New York State Testing Program 2010: English Language Arts, Grades 3–8

Technical Report

Submitted
2010

CTB/McGraw-Hill
Monterey, California 93940

Copyright

Table of Contents

List of Tables

Section I: Introduction and Overview

Introduction

An overview of the New York State Testing Program (NYSTP), Grades 3–8, English Language Arts (ELA) 2010 Operational (OP) Tests is provided in this report. The report contains information about OP test development and content, item and test statistics, validity and reliability, differential item functioning studies, test administration and scoring, scaling, and student performance.

Test Purpose

The NYSTP is an assessment system designed to measure concepts, processes, and skills taught in schools in New York. The ELA Tests target student progress toward three of the four content standards as described in Section II, "Test Design and Development," subsection "Content Rationale." The Grades 3–8 ELA Tests are written for all students to have the opportunity to demonstrate their knowledge and skills in these standards. The established cut scores classify student proficiency into one of four levels based on their test performance.

Target Population

Students in New York State public school Grades 3, 4, 5, 6, 7, and 8 (and ungraded students of equivalent age) are the target population for the Grades 3–8 testing program. Nonpublic schools may participate in the testing program, but the participation is not mandatory for them. In 2010, nonpublic schools participated in all grade tests but were not well represented in the testing program. The New York State Education Department (NYSED) made a decision to exclude these schools from the data analyses. Public school students were required to take all State assessments administered at their grade level, except for a very small percentage of students with disabilities who took the New York State Alternate Assessment (NYSAA) for students with severe disabilities. For more detail on this exemption, please refer to the *School Administrator's Manual*, available online at http://www.p12.nysed.gov/osa/manuals/.

Test Use and Decisions Based on Assessment

The Grades 3–8 ELA Tests are used to measure the extent to which individual students achieve the New York State Learning Standards in ELA and to determine whether schools, districts, and the State meet the required progress targets specified in the New York State accountability system. There are several types of scores available from the Grades 3–8 ELA Tests and these are discussed in this section.

Scale Scores

The scale score is a quantification of the ability measured by the Grades 3–8 ELA Tests at each grade level. The scale scores are comparable within each grade level but not across grades because the Grades 3–8 ELA Tests are not on a vertical scale. The test scores are reported at the individual level and can also be aggregated. Detailed information on the derivation and properties of scale scores is provided in Section VI, "IRT Scaling and Equating." The Grades 3–8 ELA Tests scores are used to determine

student progress within schools and districts, support registration of schools and districts, determine eligibility of students for additional instruction time, and provide teachers with indicators of a student's need, or lack of need, for remediation in specific content-area knowledge.

Proficiency Level Cut Scores and Classification

Students are classified as Level I (Below Standards), Level II (Meets Basic Standards), Level III (Meets Proficiency Standards), and Level IV (Exceeds Proficiency Standards). The original proficiency cut scores used to distinguish among Levels I, II, III, and IV were established during the process of Standard Setting in 2006. In 2010 a change in the test administration window between the 2008–2009 and 2009–2010 school years and a decision to align the proficiency standards with Grade 8 student performance on the NYS Regents ELA exams led to changes in the proficiency cut scores. The process of cut score adjustment after the 2010 OP test administration is described in detail in Section VII "Proficiency Level Cut Score Adjustment" of this report.

Detailed information on a process of establishing original performance cut scores and their association with test content is provided in the *Bookmark Standard Setting Technical Report 2006 for Grades 3, 4, 5, 6, 7, and 8 English Language Arts* and the *New York State ELA Measurement Review Technical Report 2006 for English Language Arts*.

Standard Performance Index Scores

Standard performance index (SPI) scores are obtained from the Grades 3–8 ELA Tests. The SPI score is an indicator of student ability and knowledge and skills in specific learning standards, and it is used primarily for diagnostic purposes to help teachers evaluate academic strengths and weaknesses of their students. These scores can be effectively used by teachers at the classroom level to modify their instructional content and format to best serve their students' specific needs. Detailed information on the properties and use of SPI scores are provided in Section VI, "IRT Scaling and Equating."

Testing Accommodations

In accordance with federal law under the Americans with Disabilities Act and Fairness in Testing, as outlined by the *Standards for Educational and Psychological Testing* (American Education Research Association, American Psychological Association, and National Council on Measurement in Education, 1999), accommodations that do not alter the measurement of any construct being tested are allowed for test takers. The allowance is in accordance with a student's individual education program (IEP) or section 504 Accommodation Plan (504 Plan). School principals are responsible for ensuring that proper accommodations are provided when necessary and that staff providing accommodations are properly trained. Details on testing accommodations can be found in the *School Administrator's Manual*.

Test Transcriptions

For the visually impaired students, large type and braille editions of the test books are provided. The students dictate and/or record their responses; the teachers transcribe student responses to multiple-choice (MC) questions onto scannable answer sheets; and the teachers transcribe the responses to the constructed-response (CR) questions onto the

regular test books. The large type editions are created by CTB/McGraw-Hill and printed by NYSED, and the braille editions are produced by Braille Publishers, Inc. The lead transcribers are members of the National Braille Association, California Transcribers and Educators of the Visually Handicapped, and the Contra Costa Braille Transcribers, and they have Library of Congress and Nemeth Code [Braille] Certifications. Braille Publishers, Inc. produced the braille editions for the previous Grades 4 and 8 Tests.

Camera-copy versions of the regular test books are provided to the braille vendor, who then produces the braille editions. Proofs of the braille editions are submitted to NYSED for review and approval prior to production.

Test Translations

Since these are assessments of proficiency in English language arts, the Grades 3–8 ELA Tests are not translated into any other language.

Section II: Test Design and Development

Test Description

The Grades 3–8 ELA Tests are New York State Learning Standards-based criterion-referenced tests composed of multiple-choice (MC) and constructed-response (CR) items. The tests were administered in New York classrooms during April 2010 over a two-day (Grades 3, 5, 7, and 8) or three-day (Grades 4 and 6) period. The tests were printed in black and white and incorporated the concepts of universal design. Copies of the OP tests are available online at http://www.nysedregents.org/elementary.html and http://www.nysedregents.org/intermediate.html. Details on the administration and scoring of these tests can be found in Section IV, "Test Administration and Scoring."

