the borderline between two score points). Trainers modeled for Readers throughout 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 assembled for constructed-response items, writing prompts, and all 2-point short-answer items for grades 3 and 4 mathematics. The ten 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 pre-determined score on at least 8 of the 10 responses) and at least 90% exact or adjacent agreement (i.e., to exactly match or be within one score point of the pre-determined score on 9 or 10 of the 10 responses), except 70% and 90%, respectively, for 6-point writing-prompt responses. In other words, Readers were allowed 1 discrepant score (i.e., 1 score of 10 that was more than one score point from the pre-determined score) provided they had at least 8 exact scores (7 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 SRs were trained in a separate training session immediately prior to Reader training. In addition to discussing the items and their responses, QAC and SR 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 SRs to handle questions from the latter.

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 more strict 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, 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 Committee-Reviewed Responses (CRRs)

Committee-Reviewed Responses (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 5 and 30 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 WP items, because these are 100% double-blind scored (defined below). Embedded CRRs were also not used for 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 SR) scoring a response after a Reader has already scored the response. The practice was applied to all open-ended 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 SR. The iScore system allowed 1, 2, or 3 responses per Reader to be placed into the read-behind queue at a time.

The SR entered his or her score into iScore before being allowed to see the Reader’s score. The SR 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 1 point), the SR’s score became the score of record. (A significant number of adjacent scores for a Reader triggered an individual scoring consultation with the SR, after which the QAC determined whether or when the Reader could resume scoring.)
- If the scores were discrepant (i.e., differed by more than 1 point), the SR’s score became the score of record. (This triggered an individual consultation with the SR, 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. 2009–10 NECAP: Examples of Read-Behind Scoring Resolutions

<i>Reader score</i>	<i>QAC/SR score</i>	<i>Score of record</i>
4	4	4
4	3	3*
4	2	2*

* QAC/SR’s score.

SRs were tasked with conducting, on average, five read-behinds per Reader throughout each half-scoring day; however, SRs 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 or not the response was to be double-blind scored. The practice was applied to all open-ended item types. Table 4-5 shows by which method(s) both common and equating open-ended item responses for each operational test were scored.

Table 4-5. 2009–10 NECAP: Frequency of Double-Blind Scoring by Grade and Content

<i>Grade</i>	<i>Content area</i>	<i>Responses double-blind scored</i>
3–8, 11	Reading	2% randomly
3–8, 11	Mathematics	2% randomly
5, 8, 11	Writing (WP)	100%
5, 8	Writing (CR)	2% randomly
All	Unreadable responses	100%
All	Blank responses	100%

If there was a discrepancy (a difference greater than 1 score point) between double-blind scores, the response was placed into an arbitration queue. Arbitration responses were reviewed by scoring leadership (SR or QAC) without knowledge of the two Readers' scores. Scoring leadership assigned the final score. Appendix D provides the NECAP 2009–10 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, Readers 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 2 of 5 responses of the first Recalibration Set, or exact on 2 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 1 of 5 responses of the first Recalibration Set, and/or exact on 3, were retrained by their SR 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 5 of 5 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 4, were allowed to begin scoring operational responses immediately, under the proviso that this Recalibration performance would be combined with that of the second Recalibration Set administered at the shift midpoint.

The results of both Recalibration Sets were combined with the expectation that Readers would have achieved an overall 80 percent-exact and 90 percent-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 WP 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. These reports are further discussed in the following section.

4.1.2.8 Reports Generated During Scoring

Due to the complexity of the 2009–10 NECAP administration, computer-generated reports were necessary to ensure that

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

The following reports were produced by iScore:

- **The Read-Behind Summary** showed the total number of read-behind responses for each Reader and noted the number and percentages of exact, adjacent, and discrepant scores with the SR/QAC. Scoring leadership could choose to generate this report by choosing options (such as “Today” “Past Week” and “Cumulative”) from a pull-down menu. The report could also be filtered to select data for a particular item or across all items. This report was used in conjunction with other reports to determine whether a Reader’s scores would be voided (i.e., sent back out to the floor to be rescored by other Readers). The benefit of this report is that it can reveal the degree to which an individual Reader agrees with their QAC or SR on how best to score live responses.
- **The Double-Blind Summary** showed the total number of double-scored responses of each Reader, and noted the number and percentages of exact, adjacent, and discrepant scores with second Readers. This report was used in conjunction with other reports to determine whether a Reader’s scores should be voided (i.e., sent back out to the floor to be rescored by other Readers).

The benefit of this report is that it can reveal the degree to which Readers are in agreement with each other about how best to score live responses.

- **The Accuracy Summary** combined read-behind and double-blind data, showing the total number for the Readers, their accuracy rates, and their score-point distributions.
- **The Embedded CRR Summary** showed, for each Reader (by item or across all items), the total number of responses scored, the number of embedded CRRs scored, and the numbers and percentages of exact, adjacent, and discrepant scores with the Chief Reader. This report was used in conjunction with other reports to determine whether a Reader's scores should be voided (i.e., sent back out to the floor to be rescored by other Readers). The benefit of this report is that it can reveal the degree to which an individual Reader agrees with their Chief Reader on how to best score live responses. Also, since embedded CRRs are administered during the first hours of scoring, this report can provide an early illustration of agreement between Readers and Chief Readers.
- **The Qualification Statistics Summary** listed each Reader by name and ID number, identified which Qualifying Set(s) they did and did not take and, for the ones taken, their pass rate. In addition to the pass rates of individuals, the report also showed numbers of Readers passing or failing a particular Qualifying Set. The QAC could use this report to determine how Readers within their scoring group performed on specific Qualifying Sets.
- **The Summary Statistics Report** showed the total number of student responses for an item, and identified, for the time at which the report was generated, the following:
 - the number of single and double-blind scorings that had been performed
 - the number of single and double-blind scorings yet to be performed

Chapter 5. CLASSICAL ITEM ANALYSES

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, 2004) and *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 1988) include standards for identifying high quality items. Test items should assess only knowledge or skills that are identified as part of the domain being measured and should avoid assessing irrelevant factors. They should also be unambiguous, and free of grammatical errors, potentially insensitive content or language, and other confounding characteristics. Further, items must not unfairly disadvantage test takers from particular racial, ethnic, or gender groups.

Both qualitative and quantitative approaches were taken to ensure that 2009–10 NECAP items met these standards. Qualitative work was discussed in Chapter 2 (“Development and Test Design”). This chapter summarizes several types of quantitative analyses that were carried out on the 2009–10 NECAP items; all analyses presented are based on the statewide administration in fall 2009:

- classical item statistics
- Differential Item Functioning (subgroup differences in item performance)
- dimensionality analyses

5.1 *Classical Difficulty and Discrimination Indices*

All 2009–10 NECAP items were evaluated in terms of item difficulty according to standard classical test theory (CTT) practice. The item p -value is the main index of item difficulty under the CTT framework. This index measures an item’s difficulty by averaging the proportion of points received across all students who took the item. Multiple-choice items were scored dichotomously (correct vs. incorrect), so for these items, the difficulty index is simply the proportion of students who correctly answered the item. Constructed response items were scored polytomously, where a student can achieve a score of 0, 1, 2, 3, or 4. 1-point short-answer items were scored 0 or 1 and 2-point short-answer items 0, 1, or 2. By computing the constructed-response and 2-point short-answer difficulty indices as the average proportion of points achieved, the indices for all item types are placed on a similar scale that ranges from 0.00 to 1.00. Although the p -value is traditionally called a measure of difficulty, it is properly interpreted as an easiness index, because larger values indicate easier items. An index of 1.00 indicates that every student received full credit for the item; such items provide little information about differences in student ability, but do indicate knowledge or skills that have been mastered by most students. Similarly, an index of 0.00 indicates that no student received credit for the item; such items provide little information about differences in student ability, but may indicate knowledge or skills that have not yet been mastered by most students.

To provide best measurement, difficulty indices generally should range from near-chance performance (i.e., 0.25 for four-option, multiple-choice items; essentially 0.00 for open-response items) to 0.90. Indices outside this range indicate items that were either too difficult or too easy for the target population. Nonetheless, 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.

Another desirable feature of an item is that the higher-achieving students perform better on the item than do lower-achieving students. The correlation between student performance on a single item and total test score is a commonly used measure of this characteristic of an 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 dichotomous items (multiple-choice and 1-point short-answer), this statistic is commonly referred to as a point-biserial correlation. The theoretical range of these statistics is -1.00 to 1.00 and their typical observed range is 0.20 to 0.60.

A discrimination index can be thought of as measuring how closely an item assesses the same knowledge and skills assessed by other items contributing to the criterion total score. That is, as a measure of construct consistency. In light of this interpretation, the selection of an appropriate criterion total score is crucial to the interpretation of the discrimination index. Because each form of the 2009–10 NECAP was constructed to be parallel in content, the criterion score selected for each item was the raw score total for each form. The analyses were conducted for each form separately.

Difficulty and discrimination indices (i.e., item level classical stats) for each item are provided in Appendix E. Item level statistics are summarized by form in Appendix F. The item difficulty and discrimination indices are within 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 near-zero discrimination indices, but none were negative.

Attempting to compare CTT difficulty indices across content areas or grade levels is a thorny proposition, because the statistics are population dependent. Such direct comparisons would require that either items or students were common across comparisons, and since that is not the case, it cannot be determined whether differences in performance are because of real differences in student ability or differences in item difficulty or both. With this caveat in mind, it appears generally that students in higher grades found their mathematics items more difficult than did students in lower grades. Mathematics items also appeared to be more difficult than items in other content areas across grades. Comparing difficulty indices across item types is also suspect, because multiple-choice items can be answered correctly by guessing. That the difficulty indices for the dichotomous items tended to be higher (i.e., the items are easier) than those for the polytomous items is not surprising. Similarly, discrimination indices for the polytomous items were larger than those for

the dichotomous items due to 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.

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 actions should be taken to make certain that differences in performance are because of construct-relevant, rather than irrelevant, factors. *Standards for Educational and Psychological Testing* (AERA, 1999) includes similar guidelines.

The standardization differential item functioning (DIF) procedure (Dorans and Kulick, 1986) 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. The criterion (matching) score for 2009–10 NECAP was computed two ways. For common items, total score was the sum of scores on common items. Total score for matrix items was the sum of item scores on common and matrix items (excluding field test items). Based on experience, this dual definition of criterion scores has worked well in identifying problematic common and matrix items.

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 and short-answer items, and the index is adjusted to the same scale for constructed-response items. Dorans and Holland (1993) suggested that index values between -0.05 and 0.05 should be considered negligible. The preponderance of 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, 0.10] range (i.e., “high” DIF) are more unusual and should be examined very carefully.²

For the 2009–10 NECAP tests, three subgroup comparisons were evaluated for DIF:

- Male versus female

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 to determine whether to include the flagged item in a future operational test administration.

- White versus African American
- White versus Hispanic

Other race/ethnicity groups (e.g., Asians) were not analyzed using DIF procedures, because limited sample sizes would have inflated type I error rates. Appendix G presents the number of items classified into each DIF category by test form and item type. Appendix H presents the number of items classified into each DIF category that favor males or females, by item type.