Test Configuration

The OP test books were administered, in order, on two to three consecutive days, depending on the grade. Table 1 provides information on the number and type of items in each book, as well as testing times. Students were administered a Reading section (Book 1, all grades; Book 3, Grades 4, 6, and 8) and a Listening section (Book 2). Students in Grades 3, 5, and 7 also completed an Editing Paragraph (in Book 2). The 2010 *Teacher's Directions* available online (http://www.p12.nysed.gov/osa/ei/directions/ela3-5-td-10.pdf and http://www.p12.nysed.gov/osa/ei/directions/ela6-8-td-10.pdf) as well as the 2010 *School Administrator's Manual* (http://www.p12.nysed.gov/osa/sam/ela/elaei-sam-10.pdf) provide details on security, scheduling, classroom organization and preparation, test materials, and administration.

Table 1. NYSTP ELA 2010 Test Configuration

Grade	Day	Book	Number of Items			Allotted Time (minutes)	
			MC	CR*	Total**	Testing	Prep
3	1	1	20	1	21	40	10
	2	2	4	3	7	35	15
	Totals		24	4	28	75	25
4	1	1	28	0	28	45	10
	2	2	0	3	3	45	15
	3	3	0	4	4	60	10
	Totals		28	7	35	150	35
5	1	1	20	1	21	45	10
	2	2	4	2	6	30	15
	Totals		24	3	27	75	25
6	1	1	26	0	26	55	10
	2	2	0	4	4	45	15
	3	3	0	4	4	60	10
	Totals		26	8	34	160	35

(Continued on next page)

Table 1. NYSTP ELA 2010 Test Configuration (cont.)

Grade	Day	Book	Number of Items			Allotted Time (minutes)	
			MC	CR*	Total**	Testing	Prep
7	1	1	26	2	28	60	10
	2	2	4	3	7	30	15
	Totals		30	5	35	90	25
8	1	1	26	0	26	55	10
	1	2	0	4	4	45	15
	2	3	0	4	4	60	10
	Totals		26	8	34	160	35

*Does not reflect cluster-scoring. **Reflects actual items in the test books.

In most cases, the test book item number is also the item number for the purposes of data analysis. The exception is that CR items from Grades 4, 6, and 8 are cluster-scored. Table 2 lists the test book item numbers and the item numbers as scored. Because analyses are based on scored data, the latter item numbers will be referred to in this *Technical Report*.

Table 2. NYSTP ELA 2010 Cluster Items

Grade	Cluster Type	Contributing Book Items	Item Number for Data Analysis
4	Listening	29, 30, 31	29
4	Reading	32, 33, 34, 35	30
4	Writing Mechanics	31, 35	31
6	Listening	27, 28, 29, 30	27
6	Reading	31, 32, 33, 34	28
6	Writing Mechanics	30, 34	29
8	Listening	27, 28, 29, 30	27
8	Reading	31, 32, 33, 34	28
8	Writing Mechanics	30, 34	29

Test Blueprint

The NYSTP Grades 3–8 ELA Tests assess students on three learning standards (S1—Information and Understanding, S2—Literary Response and Expression, and S3—Critical Analysis and Evaluation). The test items are indicators used to assess a variety of reading, writing, and listening skills against each of the three Learning Standards. Standard 1 is assessed primarily by use of test items associated with informational passages; Standard 2 is assessed primarily by use of test items associated with literary passages; and Standard 3 is assessed by use of test items associated with a combination of genres. In addition, students are also tested on writing mechanics, which is assessed independent of alignment to the Learning Standards, since writing mechanics is associated with all three Learning Standards. The distribution of score points across the Learning Standards was determined during blueprint specifications meetings held with panels of New York State educators at the start of the testing program, prior to item development. The distribution in each grade reflects the number of assessable performance indicators in each Learning Standard at that grade and the emphasis placed

on those performance indicators by the blueprint-specifications panel members. Table 3 shows the Grades 3–8 ELA Tests blueprint and actual number of score points in 2010 OP tests.

Table 3. NYSTP ELA 2010 Test Blueprint

Grade	Total Points	Writing Mechanics Points	Standard	Target Reading and Listening Points	Selected Reading and Listening Points	Target % of Test (Excluding Writing)	Selected % of Test (Excluding Writing)
3	33	3	S1	10	10	33.0	33.0
			S2	14	15	47.0	50.0
			S3	6	5	20.0	17.0
4	39	3	S1	13	12	36.0	33.0
			S2	16	17	44.5	47.0
			S3	7	7	19.5	20.0
5	31	3	S1	12	13	43.0	46.0
			S2	10	8	36.0	29.0
			S3	6	7	21.0	25.0
6	39	3	S1	13	14	36.0	39.0
			S2	16	15	44.5	42.0
			S3	7	7	19.5	19.0
7	41	3	S1	15	16	39.0	42.0
			S2	15	13	39.0	34.0
			S3	8	9	22.0	24.0
8	39	3	S1	14	14	39.0	39.0
			S2	14	13	39.0	36.0
			S3	8	9	22.0	25.0

Tables 4a–4f present Grades 3–8 ELA Test item maps with the item type indicator, the maximum number of points obtainable from each item, the Learning Standard measured by each item, and the answer key.

Table 4a. NYSTP ELA 2010 Operational Test Map, Grade 3

Question	Type	Points	Standard	Performance Indicator	Answer Key
Book 1	**Reading**				
1	Multiple Choice	1	2	Summarize main ideas and supporting details from imaginative texts	D
2	Multiple Choice	1	2	Use knowledge of story structure, story elements, and key vocabulary to interpret stories	B
3	Multiple Choice	1	2	Make predictions, draw conclusions, and make inferences about events and characters	A
4	Multiple Choice	1	2	Use specific evidence from stories to describe characters, their actions, and their motivations; relate sequences of events	B
5	Multiple Choice	1	3	Evaluate the content by identifying whether events, actions, characters, and/or settings are realistic	C
6	Multiple Choice	1	1	Identify main ideas and supporting details in informational texts	A
7	Multiple Choice	1	1	Read unfamiliar texts to collect data, facts, and ideas	D
8	Multiple Choice	1	1	Identify main ideas and supporting details in informational texts	D
9	Multiple Choice	1	1	Identify main ideas and supporting details in informational texts	C
10	Multiple Choice	1	3	Evaluate the content by identifying the author's purpose	C
11	Multiple Choice	1	2	Use knowledge of story structure, story elements, and key vocabulary to interpret stories	C
12	Multiple Choice	1	2	Use specific evidence from stories to describe characters, their actions, and their motivations; relate sequences of events	A
13	Multiple Choice	1	2	Make predictions, draw conclusions, and make inferences about events and characters	C
14	Multiple Choice	1	2	Make predictions, draw conclusions, and make inferences about events and characters	A
15	Multiple Choice	1	3	Evaluate the content by identifying the author's purpose	D
16	Multiple Choice	1	1	Locate information in a text that is needed to solve a problem	A
17	Multiple Choice	1	1	Locate information in a text that is needed to solve a problem	C
18	Multiple Choice	1	1	Read unfamiliar texts to collect data, facts, and ideas	A
19	Multiple Choice	1	3	Evaluate the content by identifying important and unimportant details	B
20	Multiple Choice	1	1	Identify a conclusion that summarizes the main idea	B
21	Short Response	2	1	Read unfamiliar texts to collect data, facts, and ideas	n/a

(Continued on next page)

Table 4a. NYSTP ELA 2010 Operational Test Map, Grade 3 (cont.)

Question	Type	Points	Standard	Performance Indicator	Answer Key
Book 2	**Listening and Writing**				
22	Multiple Choice	1	2	Identify elements of character, **plot**, and setting to understand the author's message or intent	C
23	Multiple Choice	1	2	Identify elements of **character**, plot, and setting to understand the author's message or intent	B
24	Multiple Choice	1	2	Identify elements of character, **plot**, and setting to understand the author's message or intent	A
25	Multiple Choice	1	3	Distinguish between fact and opinion	B
26	Short Response	2	2	Use note taking and graphic organizers to record and organize information and ideas recalled from stories read aloud	n/a
27	Short Response	2	2	Identify elements of **character**, plot, and setting to understand the author's message or intent	n/a
28	Editing Paragraph	3	n/a	Use basic punctuation correctly Capitalize words such as literary titles, holidays, and product names	n/a

Table 4b. NYSTP ELA 2010 Operational Test Map, Grade 4

Question	Type	Points	Standard	Performance Indicator	Answer Key
Book 1	**Reading**				
1	Multiple Choice	1	2	Use specific evidence from stories to identify themes; describe characters, their actions, and their motivations; relate a sequence of events	A
2	Multiple Choice	1	2	Use graphic organizers to record significant details about characters and events in stories	C
3	Multiple Choice	1	2	Use knowledge of story structure, story elements, and key vocabulary to interpret stories	A
4	Multiple Choice	1	3	Evaluate the content by identifying important and unimportant details	B
5	Multiple Choice	1	3	Evaluate the content by identifying the author's purpose	C
6	Multiple Choice	1	2	Determine the meaning of unfamiliar words by using context clues, dictionaries, and other classroom resources	A
7	Multiple Choice	1	1	Identify a main idea and supporting details in informational texts	A
8	Multiple Choice	1	1	Collect and interpret data, facts, and ideas from unfamiliar texts	B
9	Multiple Choice	1	1	Collect and interpret data, facts, and ideas from unfamiliar texts	D
10	Multiple Choice	1	1	Locate information in a text that is needed to solve a problem	D

(Continued on next page)

Table 4b. NYSTP ELA 2010 Operational Test Map, Grade 4 (cont.)

Question	Type	Points	Standard	Performance Indicator	Answer Key
Book 1	**Reading**				
11	Multiple Choice	1	1	Recognize and use organizational features, such as table of contents, indexes, page numbers, and chapter headings/subheadings, to locate information	B
12	Multiple Choice	1	1	Identify a conclusion that summarizes the main idea	A
13	Multiple Choice	1	2	Use specific evidence from stories to identify themes; describe characters, their actions, and their motivations; relate a sequence of events	B
14	Multiple Choice	1	2	Use specific evidence from stories to identify themes; describe characters, their actions, and their motivations; relate a sequence of events	D
15	Multiple Choice	1	2	Use specific evidence from stories to identify themes; describe characters, their actions, and their motivations; relate a sequence of events	A
16	Multiple Choice	1	2	Make predictions, draw conclusions, and make inferences about events and characters	D
17	Multiple Choice	1	2	Make predictions, draw conclusions, and make inferences about events and characters	A
18	Multiple Choice	1	2	Use specific evidence from stories to identify themes; describe characters, their actions, and their motivations; relate a sequence of events	C
19	Multiple Choice	1	2	Make predictions, draw conclusions, and make inferences about events and characters	A
20	Multiple Choice	1	2	Use knowledge of story structure, story elements, and key vocabulary to interpret stories	B
21	Multiple Choice	1	2	Make predictions, draw conclusions, and make inferences about events and characters	B
22	Multiple Choice	1	3	Evaluate the content by identifying important and unimportant details	D
23	Multiple Choice	1	1	Identify a main idea and supporting details in informational texts	D
24	Multiple Choice	1	1	Understand written directions and procedures	C
25	Multiple Choice	1	1	Locate information in a text that is needed to solve a problem	A
26	Multiple Choice	1	1	Understand written directions and procedures	A
27	Multiple Choice	1	1	Recognize and use organizational features, such as table of contents, indexes, page numbers, and chapter headings/subheadings, to locate information	D
28	Multiple Choice	1	1	Identify a conclusion that summarizes the main idea	B

(Continued on next page)

Table 4b. NYSTP ELA 2010 Operational Test Map, Grade 4 (cont.)