5.3 Dimensionality Analyses

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 presence of just such a dominant primary dimension is the psychometric assumption that provides the foundation for the unidimensional 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 2009–10 NECAP common items for mathematics and reading are reported below. (Note: only common items were analyzed since they are used for score reporting.)

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 over all possible conditioning scores. When a test is strictly unidimensional, all conditional covariances are expected to take on values within random noise of zero, indicating statistically independent item responses for examinees with equal expected scores. Non-zero conditional covariances are essentially violations of the principle of local independence, and local *dependence* implies multidimensionality. Thus, 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 fit a systematic pattern of positive conditional covariances for pairs of items from the same cluster and negative conditional covariances from different clusters. Next, the clusters from the training sample 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 weak to moderate multidimensionality, values of 0.4 to 1.0 moderate to strong multidimensionality, and values greater than 1.0 very strong multidimensionality.

DIMTEST and DETECT were applied to the 2009-10 NECAP. The data for each grade and content area were split into a training sample and a cross-validation sample. Every grade/content area combination had at least 30,000 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 15,000 students. DIMTEST was then applied to every grade/content area. 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 hypothesis tests were that the null hypothesis was strongly rejected for every dataset (p -value < 0.00005 in all cases). 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 5-1 below displays the multidimensional effect size estimates from DETECT.

Table 5-1. 2009-10 NECAP: Multidimensionality Effect Sizes by Grade and Content Area

Grade	Content area	Multidimensionality effect size	
		Fall 2008	Fall 2009
3	Mathematics	0.12	0.17
	Reading	0.22	0.18
4	Mathematics	0.14	0.13
	Reading	0.35	0.18

continued

Grade	Content area	Multidimensionality effect size	
		Fall 2008	Fall 2009
5	Mathematics	0.21	0.15
	Reading	0.18	0.18
6	Mathematics	0.17	0.16
	Reading	0.24	0.19
7	Mathematics	0.19	0.16
	Reading	0.23	0.20
8	Mathematics	0.15	0.16
	Reading	0.20	0.32
11	Mathematics	0.17	0.12
	Reading	0.31	0.28

All of the DETECT values indicated very weak to weak multidimensionality, except for grade 8 reading whose value of 0.32 is near the borderline between weak and moderate. The reading test forms tended to show slightly greater multidimensionality than did the mathematics. The average DETECT value for reading was 0.22 as compared to 0.15 for mathematics, which indicate very weak and weak multidimensionality, respectively. Also shown in Table 5-1 are the values reported in last year's dimensionality analyses. The averages for mathematics and reading are seen to be very similar to those from last year, which were 0.25 and 0.16 for reading and mathematics, respectively. We also investigated how DETECT divided the tests into clusters to see if there were any discernable patterns with respect to the item types (i.e., multiple-choice, short answer, and constructed response). The mathematics clusters showed no discernable patterns. For reading, however, there was a strong tendency for the multiple-choice items to cluster separately from the remaining items. Despite this multidimensionality between the multiple-choice items and remaining items for reading, the effect sizes were not strong enough to warrant further investigation. These trends and conclusions are the same as were reported for last year's tests.

Chapter 6. IRT SCALING AND EQUATING

6.1 Item Response Theory Scaling

All NECAP items were calibrated using item response theory (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, it is assumed that all items are independent measures of the same construct (i.e., of the same θ). Another way to think of θ is as a mathematical representation of the latent trait of interest. Several common IRT models are used to specify the relationship between θ and p (Hambleton and van der Linden, 1997; Hambleton and 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 NECAP 2009–10, the three-parameter logistic (3PL) model was used for dichotomous items and the graded-response model (GRM) was used for polytomous items. The 3PL model for dichotomous items can be defined as follows (note that for 1-point short-answer items, the c parameter is set to zero, thus the model becomes 2PL):

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

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

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

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

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

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

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

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 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 Analyses

The previous section introduced IRT and gave a thorough description of the topic. It was discussed there that all 2009–10 NECAP items were calibrated using IRT and that the calibrated item parameters were ultimately used to scale both the items and students onto a common framework. The results of those analyses are presented in Appendix I.

The tables in Appendix I give the IRT item parameters of all common items on the 2009–10 NECAP tests by grade and content area. Accompanying the parameter tables are graphs of the corresponding test characteristic curves (TCCs) and test information functions (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 that the test provides at each value of θ_j . Information functions depict test precision across the entire latent trait continuum. There is an inverse relationship between the information of a test and its standard error of measurement (SEM). For long tests, the SEM at a given θ_j is approximately equal to the inverse of the square root of the statistical information at θ_j (Hambleton, Swaminathan, & Rogers, 1991), as follows:

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

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

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 2009–10 administration of NECAP used a raw score-to-theta equating procedure in which test forms are equated every year to the theta scale of the reference test forms. (In the case of NECAP, the reference forms are those from the 2005–06 administration for grades 3 through 8 and 2007–08 for grade 11.) 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 2009–10 NECAP tests are not equivalent to the groups who took them in the reference years (2007–08 or 2005–06, as described above). IRT is particularly useful for equating scenarios that involve nonequivalent groups (Allen and 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 through utilizing a set of anchor items (also called equating items). The NECAP uses an *external* anchor test design, which means that the equating items are not counted toward students’ test scores. 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 matrixed across forms.

Item parameter estimates for 2009–10 were placed on the 2008–09 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 2008–09 and 2009–10 NECAP tests should have the same item parameters. After the item parameters for each 2009–10 NECAP mathematics and reading test were estimated using PARSCALE (Muraki and Bock, 2003), as described earlier, 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 2009–10 TCC was as close as possible to that of 2008–09. The transformation constants are presented in Table 6-1. It should be noted that grades 5 and 8 writing were excluded from the equating process; writing test forms are equated through the scoring rubric.

Table 6-1. 2009–10 NECAP: Stocking & Lord Transformation Constants

<i>Content area</i>	<i>Grade</i>	<i>A-slope</i>	<i>B-intercept</i>
	3	1.031	0.141
	4	1.085	0.083
	5	1.034	0.160
Mathematics	6	1.087	0.223
	7	1.037	0.230
	8	0.988	0.237
	11	1.004	0.119

continued

<i>Content area</i>	<i>Grade</i>	<i>A-slope</i>	<i>B-intercept</i>
Reading	3	1.028	0.128
	4	1.035	0.160
	5	1.002	0.193
	6	1.047	0.036
	7	1.056	0.109
	8	1.099	0.163
	11	1.045	0.242

A = Slope, B = Intercept

The next administration of NECAP (2010–11) will be scaled to the 2009–10 administration by the same equating method described above.

6.4 **Equating Results**

An Equating Report was submitted to the NECAP state testing directors for their approval prior to production of student reports. Various elements from the Equating Report are presented throughout this technical report and its appendices.

In addition to the equating and scaling activities described in the previous section (IRT calibrations and execution of the Stocking and Lord equating procedure) various quality control procedures were implemented within the Psychometrics Department at Measured Progress and reviewed with the NECAP state testing directors and NECAP Technical Advisory Committee. A variety of quality control activities were undertaken during the IRT calibration, equating, and scaling, and various results are presented throughout this report.

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 in order for the solution to converge fell within acceptable ranges.

Table 6-2. 2009–10 NECAP: Number of Newton Cycles Required for Convergence

<i>Content area</i>	<i>Grade</i>	<i>Cycles</i>
Mathematics	3	24
	4	42
	5	44
	6	49
	7	56
	8	58
	11	94
Reading	3	50
	4	48
	5	49
	6	47
	7	46
	8	46
	11	50

The number of items that required intervention during the IRT analysis, presented in Table 6-3, was very typical across the various grades and content areas. Appendix J presents the results from the Delta analysis. This procedure was used to evaluate adequacy of equating items, and the discard status presented in the appendix indicates whether or not the item was used 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, 0.80.

Table 6-3. 2009–10 NECAP: Number of Items that Required Intervention During IRT Calibration and Equating

<i>Content area</i>	<i>Grade</i>	<i>IREF</i>	<i>Reasons</i>	<i>Action</i>
Mathematics	3	119911	c parameter	c = 0
		119868	c parameter	c = 0
		119912	c parameter	c = 0
		121349	c parameter	c = 0
		119935	Delta Analysis	Removed from equating
		255964	Delta Analysis	Removed from equating
	4	120066	c parameter	c = 0
		120102	c parameter	c = 0
		120293	c parameter	c = 0
		120226	c parameter	c = 0
		124620	c parameter	c = 0
		120061	a parameter	a set to initial
	5	198442	Delta Analysis	Removed from equating
		120733	c parameter	c = 0
		124973	c parameter	c = 0
		203361	c parameter	c = 0
		198494	Delta Analysis	Removed from equating
	6	203367	IRT Plot Outlier	No action taken
		125025	c parameter	c = 0
		122249	c parameter	c = 0
		123501	c parameter	c = 0
		119232	c parameter	c = 0
		119326	Delta Analysis	Removed from equating
		203483	Delta Analysis	Removed from equating
	7	198622	Delta Analysis	Removed from equating
		120329	c parameter	c = 0
		125286	c parameter	c = 0
		120327	c parameter	c = 0
		120402	c parameter	c = 0
		206146	c parameter	c = 0
269069		Delta Analysis	Removed from equating	
120524		a parameter	a set to initial	
8	256118	a parameter	a set to initial	
	121056	c parameter	c = 0	
	121040	Delta Analysis	Removed from equating	
	269098	Delta Analysis	Removed from equating	
11	269172	Delta Analysis	Removed from equating	
	119423	c parameter	c = 0	
	260001	Delta Analysis	Removed from equating	

continued

<i>Content area</i>	<i>Grade</i>	<i>IREF</i>	<i>Reasons</i>	<i>Action</i>
Reading	3	117744	c parameter	c = 0
		117676	c parameter	c = 0
		117793	Delta Analysis	Removed from equating
		225242	a parameter	a set to initial
	4	117998	c parameter	c = 0
		118003	c parameter	c = 0
		117959	Delta Analysis	Removed from equating
		118028	a parameter	a set to initial
	5	118082	c parameter	c = 0
		118127	c parameter	c = 0
		118179	c parameter	c = 0
		118180	c parameter	c = 0
		118181	c parameter	c = 0
		118192	c parameter	c = 0
		118052	c parameter	c = 0
		118053	c parameter	c = 0
		128931	c parameter	c = 0
		128932	c parameter	c = 0
	6	118226	c parameter	c = 0
		118227	c parameter	c = 0
		118366	c parameter	c = 0
		118365	c parameter	c = 0
		269508	IRT Plot Outlier	Removed from equating
	7	118284	Delta Analysis	Removed from equating
118492		c parameter	c = 0	
118495		c parameter	c = 0	
118467		c parameter	c = 0	
118468		c parameter	c = 0	
118500		c parameter	c = 0	
118512		c parameter	c = 0	
118570		c parameter	c = 0	
118448		c parameter	c = 0	
118535		Delta Analysis	Removed from equating	
Reading	8	201482	c parameter	c = 0
		118743	c parameter	c = 0
		118714	c parameter	c = 0
		118716	c parameter	c = 0
		118719	c parameter	c = 0
		118588	c parameter	c = 0
		118590	c parameter	c = 0
		118593	c parameter	c = 0
		118732	c parameter	c = 0
		118748	c parameter	c = 0
		118674	c parameter	c = 0
		118676	c parameter	c = 0
		204100	c parameter	c = 0
		11	129591	c parameter
129602	c parameter		c = 0	
118812	c parameter		c = 0	
118814	c parameter		c = 0	
118848	c parameter		c = 0	
118852	c parameter		c = 0	
118836	c parameter		c = 0	
118840	c parameter		c = 0	
118843	c parameter	c = 0		

continued

<i>Content area</i>	<i>Grade</i>	<i>IREF</i>	<i>Reasons</i>	<i>Action</i>
Reading	11	118884 269456	c parameter IRT Plot Outlier	c = 0 Removed from equating

c parameter: PARSCALE had difficulty estimating a c parameter.