Question	Type	Points	Standard	Performance Indicator	Answer Key
Book 2	**Listening and Writing**				
29–31	Short and Extended Response	4	2	Listening/Writing cluster	n/a
Book 3	**Reading and Writing**				
32–35	Short and Extended Response	4	3	Reading/Writing cluster	n/a
Book 2 & Book 3	**Writing Mechanics**				
31 & 35	Extended Response	3	n/a	Writing Mechanics cluster	n/a

Table 4c. NYSTP ELA 2010 Operational Test Map, Grade 5

Question	Type	Points	Standard	Performance Indicator	Answer Key
Book 1	**Reading**				
1	Multiple Choice	1	1	Read to collect and interpret data, facts, and ideas from multiple sources	C
2	Multiple Choice	1	1	Use text features, such as headings, captions, and titles, to understand and interpret informational texts	B
3	Multiple Choice	1	1	Recognize organizational formats to assist in comprehension of informational texts	B
4	Multiple Choice	1	1	Determine the meaning of unfamiliar words by using context clues, a dictionary, or a glossary	A
5	Multiple Choice	1	1	Read to collect and interpret data, facts, and ideas from multiple sources	D
6	Multiple Choice	1	3	Evaluate information, ideas, opinions, and themes in texts by identifying a central idea and supporting details	B
7	Multiple Choice	1	2	Identify literary elements, such as **setting**, plot, and character, of different genres	D
8	Multiple Choice	1	2	Define characteristics of different genres	C
9	Multiple Choice	1	2	Read, view, and interpret literary texts from a variety of genres	C
10	Multiple Choice	1	2	Recognize how the author uses literary devices, such as simile, metaphor, and personification, to create meaning	B
11	Multiple Choice	1	1	Recognize organizational formats to assist in comprehension of informational texts	A

(Continued on next page)

Table 4c. NYSTP ELA 2010 Operational Test Map, Grade 5 (cont.)

Question	Type	Points	Standard	Performance Indicator	Answer Key
Book 1	**Reading**				
12	Multiple Choice	1	1	Read to collect and interpret data, facts, and ideas from multiple sources	A
13	Multiple Choice	1	3	Evaluate information, ideas, opinions, and themes in texts by identifying a central idea and supporting details	D
14	Multiple Choice	1	1	Distinguish between fact and opinion	D
15	Multiple Choice	1	1	Determine the meaning of unfamiliar words by using context clues, a dictionary, or a glossary	D
16	Multiple Choice	1	2	Identify literary elements, such as setting, plot, and **character**, of different genres	B
17	Multiple Choice	1	2	Identify literary elements, such as **setting**, plot, and character, of different genres	C
18	Multiple Choice	1	2	Determine the meaning of unfamiliar words by using context clues, a dictionary, or a glossary	D
19	Multiple Choice	1	2	Read, view, and interpret literary texts from a variety of genres	C
20	Multiple Choice	1	3	Evaluate information, ideas, opinions, and themes in texts by identifying a central idea and supporting details	C
21	Short Response	2	3	Evaluate information, ideas, opinions, and themes in texts by identifying a central idea and supporting details	n/a
Book 2	**Listening and Writing**				
22	Multiple Choice	1	1	Identify information that is implicit rather than stated	B
23	Multiple Choice	1	1	Identify essential details for note taking	B
24	Multiple Choice	1	1	Identify information that is implicit rather than stated	A
25	Multiple Choice	1	1	Identify essential details for note taking	C
26	Short Response	2	3	Form an opinion on a subject on the basis of information, ideas, and themes expressed in presentations	n/a
27	Editing Paragraph	3	n/a	Observe the rules of punctuation, capitalization, and spelling	n/a

Table 4d. NYSTP ELA 2010 Operational Test Map, Grade 6

Question	Type	Points	Standard	Performance Indicator	Answer Key
Book 1	**Reading**				
1	Multiple Choice	1	1	Use text features, such as headings, captions, and titles, to understand and interpret informational texts	D
2	Multiple Choice	1	1	Read to collect and interpret data, facts, and ideas from multiple sources	B
3	Multiple Choice	1	1	Identify information that is implied rather than stated	D
4	Multiple Choice	1	1	Identify information that is implied rather than stated	B
5	Multiple Choice	1	3	Identify different perspectives, such as social, cultural, ethnic, and historical, on an issue presented in one or more than one text	D
6	Multiple Choice	1	2	Identify literary elements (e.g., **setting,** plot, character, rhythm, and rhyme) of different genres	B
7	Multiple Choice	1	2	Identify literary elements (e.g., setting, **plot,** character, rhythm, and rhyme) of different genres	A
8	Multiple Choice	1	2	Identify literary elements (e.g., setting, plot, **character,** rhythm, and rhyme) of different genres	B
9	Multiple Choice	1	2	Identify the ways in which characters change and develop throughout a story	B
10	Multiple Choice	1	2	Identify literary elements (e.g., setting, **plot,** character, rhythm, and rhyme) of different genres	A
11	Multiple Choice	1	2	Define characteristics of different genres	C
12	Multiple Choice	1	1	Read to collect and interpret data, facts, and ideas from multiple sources	B
13	Multiple Choice	1	1	Distinguish between fact and opinion	D
14	Multiple Choice	1	1	Read to collect and interpret data, facts, and ideas from multiple sources	C
15	Multiple Choice	1	1	Read to collect and interpret data, facts, and ideas from multiple sources	B
16	Multiple Choice	1	1	Determine the meaning of unfamiliar words by using context clues, a dictionary, or a glossary	D
17	Multiple Choice	1	2	Identify literary elements (e.g., setting, **plot,** character, rhythm, and rhyme) of different genres	A
18	Multiple Choice	1	2	Identify literary elements (e.g., setting, **plot,** character, rhythm, and rhyme) of different genres	D
19	Multiple Choice	1	2	Identify the ways in which characters change and develop throughout a story	C
20	Multiple Choice	1	2	Define characteristics of different genres	B
21	Multiple Choice	1	3	Evaluate information, ideas, opinions, and themes by identifying a central idea and supporting details	A
22	Multiple Choice	1	1	Read to collect and interpret data, facts, and ideas from multiple sources	B

(Continued on next page)

Table 4d. NYSTP ELA 2010 Operational Test Map, Grade 6 (cont.)

Question	Type	Points	Standard	Performance Indicator	Answer Key
Book 1	**Reading**				
23	Multiple Choice	1	1	Read to collect and interpret data, facts, and ideas from multiple sources	A
24	Multiple Choice	1	1	Read to collect and interpret data, facts, and ideas from multiple sources	C
25	Multiple Choice	1	1	Read to collect and interpret data, facts, and ideas from multiple sources	A
26	Multiple Choice	1	1	Determine the meaning of unfamiliar words by using context clues, a dictionary, or a glossary	D
Book 2	**Listening and Writing**				
27–30	Short and Extended Response	5	2	Listening/Writing cluster	n/a
Book 3	**Reading and Writing**				
31–34	Short and Extended Response	5	3	Reading/Writing cluster	n/a
Book 2 & Book 3	**Writing Mechanics**				
30 & 34	Extended Response	3	n/a	Writing Mechanics cluster	n/a

Table 4e. NYSTP ELA 2010 Operational Test Map, Grade 7

Question	Type	Points	Standard	Performance Indicator	Answer Key
Book 1	**Reading**				
1	Multiple Choice	1	2	Identify the author's point of view, such as first-person narrator and omniscient narrator	C
2	Multiple Choice	1	2	Determine how the use and meaning of literary devices (e.g., symbolism, metaphor and simile, alliteration, personification, flashback, and foreshadowing) convey the author's message or intent	D
3	Multiple Choice	1	2	Recognize how the author's use of language creates images or feelings	B
4	Multiple Choice	1	2	Interpret characters, **plot**, setting, and theme, using evidence from the text	D
5	Multiple Choice	1	2	Interpret **characters**, plot, setting, and theme, using evidence from the text	B
6	Multiple Choice	1	2	Determine the meaning of unfamiliar words by using context clues, a dictionary, a glossary, and structural analysis (i.e., looking at roots, prefixes, and suffixes of words)	A

(Continued on next page)

Table 4e. NYSTP ELA 2010 Operational Test Map, Grade 7 (cont.)

Question	Type	Points	Standard	Performance Indicator	Answer Key
Book 1	**Reading**				
7	Multiple Choice	1	1	Use knowledge of structure, content, and vocabulary to understand informational text	C
8	Multiple Choice	1	1	Interpret data, facts, and ideas from informational texts by applying thinking skills, such as define, classify, and infer	C
9	Multiple Choice	1	1	Interpret data, facts, and ideas from informational texts by applying thinking skills, such as define, classify, and infer	A
10	Multiple Choice	1	3	Evaluate the validity and accuracy of information, ideas, themes, opinions, and experiences in text to evaluate examples, details, or reasons used to support ideas	B
11	Multiple Choice	1	2	Interpret characters, plot, setting, and **theme**, using evidence from the text	D
12	Multiple Choice	1	2	Identify poetic elements, such as repetition, rhythm, and rhyming patterns, in order to interpret poetry	B
13	Multiple Choice	1	3	Evaluate the validity and accuracy of information, ideas, themes, opinions, and experiences in text to identify multiple levels of meaning	C
14	Multiple Choice	1	2	Determine the meaning of unfamiliar words by using context clues, a dictionary, a glossary, and structural analysis (i.e., looking at roots, prefixes, and suffixes of words)	A
15	Multiple Choice	1	2	Identify the author's point of view, such as first-person narrator and omniscient narrator	A
16	Multiple Choice	1	2	Interpret **characters**, plot, setting, and theme, using evidence from the text	C
17	Multiple Choice	1	2	Recognize how the author's use of language creates images or feelings	B
18	Multiple Choice	1	2	Interpret characters, **plot**, setting, and theme, using evidence from the text	B
19	Multiple Choice	1	3	Evaluate the validity and accuracy of information, ideas, themes, opinions, and experiences in text to identify multiple levels of meaning	B
20	Multiple Choice	1	1	Interpret data, facts, and ideas from informational texts by applying thinking skills, such as define, classify, and infer	A
21	Multiple Choice	1	1	Draw conclusions and make inferences on the basis of explicit and implied information	B
22	Multiple Choice	1	1	Draw conclusions and make inferences on the basis of explicit and implied information	C

(Continued on next page)

Table 4e. NYSTP ELA 2010 Operational Test Map, Grade 7 (cont.)

Question	Type	Points	Standard	Performance Indicator	Answer Key
Book 1	**Reading**				
23	Multiple Choice	1	3	Evaluate the validity and accuracy of information, ideas, themes, opinions, and experiences in text to consider the background and qualifications of the writer	C
24	Multiple Choice	1	1	Use knowledge of structure, content, and vocabulary to understand informational text	B
25	Multiple Choice	1	3	Evaluate the validity and accuracy of information, ideas, themes, opinions, and experiences in text to evaluate examples, details, or reasons used to support ideas	B
26	Multiple Choice	1	1	Determine the meaning of unfamiliar words by using context clues, a dictionary, a glossary, and structural analysis (i.e., looking at roots, prefixes, and suffixes of words)	A
27	Short Response	2	1	Interpret data, facts, and ideas from informational texts by applying thinking skills, such as define, classify, and infer	n/a
28	Short Response	2	3	Evaluate the validity and accuracy of information, ideas, themes, opinions, and experiences in text to identify cultural and ethnic values and their impact on content	n/a
Book 2	**Listening and Writing**				
29	Multiple Choice	1	1	Draw conclusions and make inferences on the basis of explicit information	A
30	Multiple Choice	1	1	Draw conclusions and make inferences on the basis of explicit information	A
31	Multiple Choice	1	1	Draw conclusions and make inferences on the basis of explicit information	C
32	Multiple Choice	1	1	Recall significant ideas and details, and describe the relationships between and among them	B
33	Short Response	2	1	Support ideas with examples, definitions, analogies, and direct references to the text	n/a
34	Short Response	2	3	Present clear analysis, using examples, details, and reasons from text	n/a
35	Editing Paragraph	3	n/a	Observe rules of punctuation, italicization, capitalization, and spelling Use correct grammatical construction	n/a