Delta Analysis: Item was flagged for removal in the delta analysis.

a-parameter: PARSCALE had difficulty estimating an a parameter

c = X: c parameter fixed to X during item calibration.

Removed from equating: Item was excluded from equating item set.

A set to initial: a parameter value was fixed to initial estimate based on classical statistics.

No action taken: no action outside of normal procedures was taken due to the listed reason.

6.5 Achievement Standards

NECAP standards to establish achievement level cut scores in reading, mathematics, and writing for grades 3 through 8 were set in January 2006 and for grade 11 in January 2008. The standard setting meetings and results were discussed in the technical reports of those years. As alluded to in the discussion of equating above, the respective NECAP reporting scales were established during those base years, and the forms serve as the reference for subsequent equating. The θ -metric cut scores that emerged from the standard setting meetings will remain fixed throughout the assessment program unless standards are reset for any reason.

6.6 Reported Scaled Scores

6.6.1 Description of Scale

Because the θ scale used in IRT calibrations is not readily understood by most stakeholders, reporting scales were developed for the NECAP tests. The reporting scales are simple linear transformations of the underlying θ scale. The scales were developed such that they ranged 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 was set 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 2009–10 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 2009–10 NECAP tests can be expressed in raw or scaled scores.

In Figure 6-1, two-way arrows depict how raw scores (the vertical axis) map through the S-shaped TCC to corresponding scores on the θ scale (horizontal axis), which in turn map directly to scaled scores.

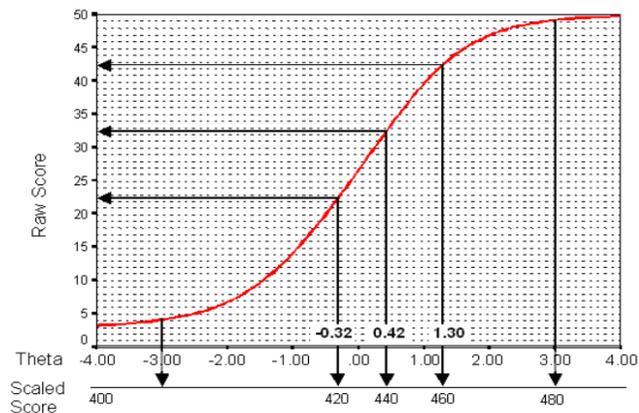


Figure 6-1. Conceptual Illustration of Raw Score to Theta Scaled Score Transformation Using a Test Characteristic Curve

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 content areas. The raw cut score between *Partially Proficient* and *Proficient* could be, for example, 35 in mathematics but 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.

6.6.2 Calculations

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

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

where
 m is the slope, and
 b is the intercept

A separate linear transformation is used for each grade content combination. For NECAP tests, each line is determined by fixing both the *Partially Proficient/Proficient* cut score and the bottom of the scale; that is, the X40 value and the X00 value (e.g., 340 and 300, respectively, for grade 3). The lowest scaled score is at a location on the θ scale beyond the scaling of all the items across the various grade content combinations. To determine its location, a chance raw score (approximately equal to a student's expected performance by

guessing), and a raw score of 0, are both mapped to a θ value of -4.0 . At the other extreme, the maximum possible raw score is assigned the scaled score of X80 (e.g., 380 in the case of grade 3).

Because only the *Partially Proficient/Proficient* cut score is fixed within the θ scaled score space, the cut scores between *Substantially Below Proficient* and *Partially Proficient* (SBP/PP) and between *Proficient* and *Proficient With Distinction* (P/PWD) vary across the grade/content combinations.

Table 6-4 presents the scaled score cuts for each grade content combination (i.e., the minimum scaled score for getting into the next achievement level). It is important to repeat that the values in Table 6-4 do not change from year to year, because the cut scores along the θ scale do not change unless standards are reset. Also, in a given year it may not be possible to attain a particular scaled score, but the scaled score cuts will remain the same.

Table 6-4. 2009–10 NECAP: Scaled Score Cuts and Minimum and Maximum Scores by Grade and Content Area

Content area	Grade	Min	Scaled score cuts			Max
			SBP/PP	PP/P	P/PWD	
Mathematics	3	300	332	340	353	380
	4	400	431	440	455	480
	5	500	533	540	554	580
	6	600	633	640	653	680
	7	700	734	740	752	780
	8	800	834	840	852	880
	11	1100	1134	1140	1152	1180
Reading	3	300	331	340	357	380
	4	400	431	440	456	480
	5	500	530	540	556	580
	6	600	629	640	659	680
	7	700	729	740	760	780
	8	800	828	840	859	880
	11	1100	1130	1140	1154	1180

SBP = *Substantially Below Proficient*, PP = *Partially Proficient*; P = *Proficient*, PWD = *Proficient With Distinction* *Scaled scores are not produced for grade 11 writing.

Table 6-5 shows the cut scores on θ and the slope and intercept terms used to calculate the scaled scores. Note that the values in Table 6-5 will not change unless the standards are reset.

Table 6-5. 2009–10 NECAP: Cut Scores (on θ Metric), Intercept, and Slope by Grade and Content Area

Content area	Grade	θ Cuts			Intercept	Slope
		SBP/PP	PP/P	P/PWD		
Mathematics	3	-1.0381	-0.2685	0.9704	342.8782	10.7195
	4	-1.1504	-0.3779	0.9493	444.1727	11.0432
	5	-0.9279	-0.2846	1.0313	543.0634	10.7659
	6	-0.8743	-0.2237	1.0343	642.3690	10.5922
	7	-0.7080	-0.0787	1.0995	740.8028	10.2007
	8	-0.6444	-0.0286	1.1178	840.2881	10.0720
	11	-0.1169	0.6190	2.0586	1134.640	8.6600

continued

Content area	Grade	θ Cuts			Intercept	Slope
		SBP/PP	PP/P	P/PWD		
Reading	3	-1.3229	-0.4970	1.0307	345.6751	11.4188
	4	-1.1730	-0.3142	1.1473	443.4098	10.8525
	5	-1.3355	-0.4276	1.0404	544.7878	11.1970
	6	-1.4780	-0.5180	1.1255	645.9499	11.4875
	7	-1.4833	-0.5223	1.2058	746.0074	11.5019
	8	-1.5251	-0.5224	1.1344	846.0087	11.5022
	11	-1.2071	-0.3099	1.0038	1143.3600	10.8399

SBP = Substantially Below Proficient; PP = Partially Proficient; P = Proficient; PWD = Proficient With Distinction

Table 6-6 shows the raw scores associated with the cut scores for each performance level by grade and content area. In order to evaluate changes in test difficulty, the results can be compared to the previous year's results which are also reflected in Table 6-6.

Table 6-6. 2009–10 NECAP: Cut Scores on Raw Score Metric for each Performance Level, by Grade and Content Area

Content area	Grade	Year 1				Year 2			
		SbP/PP	PP/P	P/PwD	Max	SbP/PP	PP/P	P/PwD	Max
Mathematics	3	28	39	55	65	28	39	55	65
	4	28	39	55	65	30	40	55	65
	5	23	31	52	66	22	30	50	66
	6	19	27	47	66	19	28	47	66
	7	19	27	47	66	18	26	44	66
	8	20	29	49	66	18	27	47	66
	11	19	32	56	64	19	31	54	64
Reading	3	21	30	44	52	21	31	45	52
	4	21	30	42	52	23	31	43	52
	5	17	26	38	52	20	28	39	52
	6	20	30	42	52	20	30	41	52
	7	21	30	42	52	19	29	42	52
	8	22	31	42	52	20	30	42	52
	11	22	31	42	52	21	30	41	52

SBP = Substantially Below Proficient; PP = Partially Proficient; P = Proficient; PWD = Proficient With Distinction

Note: The values presented are not the cut scores per se. The cut scores are defined on the θ metric and do not change from year to year. The values in this table represent the raw scores associated with the cut scores, and these values are found via a TCC mapping.

Appendix K contains raw score to scaled score lookup tables. These are the actual tables that were used to determine student scaled scores, error bands, and achievement levels.

6.6.3 Distributions

Appendix L contains scaled score distributions 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 impact a student's score are referred to as measurement error. Any assessment includes some amount of measurement error; that is, no measurement is perfect. This is true of all academic assessments—some students will receive scores that underestimate their true ability, and other students will receive scores that overestimate their true ability. When tests have a high amount of measurement error, student scores are very unstable. Students with high ability may get low scores or vice versa. Consequently, one cannot reliably measure a student's true level of ability with such a test. Assessments that have less measurement error (i.e., errors made are small on average and student scores on such a test will consistently represent their ability) are described as reliable.

There are a number of ways to estimate an assessment's reliability. One possible approach is to give the same test to the same students at two different points in time. If students receive the same scores on each test, then the extraneous factors affecting performance are small and the test is reliable. (This is referred to as "test-retest reliability.") A potential problem with this approach is that students may remember items from the first administration or may have gained (or lost) knowledge or skills in the interim between the two administrations. A solution to the "remembering items" problem is to give a different, but parallel test at the second administration. If student scores on each test correlate highly the test is considered reliable. (This is known as "alternate forms reliability," because an alternate form of the test is used in each administration.) This approach, however, does not address the problem that students may have gained (or lost) knowledge or skills in the interim between the two administrations. In addition, the practical challenges of developing and administering parallel forms generally preclude the use of parallel forms reliability indices. One way to address the latter 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 of 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 halves will result in a different correlation. Another problem with the split-half method of calculating reliability is that it underestimates reliability, because test length is cut in half. All else being equal, a shorter

test is less reliable than a longer test. Cronbach (1951) provided a statistic, α (alpha), which eliminates the problem of the split-half method by comparing individual item variances to total test variance. Cronbach's α was used to assess the reliability of the 2009–10 NECAP tests:

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

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 content area and grade. (Statistics are based on common items only.)