Table 4f. NYSTP ELA 2010 Operational Test Map, Grade 8

Question	Type	Points	Standard	Performance Indicator	Answer Key
Book 1	**Reading**				
1	Multiple Choice	1	2	Identify the author's point of view, such as first-person narrator and omniscient narrator	A
2	Multiple Choice	1	2	Interpret characters, **plot,** setting, theme, and dialogue, using evidence from the text	D
3	Multiple Choice	1	2	Determine the meaning of unfamiliar words by using context clues, a dictionary, a glossary, and structural analysis (i.e., looking at roots, prefixes, and suffixes of words)	C
4	Multiple Choice	1	2	Recognize how the author's use of language creates images or feelings	B
5	Multiple Choice	1	2	Interpret **characters,** plot, setting, theme, and dialogue, using evidence from the text	D
6	Multiple Choice	1	2	Interpret characters, **plot,** setting, theme, and dialogue, using evidence from the text	C
7	Multiple Choice	1	1	Draw conclusions and make inferences on the basis of explicit and implied information	B
8	Multiple Choice	1	1	Determine the meaning of unfamiliar words by using context clues, a dictionary, a glossary, and structural analysis (i.e., looking at roots, prefixes, and suffixes of words)	A
9	Multiple Choice	1	1	Draw conclusions and make inferences on the basis of explicit and implied information	C
10	Multiple Choice	1	2	Identify social and cultural contexts and other characteristics of the time period in order to enhance understanding and appreciation of text	C
11	Multiple Choice	1	1	Draw conclusions and make inferences on the basis of explicit and implied information	A
12	Multiple Choice	1	1	Determine the meaning of unfamiliar words by using context clues, a dictionary, a glossary, and structural analysis (i.e., looking at roots, prefixes, and suffixes of words)	D
13	Multiple Choice	1	2	Interpret characters, plot, **setting,** theme, and dialogue, using evidence from the text	A
14	Multiple Choice	1	3	Evaluate the validity and accuracy of information, ideas, themes, opinions, and experiences in texts to evaluate examples, details, or reasons used to support ideas	C
15	Multiple Choice	1	2	Determine how the use and meaning of literary devices, such as symbolism, metaphor and simile, illustration, personification, flashback, and foreshadowing convey the author's message or intent	D
16	Multiple Choice	1	2	Determine how the use and meaning of literary devices, such as symbolism, metaphor and simile, illustration, personification, flashback, and foreshadowing convey the author's message or intent	B

(Continued on next page)

Table 4f. NYSTP ELA 2010 Operational Test Map, Grade 8 (cont.)

Question	Type	Points	Standard	Performance Indicator	Answer Key
Book 1	**Reading**				
17	Multiple Choice	1	2	Determine the meaning of unfamiliar words by using context clues, a dictionary, a glossary, and structural analysis (i.e., looking at roots, prefixes, and suffixes of words)	A
18	Multiple Choice	1	3	Evaluate the validity and accuracy of information, ideas, themes, opinions, and experiences in texts to question the writer's assumptions, beliefs, intentions, and biases	B
19	Multiple Choice	1	2	Identify the author's point of view, such as first-person narrator and omniscient narrator	D
20	Multiple Choice	1	3	Evaluate the validity and accuracy of information, ideas, themes, opinions, and experiences in texts to identify multiple levels of meaning	B
21	Multiple Choice	1	2	Determine how the use and meaning of literary devices, such as symbolism, metaphor and simile, illustration, personification, flashback, and foreshadowing convey the author's message or intent	C
22	Multiple Choice	1	1	Distinguish between relevant and irrelevant information	B
23	Multiple Choice	1	1	Draw conclusions and make inferences on the basis of explicit and implied information	B
24	Multiple Choice	1	3	Evaluate the validity and accuracy of information, ideas, themes, opinions, and experiences in texts to evaluate examples, details, or reasons used to support ideas	C
25	Multiple Choice	1	1	Determine the meaning of unfamiliar words by using context clues, a dictionary, a glossary, and structural analysis (i.e., looking at roots, prefixes, and suffixes of words)	C
26	Multiple Choice	1	1	Use indexes to locate information and glossaries to define terms	A
Book 2	**Listening and Writing**				
27–30	Short and Extended Response	5	1	Listening/Writing cluster	n/a
Book 3	**Reading and Writing**				
31–34	Short and Extended Response	5	3	Reading/Writing cluster	n/a
Book 2 & Book 3	**Writing Mechanics**				
30 & 34	Extended Response	3	n/a	Writing Mechanics cluster	n/a

2010 Item Mapping by New York State Standards

Table 5. NYSTP ELA 2010 Standard Coverage

Grade	Standard	MC Item #s	CR Item #s	Total Items	Total Points
3	S1	6, 7, 8, 9, 16, 17, 18, 20	21	9	10
3	S2	1, 2, 3, 4, 11, 12, 13, 14, 22, 23, 24	26, 27	13	15
3	S3	5, 10, 15, 19, 25	n/a	5	5
4	S1	7, 8, 9, 10, 11, 12, 23, 24, 25, 26, 27, 28	n/a	12	12
4	S2	1, 2, 3, 6, 13, 14, 15, 16, 17, 18, 19, 20, 21	29, 30, 31	16	17
4	S3	4, 5, 22	32, 33, 34, 35	7	7
5	S1	1, 2, 3, 4, 5, 11, 12, 14, 15, 22, 23, 24, 25	n/a	13	13
5	S2	7, 8, 9, 10, 16, 17, 18, 19	n/a	8	8
5	S3	6, 13, 20	21, 26	5	7
6	S1	1, 2, 3, 4, 12, 13, 14, 15, 16, 22, 23, 24, 25, 26	n/a	14	14
6	S2	6, 7, 8, 9, 10, 11, 17, 18, 19, 20	27, 28, 29, 30	14	15
6	S3	5, 21	31, 32, 33, 34	6	7
7	S1	7, 8, 9, 20, 21, 22, 24, 26, 29, 30, 31, 32	27, 33	14	16
7	S2	1, 2, 3, 4, 5, 6, 11, 12, 14, 15, 16, 17, 18	n/a	13	13
7	S3	10, 13, 19, 23, 25	28, 34	7	9
8	S1	7, 8, 9, 11, 12, 22, 23, 25, 26	27, 28, 29, 30	13	14
8	S2	1, 2, 3, 4, 5, 6, 10, 13, 15, 16, 17, 19, 21	n/a	13	13
8	S3	14, 18, 20, 24	31, 32, 33, 34	8	9

New York State Educators' Involvement in Test Development

New York State educators are actively involved in ELA Test development at different test stages, including the following events: passage review, item review, rangefinding, and test form final-eyes review. These events are described in details in the later sections of this report. The New York State Education Department gathers a diverse group of educators to review all test materials in order to create fair and valid tests. The participants are selected for each testing event based on:

- certification and appropriate grade-level experience
- geographical region
- gender
- ethnicity

The selected participants must be certified and have both teaching and testing experience. The majority of them are classroom teachers, but specialists, such as reading coaches, literacy coaches, as well as special education and bilingual instructors, also participate. Some participants are also recommended by principals, professional organizations, Big Five Cities, the Staff and Curriculum Development Network (SCDN), etc. Other criteria are also considered, such as gender, ethnicity, geographic location, and type of school (urban, suburban, and rural). A file of participants is maintained and is routinely updated, with current participant information and the addition of possible future participants as recruitment forms are received. This gives many educators the opportunity to participate in the test development process. Every effort is made to have diverse groups of educators participate in each testing event.