Table 7-1. 2009–10 NECAP: Common Item Raw Score Descriptives, Reliability, and SEM by Grade and Content Area

Grade	Content area	N	Possible score	Min score	Max score	Mean score	Score SD	Reliability (α)	SEM
3	Mathematics	44,395	65	0	65	42.81	12.77	0.93	3.42
	Reading	44,280	52	0	52	35.64	9.63	0.89	3.26
4	Mathematics	44,069	65	0	65	43.24	12.01	0.92	3.41
	Reading	43,956	52	0	52	34.34	8.83	0.89	2.99
5	Mathematics	44,295	66	0	66	36.27	13.91	0.92	3.90
	Reading	44,178	52	0	51	31.66	7.75	0.87	2.85
6	Mathematics	46,266	66	0	66	34.03	14.08	0.92	3.96
	Reading	46,159	52	0	52	32.68	8.31	0.89	2.82
7	Mathematics	46,632	66	0	66	30.63	13.76	0.92	3.82
	Reading	46,538	52	0	52	32.58	8.70	0.90	2.81
8	Mathematics	47,188	66	0	66	31.80	14.39	0.93	3.90
	Reading	47,099	52	0	52	33.84	8.76	0.88	3.02
11	Mathematics	32,635	64	0	64	24.41	13.51	0.93	3.56
	Reading	32,720	52	0	52	33.61	9.12	0.90	2.92
	Writing	32,730	12	0	12	6.57	2.05		

For mathematics, the reliability coefficient ranged from 0.92 to 0.93 and for reading, from 0.87 to 0.90. 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. The grades 5 and 8 writing tests were omitted from the reliability analysis as there were no constructed test forms per se. This year saw only pilot testing for

future writing assessments. Reliability is not computed for grade 11 writing, as the test design does not support this statistic.

The α coefficients, broken down by subgroup, item type, and reporting category, are presented in Appendix M. These detailed α coefficient results are discussed in sections 7.2 through 7.4.

7.2 *Subgroup Reliability*

The reliability coefficients discussed in the previous section were based on the overall population of students who took the 2009–10 NECAP tests. Subgroup Cronbach’s α ’s were calculated using the formula defined above using only the members of the subgroup in question in the computations. For mathematics, subgroup reliabilities ranged from 0.87 to 0.96, for reading from 0.71 to 0.94

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 M 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 and 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 *Item Type Reliability*

Another approach to estimating the reliability for a test with differing item types (i.e., multiple-choice and constructed-response) is to assume that at least a small, but important, degree of unique variance is associated with item type (Feldt and Brennan, 1989), in contrast to Cronbach’s α , which assumes that there are no such local or clustered dependencies. A stratified version of coefficient α corrects for this problem by using the following formula:

$$\alpha_{strat} = 1 - \frac{\sum_{j=1}^k \sigma_{x_j}^2 (1 - \alpha_j)}{\sigma_x^2}$$

where

j indexes the subtests or categories,

$\sigma_{x_j}^2$ represents the variance of each of the k individual subtests or categories,

α_j is the unstratified Cronbach’s α coefficient for each subtest, and

σ_x^2 represents the total test variance.

Stratified α was calculated separately for all grades and content areas except grade 11 writing. The results of stratification based on item type (multiple-choice versus constructed-response) are presented in Table 7-2. Not surprisingly, reliabilities were higher on the full test than on subsets of items. Similar analyses done separately by form can be found in Appendix M.

Table 7-2. 2009–10 NECAP: Common Item Cronbach’s Alpha Reliability by Grade and Content Area—Overall by Item Type and Stratified by Item Type

Grade	Content area	All	MC	CR			Stratified α
		α	α	N	α	N (poss.)	
3	Mathematics	0.93	0.88	35	0.86	20 (30)	0.93
	Reading	0.89	0.87	28	0.74	6 (24)	0.90
4	Mathematics	0.92	0.87	35	0.84	20 (30)	0.92
	Reading	0.89	0.87	28	0.72	6 (24)	0.89
5	Mathematics	0.92	0.87	32	0.86	16 (34)	0.93
	Reading	0.86	0.81	28	0.83	6 (24)	0.88
6	Mathematics	0.92	0.87	32	0.86	16 (34)	0.93
	Reading	0.88	0.85	28	0.85	6 (24)	0.91
7	Mathematics	0.92	0.86	32	0.87	16 (34)	0.93
	Reading	0.90	0.86	28	0.87	6 (24)	0.92
8	Mathematics	0.93	0.87	32	0.87	16 (34)	0.93
	Reading	0.88	0.83	28	0.88	6 (24)	0.91
11	Mathematics	0.93	0.83	24	0.9	22 (40)	0.93
	Reading	0.90	0.85	28	0.91	6 (24)	0.93

All = MC and CR; MC = multiple-choice; CR = constructed-response= number of items; poss. = total possible constructed-response points

7.4 Reporting Categories Reliability

In Section 7.3, the reliability coefficients were calculated based on form and item type. Item type represents just one way of breaking an overall test into subtests. Of even more interest are reliabilities for the reporting categories within NECAP content areas described in Chapter 2. Cronbach’s α coefficients for reporting categories were calculated via the same alpha formula defined at the beginning of Chapter 7 using just the items of a given reporting category in the computations. These results are presented in Appendix M. Once again, as expected, because they are based on a subset of items rather than the full test, computed reporting category reliabilities were lower (sometimes substantially so) than were overall test reliabilities, and interpretations should take this into account.

For mathematics, reporting category reliabilities ranged from 0.58 to 0.89, for reading from 0.51 to 0.84. In general, the reporting category reliabilities were lower than those based on the total test and approximately to the degree one would expect based on CTT. Qualitative differences between grades and content areas once again preclude valid inferences about the quality of the full test based on statistical comparisons among subtests.

7.5 **Reliability of Achievement Level Categorization**

All test scores contain measurement error; thus, classifications based on test scores are also subject to measurement error. After the 2009–10 NECAP achievement levels were specified and students classified into those levels, empirical analyses were conducted to determine the statistical accuracy and consistency of the classifications. For every 2009–10 NECAP grade and content area, each student was classified into one of the following achievement levels: *Substantially Below Proficient* (SBP), *Partially Proficient* (PP), *Proficient* (P), or *Proficient With Distinction* (PWD). This section of the report explains the methodologies used to assess the reliability of classification decisions and presents the results.

7.5.1 **Accuracy and Consistency**

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 2009–10 NECAP because it is easily adaptable to all types of testing formats, including mixed format tests.

7.5.2 **Calculating Accuracy**

The accuracy and consistency estimates reported below make use of “true scores” in the CTT 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 method, estimated true scores are used to classify students into their “true” achievement level.

For the 2009-10 NECAP, after various technical adjustments were made (described in Livingston and Lewis, 1995), a 4×4 contingency table of accuracy was created for each content area and grade, where cell $[i, j]$ represented the estimated proportion of students whose true score fell into achievement level i (where $i = 1 - 4$) and whose observed score fell into achievement level j (where $j = 1 - 4$). The sum of the diagonal entries, i.e., the proportion of students whose true and observed achievement levels matched one another, signified overall accuracy.

7.5.3 Calculating Consistency

To estimate 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 4×4 contingency table was created for each content area and grade and populated by the proportion of students who would be classified into each combination of achievement levels 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 achievement level i (where $i = 1 - 4$), and whose observed score on the second form would fall into achievement level j (where $j = 1 - 4$). The sum of the diagonal entries, i.e., the proportion of students classified by the two forms into exactly the same achievement level, signified overall consistency.

7.5.4 Calculating Kappa

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

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

where

C_i is the proportion of students whose observed achievement level would be Level i (where $i = 1 - 4$) on the first hypothetical parallel form of the test;

C_i is the proportion of students whose observed achievement level would be Level i (where $i = 1 - 4$) on the second hypothetical parallel form of the test;

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.

7.5.5 Results of Accuracy, Consistency, and Kappa Analyses

The accuracy and consistency analyses described above are provided in Table N-1 of Appendix N. The table includes overall accuracy and consistency indices, including kappa. Accuracy and consistency values conditional upon achievement level are also provided in Table N-1. For these calculations, the denominator is the proportion of students associated with a given achievement level. For example, the conditional accuracy value is 0.70 for the PP achievement level for mathematics grade 3. This figure indicates that among the students whose true scores placed them in the PP achievement level, 70% of them would be expected to be in the PP achievement level when categorized according to their observed score. Similarly, the corresponding consistency value of 0.61 indicates that 61% of students with observed scores in PP would be expected to score in the PP achievement level again if a second, parallel test form were used.

For some testing situations, the greatest concern may be decisions around level thresholds. For example, if a college gave credit to students who achieved an Advanced Placement test score of 4 or 5, but not to students with scores of 1, 2, or 3, one might be interested in the accuracy of the dichotomous decision below-4 versus 4-or-above. For the 2009–10 NECAP, Table N-2 in Appendix N provides accuracy and consistency estimates at each cut point as well as false positive and false negative decision rates. (A false positive is the proportion of students whose observed scores were above the cut and whose true scores were below the cut. A false negative is the proportion of students whose observed scores were below the cut and whose true scores were above the cut.)

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

Descriptive statistics relating to the decision accuracy and consistency (DAC) of the 2009–10 NECAP tests can be derived from Table N-1. For mathematics, overall accuracy ranged from 0.80 to 0.84; overall consistency ranged from 0.73 to 0.77; the kappa statistic ranged from 0.61 to 0.67. For reading, overall accuracy ranged from 0.78 to 0.83; overall consistency ranged from 0.70 to 0.76; the kappa statistic ranged from 0.55 to 0.64. Finally, for writing grades 5 and 8, DAC analysis was not possible because there were no cut scores, and for writing grade 11, DAC analysis was not possible because the assessment consisted of only one writing prompt.

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 2009 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 2009–10 NECAP:

- Student Report
- Item Analysis Report
- School and District Results Report
- School and District Summary Report
- School and District Student-Level Data File

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 O.

8.3 *Student Report*

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 student report and the achievement level definitions. The reverse side of the student report provides a complete picture of an individual student’s performance on the NECAP, 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 11 writing only reports extended-response as a category.

During scoring of the writing prompt, each scorer selects up to three comments about the student's writing performance. The comments are selected from a predetermined list produced by the writing representatives from each state's Department of Education. These scorers' comments are presented in a box next to the writing results.

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 percentage correct for each multiple-choice item and average scores for the short-answer and constructed-response items are shown for the school, district, and state.

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 (page 2), the results for the content areas (pages 3, 5, and 7), and the disaggregated content area results (pages 4, 6, and 8).

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, 2008. The total number of students enrolled is defined as the number of students tested plus the number of students not tested.

Because students who were not tested did not participate, average school scores were not affected by non-tested 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 testing program.

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 2008–09 school year.

- **First-Year LEP:** Students in this category are defined as being new to the United States after October 1, 2008 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, 2009 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, 2009 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.

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). The purpose of these sections is to help schools to determine the extent to which their curricula are effective in helping students to achieve the particular standards and benchmarks contained in the *Grade-Level and Grade-Span Expectations*. Information about each content area (reading and mathematics for all grades and writing for grade 11) 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.

Information about each content area reporting category for reading and mathematics in all grades and writing in grade 11 includes the following:

- The total possible points for that category. 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 percent of total possible points for the school, state, and district. In this graphic display, there are symbols representing school, district, and state performance. 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).
- For grade 11 writing only, a column showing the number of prompts for each subtopic (strand) is also provided, as well as the distribution of score points across prompts within each strand in terms of percentages for the school, district, and state.

The disaggregated content area results pages present the relationship between performance and student reporting variables (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 by each reporting category.

The list of student reporting categories is as follows:

- All Students
- Gender
- Primary Race/Ethnicity
- LEP Status (limited English proficiency)
- IEP
- SES (socioeconomic status)
- Migrant
- Title I
- 504 Plan

The data for achievement levels and mean scaled score 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 for New Hampshire and Vermont, no data were reported for the 504 Plan in any of the content areas. In addition, for Vermont, no data were reported for Title I in any of the content areas.

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 *Grade-Level Expectations* and *Grad-Span Expectations*.

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 report 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 two 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 (for example, 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 2009–10 test administration were accessible online via the new 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. Sample interactive reports are provided in Appendix O. 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 include 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/content area. A more detailed description of the information included on this report can be found in section 8.4 of this document. Please note that when advanced filtering criteria are applied by the user, the School and District Percent Correct/Average Score rows at the bottom of the report are blanked out and only the Group row and the State row for the group selected will contain data. This report can be saved, printed, or exported as a PDF.

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/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/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/percent 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 this report. If a user clicks on a particular magnifying glass icon next to a released item number, a pop-up box will open displaying the released item.

8.8.1.4 Longitudinal Data Report

The Longitudinal Data report is a confidential student-level report that provides individual student performance data for multiple test administrations. Fall 2009 NECAP scores and achievement levels are provided for each tested student in reading, mathematics, and writing. In addition, fall NECAP 2008 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 only show fall 2009 NECAP scores and achievement levels in reading and mathematics since this is the first 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 in their school such as teachers, curriculum coordinators or special education coordinators. 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 2009–10 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 2009 administration of the NECAP is found in Appendix P.

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.

8.10 ***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 Services and Static Reporting (DSSR) and Psychometrics and Research (P&R) 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 receives scaled scores and/or is included in different levels of aggregation are also parallel processed. Using the decision rules document, two data analysts independently write a computer program that assigns students' exclusions. For each 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 2009–10 *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.

The NECAP tests are based on and aligned with the content standards and performance indicators in the GLEs and GSEs for mathematics, reading, and writing. Inferences about student achievement on the content standards are intended from NECAP results, which in turn serve the evaluation of school accountability and inform the improvement of programs and instruction.

Standards for Educational and Psychological Testing (AERA, 1999) provides a framework for describing sources of evidence that should be considered when evaluating validity. These sources include evidence on the following five general areas: test content, response processes, internal structure, consequences of testing, and relationship to other variables. Although each of these sources may speak to a different *aspect* of validity, they are not distinct *types* of validity. Instead, each contributes to a body of evidence about the comprehensive validity of score interpretations.

A measure of test content validity is to determine how well the test’s tasks represent the curriculum and standards for each content area and grade level. This is informed by the item development process, including how test blueprints and test items align with the curriculum and standards. Validation through this content lens was extensively described in Chapter 2. In other words, the element’s components discussed in the chapter—item alignment with 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 content-based validity evidence. Every NECAP test question or prompt was aligned by educators to specific content standards and underwent several rounds of review for content fidelity and appropriateness. Items of multiple formats (multiple-choice, short-answer, and constructed-response) were presented to students. Finally, tests were administered according to mandated standardized procedures, with allowable accommodations, and all test coordinators and test administrators were required to familiarize themselves with and adhere to all of the procedures outlined in the *NECAP Principal/Test Coordinator Manual* and *Test Administrator Manuals*.

The scoring information in Chapter 4 describes the steps taken to train and monitor hand-scorers as well as the quality control procedures related to machine scanning and scoring. Additional studies might be informative on student response processes. For example, think-aloud protocols could be used to investigate students’ cognitive processes when confronting test items.

Evidence on internal structure is extensively detailed in the chapters on item analyses, scaling and equating, and reliability (Chapters 5–7). Technical characteristics of the internal structure of the tests were presented in terms of classical item statistics (p -values and discriminations), DIF analyses, several reliability coefficients, SEMs, multidimensionality hypothesis testing and effect size estimation, and IRT analyses. In general, item difficulty indices were within acceptable and expected ranges. Chapter 6 also describes the procedures used to equate the 2009–10 test to the 2008–09 scales.

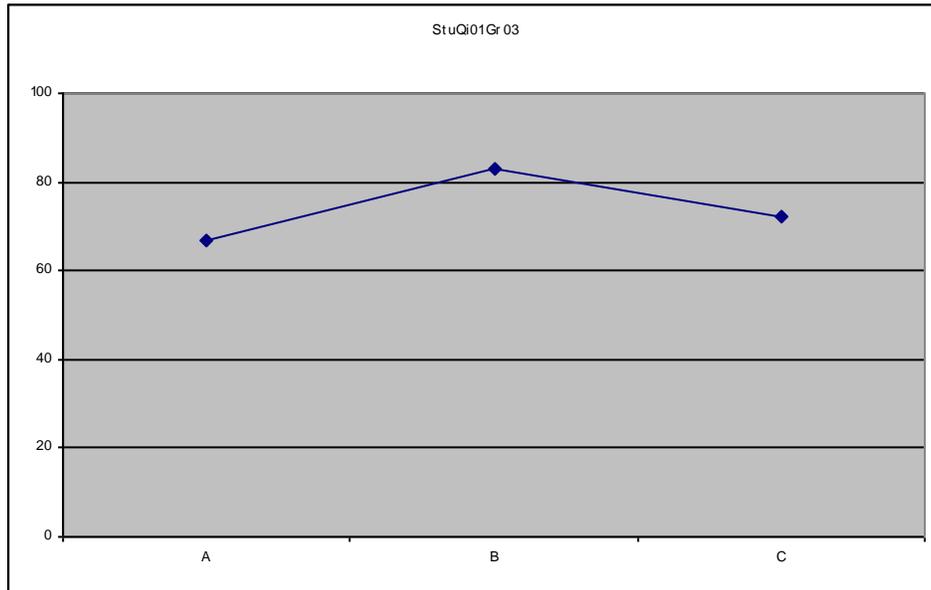
Evidence on the consequences of testing is addressed in information on scaled score and reporting in Chapters 6 and 9 and in the *Guide to Using the 2009 NECAP Reports*, which is a separate document. Each of these speaks to efforts undertaken for providing the public with accurate and clear test score information. Scaled scores simplify results reporting across content areas, grade levels, and successive years. Achievement levels give reference points for mastery at each grade level, another useful and simple way to interpret scores. Several different standard reports were provided to stakeholders. Evidence on the consequences of testing could be supplemented with broader research on the impact on student learning of NECAP testing.

9.1 Questionnaire Data

A measure of external validity was provided by comparing student performance with answers to a questionnaire administered at the end of test. The grades 3 through 8 questionnaire contained 34 questions while the grade 11 questionnaire contained 36 questions. Most of the questions were designed to gather information about students and their study habits; however, a subset could be utilized in the evaluation of external validity. The results from 16 of these questions follow. The graphs show, for each question, what percentage of the students who chose each response scored Proficient or above on the applicable content area test. For example, for reading question 1, the graph shows that approximately 67% of grade 3 students who answered A were Proficient or above in reading, approximately 83% who answered B were Proficient or above, and approximately 72% who answered C were Proficient or above.

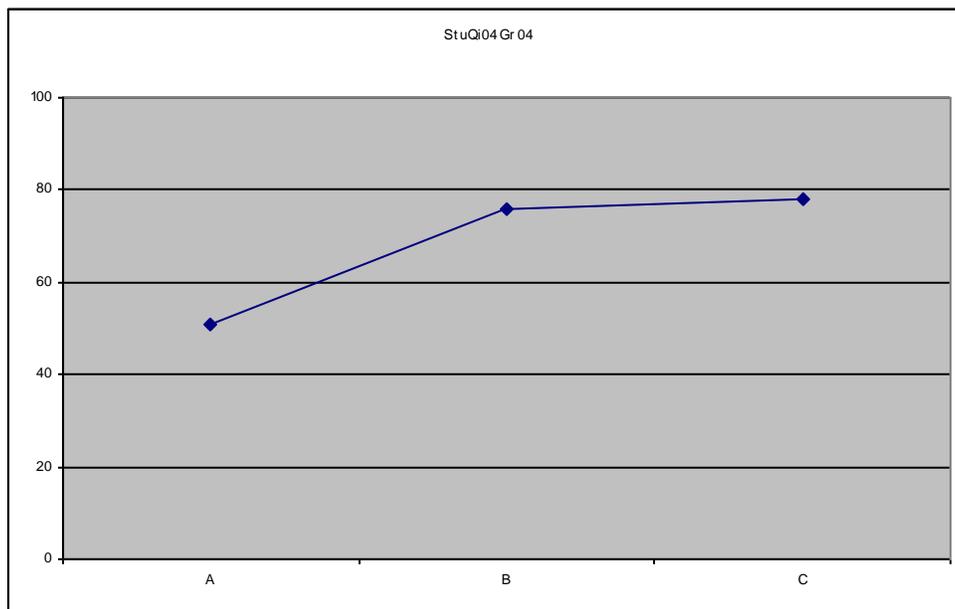
Grades 3–8 Reading Questions

Question 1: How difficult was the reading test?



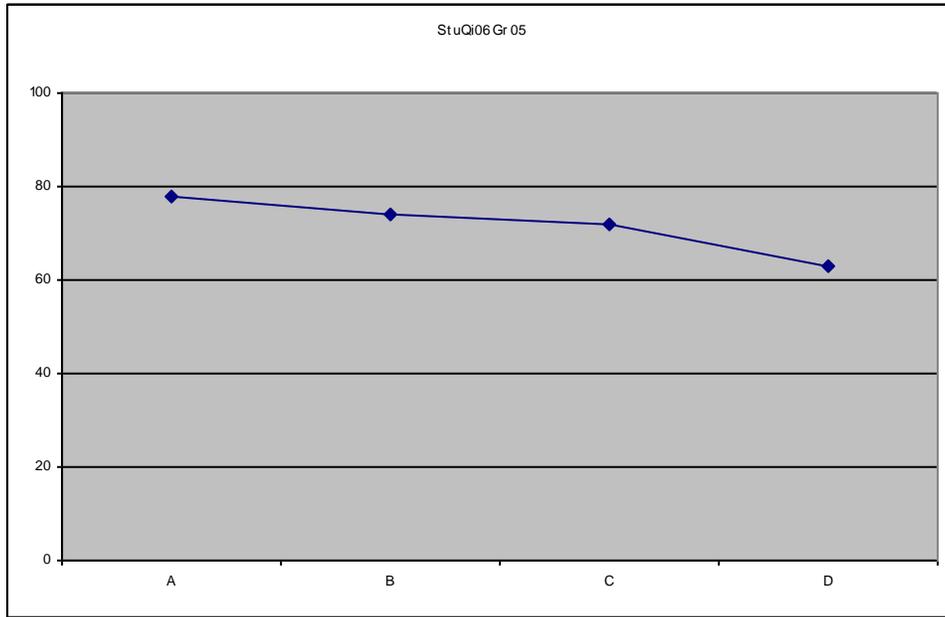
- A. harder than my regular reading school work
- B. about the same as my regular reading school work
- C. easier than my regular reading school work

Question 4: How difficult were the reading passages on the test?