Content Rationale

In June 2004, CTB/McGraw-Hill facilitated test specifications meetings in Albany, New York, during which committees of state educators, along with NYSED staff, reviewed the standards and performance indicators to make the following determinations:

- which performance indicators were to be assessed
- which item types were to be used for the assessable performance indicators (For example, some performance indicators lend themselves more easily to assessment by CR items than others.)
- how much emphasis to place on each assessable performance indicator (For example, some performance indicators encompass a wider range of skills than others, necessitating a broader range of items in order to fully assess the performance indicator.)
- how the limitations, if any, were to be applied to the assessable performance indicators (For example, some portions of a performance indicator may be more appropriately assessed in the classroom than on a paper-and-pencil test.)
- what general examples of items could be used
- what the test blueprint was to be for each grade

The committees were composed of teachers from around the state who were selected for their grade-level expertise, were grouped by grade band (i.e., Grades 3/4, 5/6, 7/8), and met for four days. The committees were composed of approximately ten participants per grade band. Upon completion of the committee meetings, NYSED reviewed the committees' determinations and approved them, with minor adjustments when necessary to maintain consistency across the grades.

Item Development

The first step in the process of item development for the 2010 Grades 3–8 ELA Tests was selection of passages to be used. The CTB/McGraw-Hill passage selectors were provided

with specifications based on the test design (see Appendix A). After an internal CTB/McGraw-Hill editorial and supervisory review, the passages were submitted to NYSED for their approval and then brought to a formal passage review meeting in Albany, New York, in March 2008. The purpose of the meeting was for committees of New York educators to review and decide whether to approve the passages. CTB/McGraw-Hill and NYSED staff were both present, with CTB/McGraw-Hill staff facilitating. After the committees completed their reviews, NYSED reviewed and approved the committees' decisions regarding the passages.

The lead-content editors at CTB/McGraw-Hill then selected from the approved passages those passages that would best elicit the types of items outlined during the test specifications meetings and distributed writing assignments to experienced item writers. The writers' assignments outlined the number and type of items (including depth-of-knowledge or thinking skill level) to write for each passage. Writers were trained in the New York State Testing Program and in the test specifications. This training entailed specific assignments that spelled out the performance indicators and depth-of-knowledge levels to assess for each passage. In addition, item writers were trained in the New York State Learning Standards and specifications (which provide information such as limitations and examples for assessing performance indicators) and were provided with item-writing guidelines (see Appendix B), sample New York State test items, and the New York State Style Guide.

CTB/McGraw-Hill editors and supervisors reviewed the items to verify that they met the specifications and criteria outlined in the writing assignments and, as necessary, revised them. After all revisions from CTB/McGraw-Hill staff had been incorporated, the items were submitted to NYSED staff for their review and approval. CTB/McGraw-Hill incorporated any necessary revisions from NYSED and prepared the items for a formal item review.

Item Review

As was done for the specifications and passage review meetings, the item review committees were composed of New York State educators selected for their content and grade-level expertise. Each committee was composed of approximately 10 participants per grade band (i.e., Grades 3/4, 5/6, and 7/8). The committee members were provided with the test items, the New York State Learning Standards, and the test specifications, and they considered the following elements as they reviewed the test items:

- the accuracy and grade-level appropriateness of the items
- the mapping of the items to the assigned performance indicators
- the accompanying exemplary responses (CR items)
- the appropriateness of the correct response and distractors (MC items)
- the conciseness, preciseness, clarity, and reading load of the items
- the existence of any ethnic, gender, regional, or other possible bias evident in the items

Upon completion of the committee work, NYSED reviewed the decisions of the committee members; NYSED either approved the changes to the items or suggested additional revisions so that the nature and format of the items were consistent across

grades and with the format and style of the testing program. All approved changes were then incorporated into the items prior to field testing.

Materials Development

Following item review, CTB/McGraw-Hill staff assembled the approved passages and items into field test (FT) forms and submitted the FT forms to NYSED for their review and approval. The Grades 3–5 ELA FTs were administered to students across New York State during January 26–30, 2009, and the Grades 6–8 ELA FTs were administered during February 2–6, 2009, using the State Sampling Matrix to ensure appropriate sampling of students. In addition, CTB/McGraw-Hill, in conjunction with NYSED test specialists, developed a FT *Teacher's Directions and School Administrator's Manual* to help ensure that the FTs were administered in a uniform manner to all participating students. FT forms were assigned to participants at the school (grade) level while balancing the demographic statistics across forms, in order to proactively sample the students.

After administration of the FTs, rangefinding sessions were conducted in March 2009 in New York State to examine a sampling of student responses to the short- and extended-response items. Committees of New York State educators with content and grade-level expertise were again assembled. Each committee was composed of approximately eight to ten participants per grade level. CTB/McGraw-Hill staff facilitated the meetings, and NYSED staff reviewed the decisions made by the committees and verified that the decisions made were consistent across grades. The committees' charge was to select student responses that exemplified each score point of each CR item. These responses, in conjunction with the rubrics, were then used by CTB/McGraw-Hill scoring staff to score the CR FT items.

Item Selection and Test Creation (Criteria and Process)

The fifth year of OP NYSTP Grades 3–8 ELA Tests were administered in April 2010. The test items were selected from the pool of items primarily field-tested in 2006, 2007, 2008, and 2009, using the data from those FTs, CTB/McGraw-Hill made preliminary selections for each grade. The selections were reviewed for alignment with the test design, blueprint, and the research guidelines for item selection (Appendix C). Item selection for the NYSTP Grades 3–8 ELA Tests was based on the classical and item response theory (IRT) statistics of the test items. Selection was conducted by content experts from CTB/McGraw-Hill and NYSED and reviewed by psychometricians at CTB/McGraw-Hill and at NYSED. Final approval of the selected items was given by NYSED. Two criteria governed the item selection process. The first of these was to meet the content specifications provided by NYSED. Second, within the limits set by these requirements, developers selected items with the best psychometric characteristics from the FT item pool.