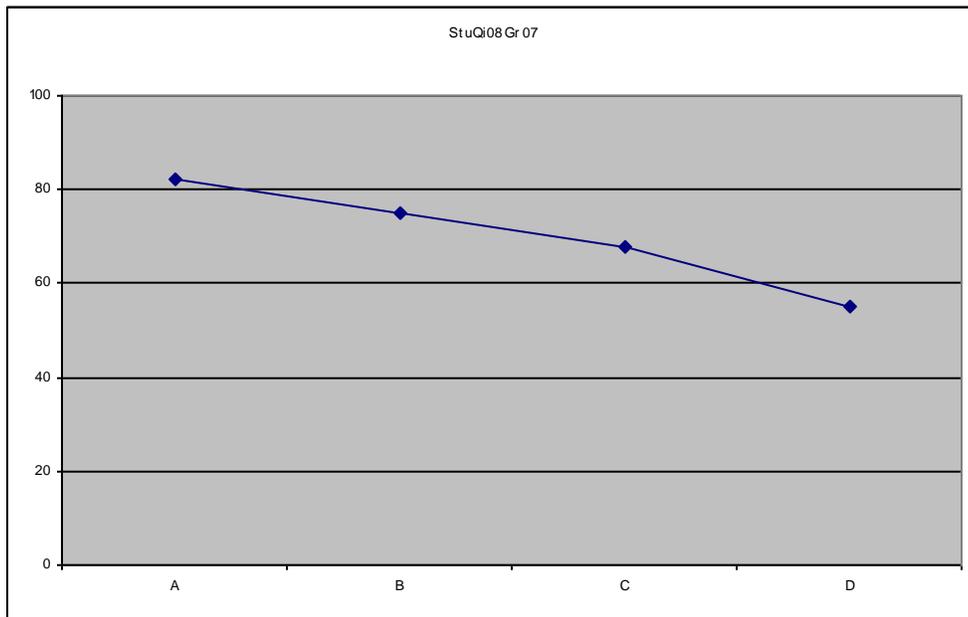
- A. Most of the passages were more difficult than what I normally read for school.
- B. Most of the passages were about the same as what I normally read for school.
- C. Most of the passages were easier than what I normally read for school.

Question 6: How often do you have language arts/reading homework?



- A. almost every day
- B. a few times a week
- C. a few times a month
- D. never or almost never

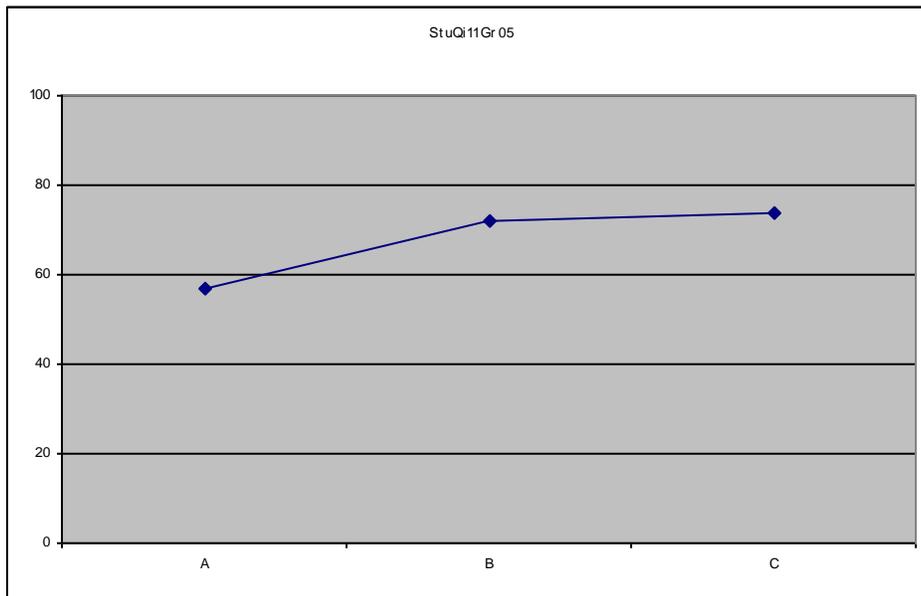
Question 8: How often do you choose to read in your free time?



- A. almost every day
- B. a few times a week
- C. a few times a month
- D. never or almost never

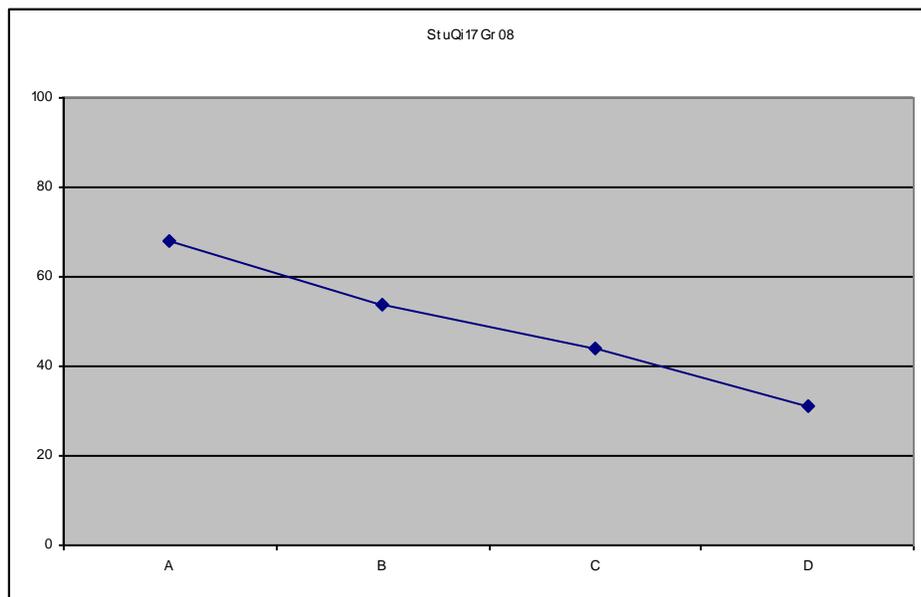
Grades 3–8 Mathematics Question

Question 11: How difficult was the mathematics test?



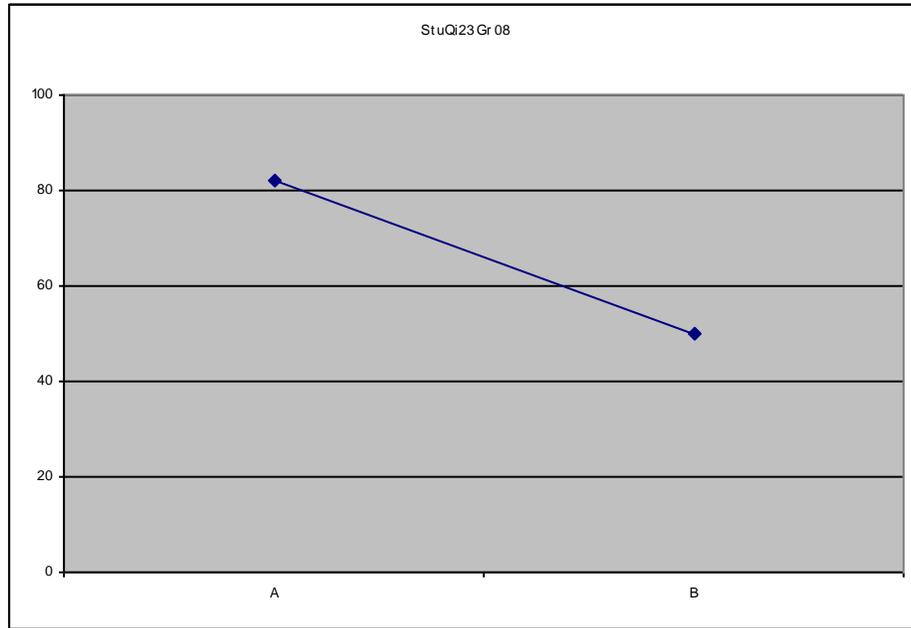
- A. harder than my regular mathematics school work
- B. about the same as my regular mathematics school work
- C. easier than my regular mathematics school work

Question 17: How often do you have mathematics homework?



- A. almost every day
- B. a few times a week
- C. a few times a month
- D. never or almost never

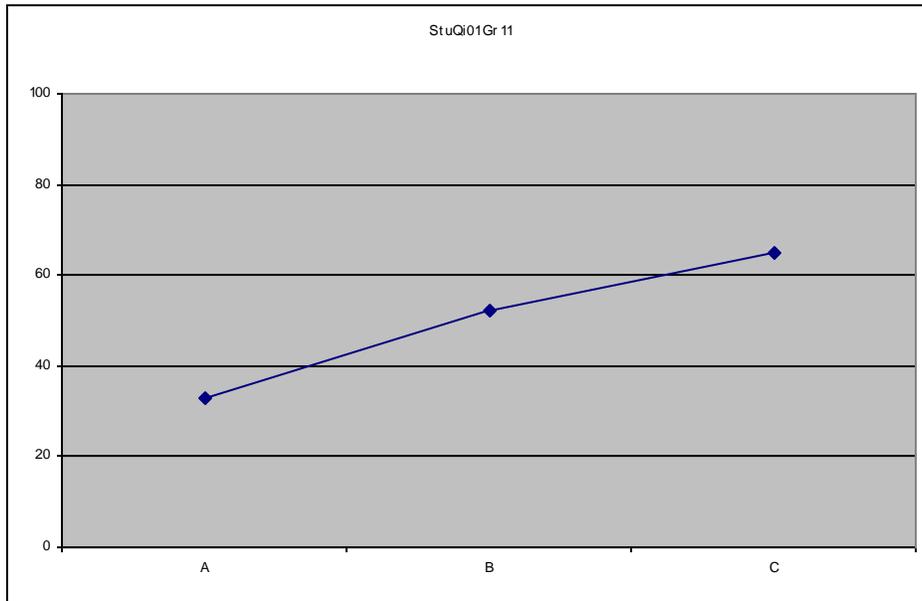
Question 23: Are you currently enrolled in an Algebra I or higher mathematics class?



- A. Yes
- B. No

Grade 11 Writing Question

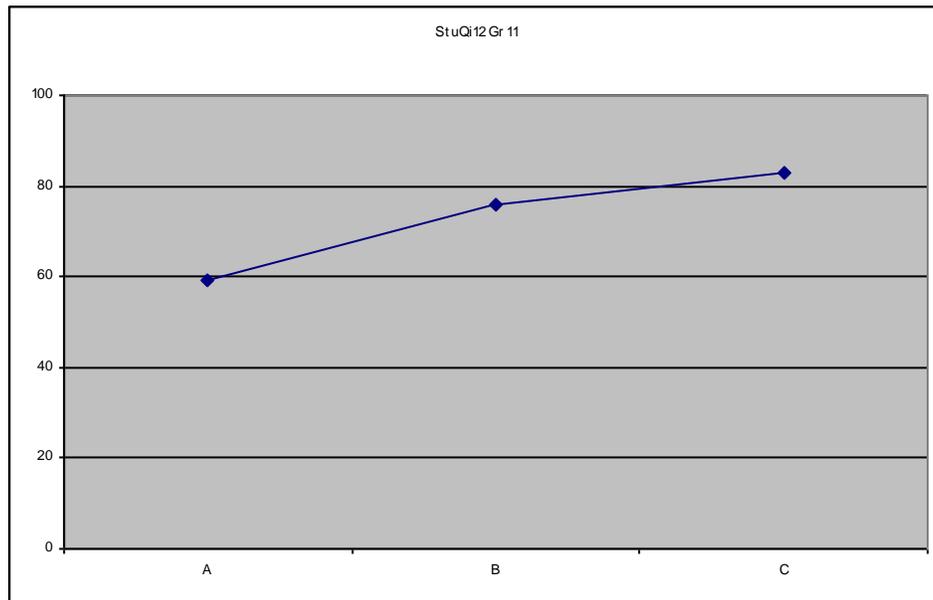
Question 1: How difficult was the writing test?



- A. harder than my regular writing work
- B. about the same as my regular writing work
- C. easier than my regular writing work

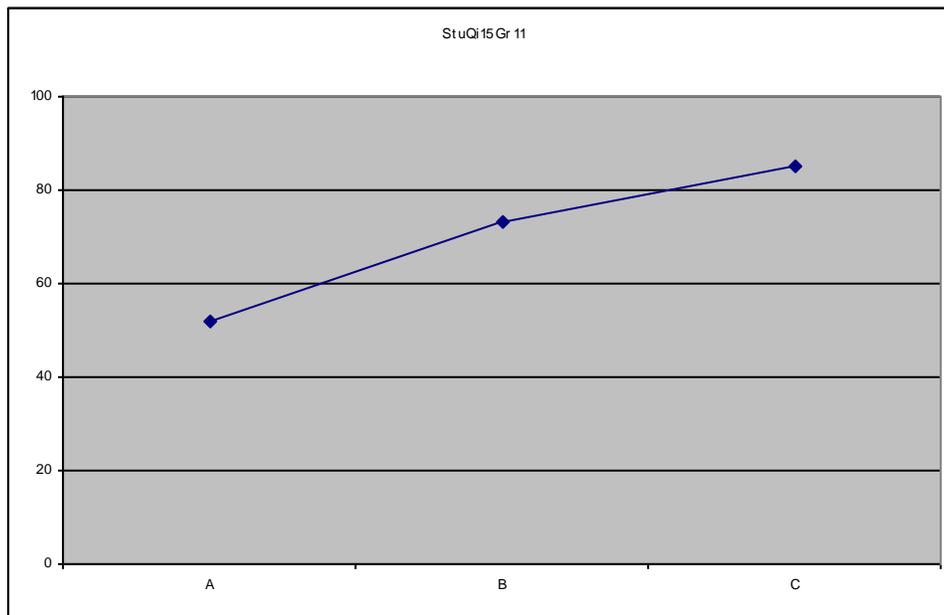
Grade 11 Reading Question

Question 12: How difficult was the reading test?



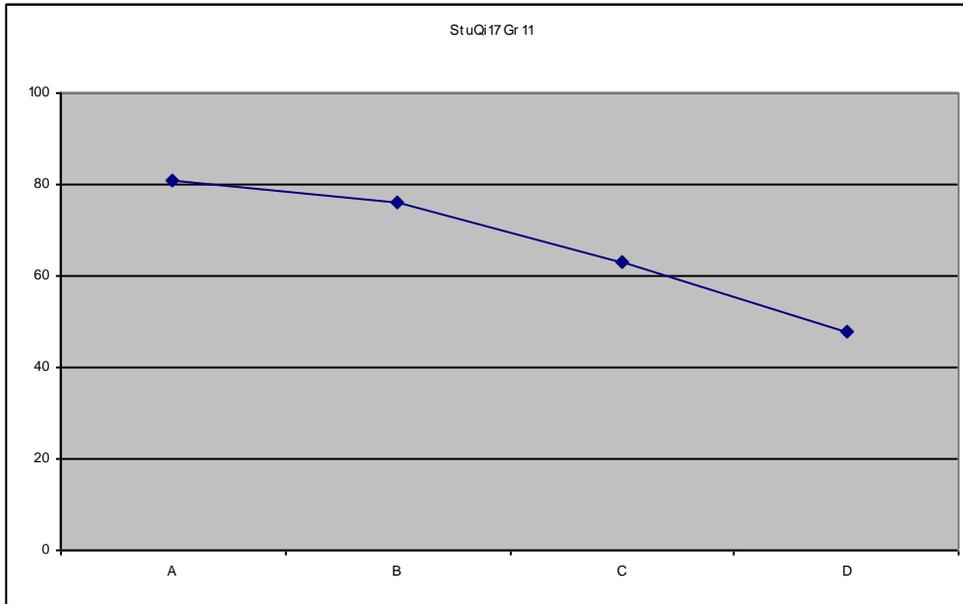
- A. harder than my regular reading work
- B. about the same as my regular reading work
- C. easier than my regular reading work

Question 15: How difficult were the reading passages on the test?



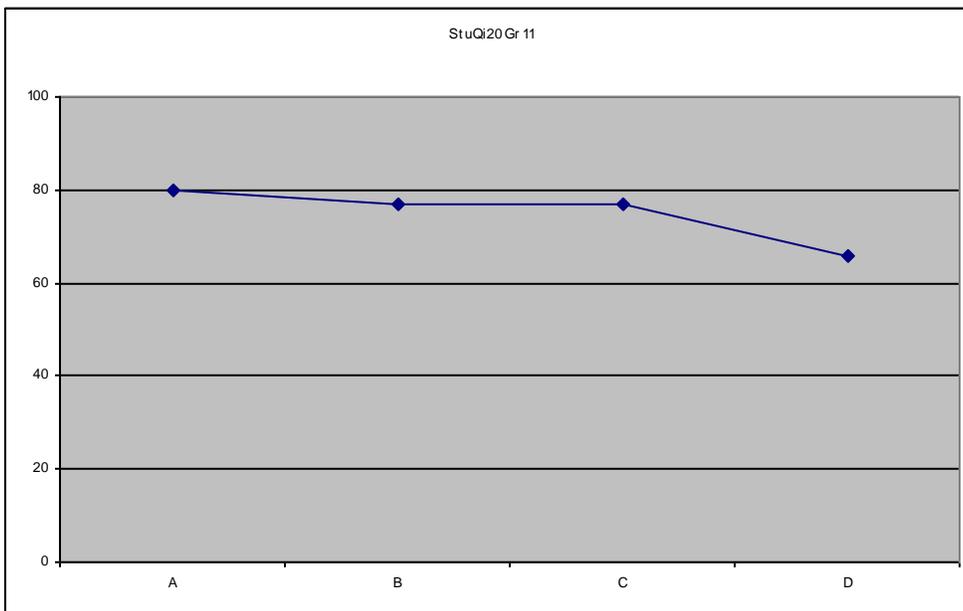
- A. Most of the passages were more difficult than what I normally read for school.
- B. Most of the passages were about the same as what I normally read for school.
- C. Most of the passages were easier than what I normally read for school.

Question 17: How often do you have reading homework in English class?



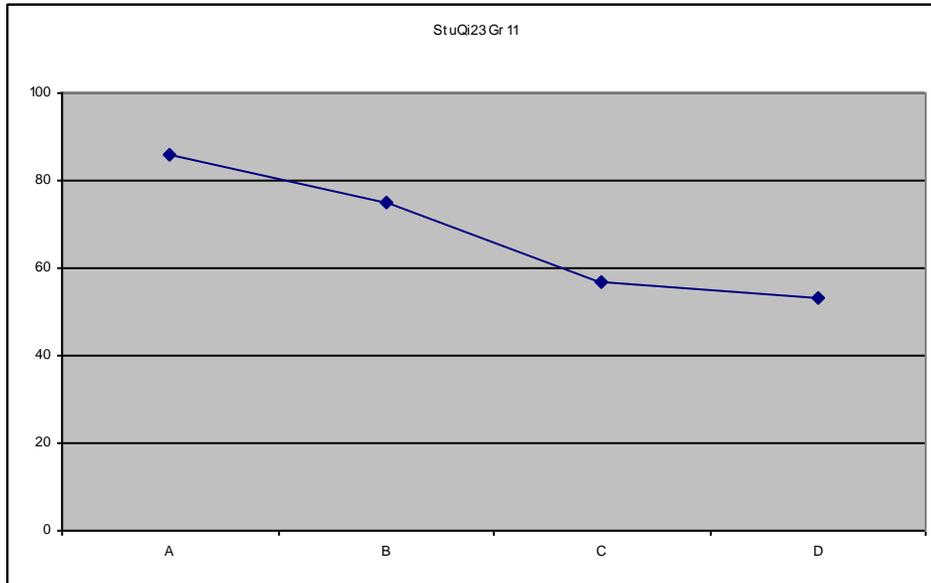
- A. almost every day
- B. a few times a week
- C. a few times a month
- D. I usually don't have reading homework in English class.

Question 20: How often do you choose to read in your free time?



- A. almost every day
- B. a few times a week
- C. a few times a month
- D. I almost never read.

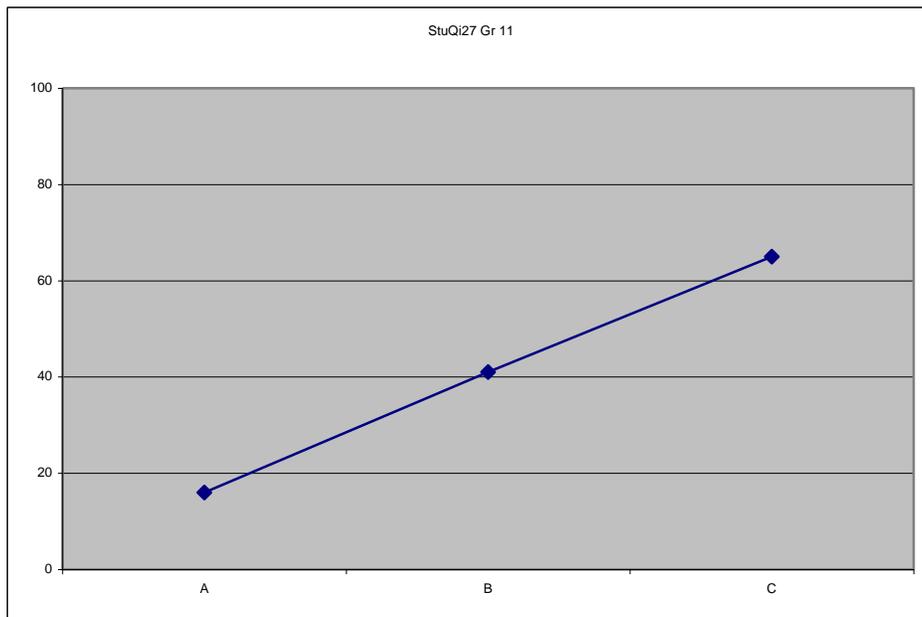
Question 23: What grade did you receive in the last English course you completed?



- A. A
- B. B
- C. C
- D. lower than C

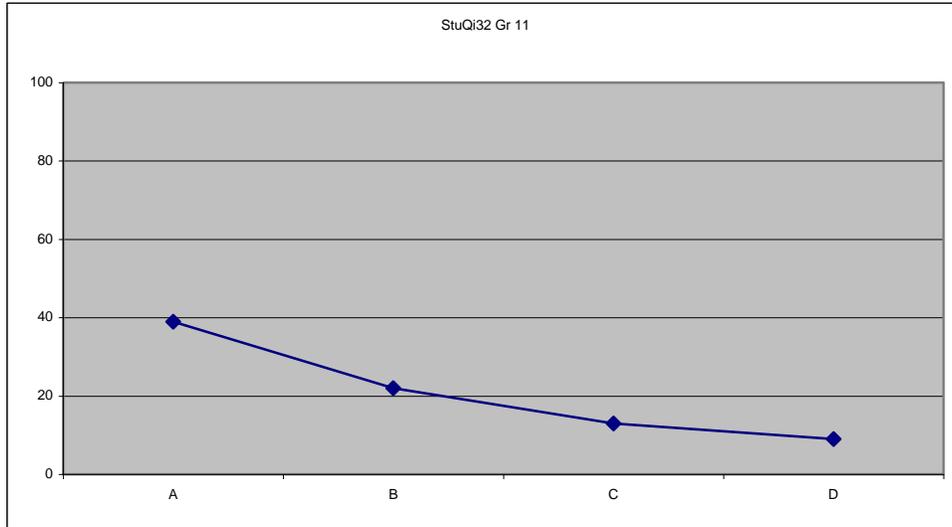
Grade 11 Mathematics Question

Question 27: How difficult was the mathematics test compared to your current or most recent mathematics class?



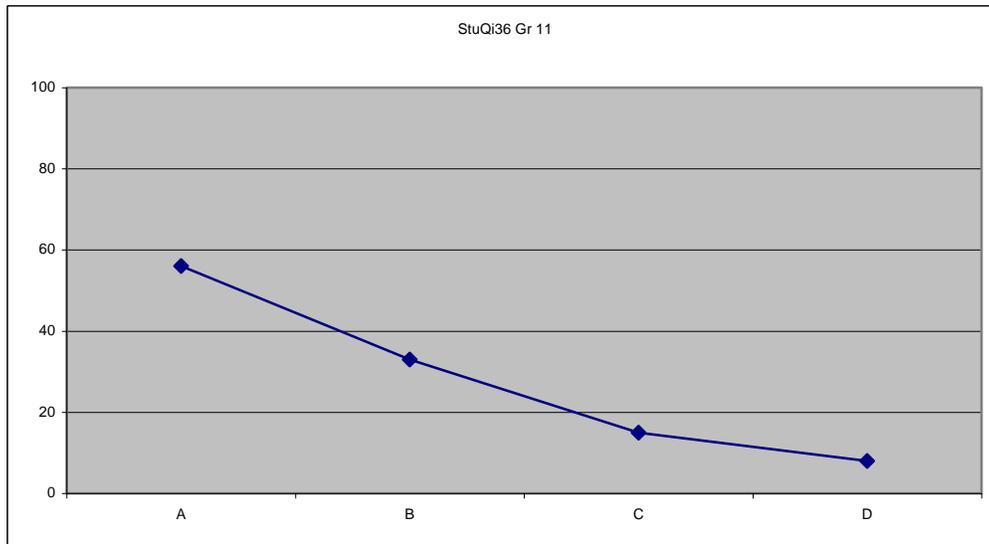
- A. more difficult
- B. about the same
- C. less difficult

Question 32: How often do you have mathematics homework assignments?



- A. almost every day
- B. a few times a week
- C. a few times a month
- D. I usually don't have homework in mathematics.

Question 36: What grade did you receive in the last mathematics course you completed?



- A. A
- B. B
- C. C
- D. lower than C

In virtually all of the graphs presented above, the relationship between the questionnaire data and performance on the NECAP was consistent with expectations; this provides evidence of external validity. See

Appendix Q for a copy of the questionnaire and complete data comparing questionnaire items and test performance.

9.2 Validity Studies Agenda

The remaining part of this chapter describes further studies of validity that could enhance the investigations of validity that have already been performed. The proposed areas of validity to be examined fall into four categories: *external validity*, *convergent and discriminant validity*, *structural validity*, and *procedural validity*. These will be discussed in turn.

9.2.1 External Validity

In the future, investigations of external validity could involve targeted examination of variables which one might expect to correlate with NECAP results, like classroom grades or classroom test scores in the same content areas as the NECAP test in question.

Further evidence of external validity might come from correlating NECAP scores with scores on another standardized test, such as the Iowa Test of Basic Skills (ITBS). As with the study of concordance between NECAP scores and grades, this investigation would compare scores in analogous content areas (e.g., NECAP reading and ITBS reading comprehension). All tests taken by each student would be appropriate to the student's grade level.

9.2.2 Convergent and Discriminant Validity

The concepts of convergent and discriminant validity were defined by Campbell and Fiske (1959) as specific types of validity that fall under the umbrella of *construct validity*. The notion of convergent validity states that measures or variables that are intended to align with one another should actually be aligned in practice. Discriminant validity, on the other hand, is the idea that measures or variables that are intended to differ from one another should not be too highly correlated. Evidence for validity comes from examining whether the correlations among variables are as expected in direction and magnitude.

Campbell and Fiske (1959) introduced the study of different *traits* and *methods* as the means of assessing convergent and discriminant validity. Traits refer to the constructs that are being measured (e.g., mathematical ability), and methods are the instruments of measuring them (e.g., a mathematics test or grade). To utilize the framework of Campbell and Fiske, it is necessary that more than one trait and more than one method be examined. Analysis is performed through the multi-trait/multi-method matrix, which gives all possible correlations of the different combinations of traits and methods. Campbell and Fiske defined four properties of the multi-trait/multi-method matrix that serve as evidence of convergent and discriminant validity:

- The correlation among different methods of measuring the same trait should be sufficiently different from zero. For example, scores on a mathematics test and grades in a mathematics class should be positively correlated.
- The correlation among different methods of measuring the same trait should be higher than that of different methods of measuring different traits. For example, scores on a mathematics test and grades in a mathematics class should be more highly correlated than scores on a mathematics test and grades in a reading class.
- The correlation among different methods of measuring the same trait should be higher than the same method of measuring different traits. For example, scores on a mathematics test and grades in a mathematics class should be more highly correlated than scores on a mathematics test and scores on an analogous reading test.
- The pattern of correlations should be similar across comparisons of different traits and methods. For example, if the correlation between test scores in reading and writing is higher than the correlation between test scores in reading and mathematics, it is expected that the correlation between grades in reading and writing would also be higher than the correlation between grades in reading and mathematics.

For NECAP, convergent and discriminant validity could be examined by constructing a multi-trait/multi-method matrix and analyzing the four pieces of evidence described above. The traits examined would be mathematics, reading, and writing; different methods could include NECAP score and such variables as grades, teacher judgments, scores on another standardized test, etc.

9.2.3 Structural Validity

Though the previous types of validity examine the concurrence between different measures of the same content area, structural validity focuses on the relation between strands *within* a content area, thus supporting *content validity*. Standardized tests are carefully designed to ensure that all appropriate strands of a content area are adequately covered in a test, and structural validity is the degree to which related elements of a test are correlated in the intended manner. For instance, it is desired that performance on different strands of a content area be positively correlated; however, as these strands are designed to measure distinct components of the content area, it is reasonable to expect that each strand would contribute a unique component to the test. Additionally, it is desired that the correlation between different item types (multiple-choice, short-answer, and constructed-response) of the same content area be positive.

As an example, an analysis of NECAP structural validity would investigate the correlation between performance in Geometry and Measurement and performance in Functions and Algebra. Additionally, the concordance between performance on multiple-choice items and constructed-response items would be examined. Such a study would address the consistency of NECAP tests within each grade and content area. In

particular, the dimensionality analyses of Chapter 5 could be expanded to include confirmatory analyses addressing these concerns.

9.2.4 Procedural Validity

As mentioned earlier, the *NECAP Principal/Test Coordinator Manual* and *Test Administrator Manual* delineated the procedures to which all NECAP test coordinators and test administrators were required to adhere. A study of procedural validity would provide a comprehensive documentation of the procedures that were followed throughout the NECAP administration. The results of the documentation would then be compared to the manuals, and procedural validity would be confirmed to the extent that the two were in alignment. Evidence of procedural validity is important because it verifies that the actual administration practices were in accord with the intentions of the design.

Possible instances where discrepancies can exist between design and implementation include the following: a teacher spirals test forms incorrectly within a classroom; cheating among students occurs; answer documents are scanned incorrectly. These are examples of *administration error*. A study of procedural validity involves capturing any administration errors and presenting them within a cohesive document for review.

All potential tests of validity that have been introduced in this chapter should be considered by the NECAP Technical Advisory Committee (NECAP TAC) during 20010–11. With the advice of the NECAP TAC (see Appendix A for list of members), the states will develop short term and longer term (e.g., 2 year to 5 year) plans for validity studies.

REFERENCES

- Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Belmont, CA: Wadsworth, Inc.
- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Marcel Dekker, Inc.
- Brown, F. G. (1983). *Principles of educational and psychological testing* (3rd ed.). Fort Worth: Holt, Rinehart and Winston.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81–105.
- 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, NJ: Lawrence Erlbaum Associates.
- Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, *23*, 355–368.
- Draper, N. R. & Smith, H. (1998). *Applied regression analysis* (3rd ed.). New York: John Wiley and Sons, Inc.
- Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105-146). New York: Macmillan Publishing Company.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer Academic Publishers.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.
- Hambleton, R. K., & van der Linden, W. J. (1997). *Handbook of modern item response theory*. New York, NY: Springer-Verlag.
- Joint Committee on Testing Practices (2004). *Code of fair testing practices in education*. Washington, DC: Joint Committee on Testing Practices. Available from www.apa.org/science/programs/testing/fair-code.aspx
- Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, *32*, 179-197.

- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Muraki, E. & R. D. Bock (2003). PARSCALE 4.1. Lincolnwood, IL: Scientific Software International.
- 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