Item selection for the OP tests was facilitated using the proprietary program ITEMWIN (Burket, 1988). This program creates an interactive connection between the developer selecting the test items and the item database. This program monitors the impact of each decision made during the item selection process and offers a variety of options for grouping, classifying, sorting, and ranking items to highlight key information as it is needed (see Green, Yen, and Burket, 1989).

The program has three parts. The first part of the program selects a working item pool of manageable size from the larger pool. The second part uses this selected item pool to perform the final test selection. The third part of the program includes a table showing the expected number correct and the standard error of ability estimate (a function of scale score), as well as statistical and graphic summaries on bias, fit, and the standard error of the final test. Any fault in the final selection becomes apparent as the final statistics are generated. Examples of possible faults that may occur are cases where the test is too easy or too difficult, contains items demonstrating differential item functioning (DIF), or does not adequately measure part of the range of performance. A developer detecting any such problems can then return to the second stage of the program and revise the selection. The flexibility and utility of the program encourages multiple attempts at fine-tuning the item selection. After preliminary selections were completed, the items were reviewed for alignment with the test design, blueprint, and research guidelines for item selection (see Appendix C).

The NYSED staff (including content and research experts) traveled to CTB/McGraw-Hill in Monterey in July 2009 to finalize item selection and test creation. There, they discussed the content and data of the proposed selections, explored alternate selections for consideration, determined the final item selections, and ordered those items (assigned positions) in the OP test books. The final test forms were approved by the final eyes committee that consisted of approximately 20 participants across all grade levels. After the approval by NYSED, the tests were produced and administered in April 2010.

In addition to the test books, CTB/McGraw-Hill and NYSED produced a *School Administrator's Manual*, as well as two *Teacher's Directions*, one for Grades 3, 4, and 5 and one for Grades 6, 7, and 8, so that the tests were administered in a standardized fashion across the state. These documents are located at the following web site: http://www.p12.nysed.gov/osa/english/home.html#ei

Proficiency and Performance Standards

The original proficiency cut score recommendations and the drafting of performance standards occurred at the NYSTP ELA standard setting review held in Albany in June 2006. In 2010, change in the test administration window between the 2008–2009 and 2009–2010 school years and a decision to align the proficiency standards with Grade 8 student performance on the NYS Regents ELA exams led to changes in the proficiency cut scores. The results were reviewed by the NYS Technical Advisory Group and were approved by the Board of Regents in July 2010. For each grade level, there are four proficiency levels. Three cut points demarcate the performance standards needed to demonstrate each ascending level of proficiency.

Section III: Validity

Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests. Test validation is an ongoing process of gathering evidence from many sources to evaluate the soundness of the desired score interpretation or use. This evidence is acquired from studies of the content of the test as well as from studies involving scores produced by the test. Additionally, reliability is a necessary test to conduct before considerations of validity are made. A test cannot be valid if it is not also reliable.

The American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) *Standards for Educational and Psychological Testing* (1999) addressed the concept of validity in testing. Validity is the most important consideration in test evaluation. The concept refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores. Test validation is the process for accumulating evidence to support any particular inference. Validity, however, is a unitary concept. Although evidence may be accumulated in many ways, validity refers to the degree to which evidence supports the inferences made from test scores.

Content Validity

Generally, achievement tests are used for student-level outcomes, either for making predictions about students or for describing students' performances (Mehrens and Lehmann, 1991). In addition, tests are now also used for the purpose of accountability and adequate yearly progress (AYP). NYSED uses various assessment data in reporting AYP. Specific to student-level outcomes, NYSTP documents student performance in the area of ELA as defined by the New York State ELA Learning Standards. To allow test score interpretations appropriate for this purpose, the content of the test must be carefully matched to the specified standards. The 1999 AERA/APA/NCME standards state that content-related evidence of validity is a central concern during test development. Expert professional judgment should play an integral part in developing the definition of what is to be measured, such as describing the universe of the content, generating or selecting the content sample, and specifying the item format and scoring system.

Logical analysis of test content indicates the degree to which the content of a test covers the domain of content the test is intended to measure. In the case of the NYSTP, the content is defined by detailed, written specifications and blueprints that describe New York State content standards and define the skills that must be measured to assess these content standards (see Tables 3–5 in Section II). The test development process requires specific attention to content representation and the balance within each test form. New York State educators were involved in test constructions in various test development stages. For example, during the item review process, they reviewed FTs for their alignment with the test blueprint. Educators also participated in a process of establishing scoring rubrics (during rangefinding sessions) for CR items. Section II, "Test Design and Development," contains more information specific to the item review process. An independent study of alignment between the New York State curriculum and the New York State Grades 3–8 ELA Tests was conducted using Norman Webb's method. The

results of the study provided additional evidence of test content validity (refer to *An External Alignment Study for New York State's Assessment Program*, April 2006, Educational Testing Services).

Construct (Internal Structure) Validity

Construct validity, what scores mean and what kind of inferences they support, is often considered the most important type of test validity. Construct validity of the New York State Grades 3–8 ELA Tests is supported by several types of evidence that can be obtained from the ELA test data.

Internal Consistency

Empirical studies of the internal structure of the test provide one type of evidence of construct validity. For example, high internal consistency constitutes evidence of validity. This is because high coefficients imply that the test questions are measuring the same domain of skill and are reliable and consistent. Reliability coefficients of the tests for total populations and subgroups of students are presented in Section VIII, "Reliability and Standard Error of Measurement." For the total population, the reliability coefficients (Cronbach's alpha) ranged from 0.83–0.88, and for most subgroups the reliability coefficient was equal or greater than 0.80 (the exceptions were for Grade 4 students from districts classified as Charter and Grades 5 and 8 students from districts classified as Charter and Low Needs.). Overall, high internal consistency of the New York State ELA Tests provided sound evidence of construct validity.

Unidimensionality

Other evidence comes from analyses of the degree to which the test questions conform to the requirements of the statistical models used to scale and equate the tests, as well as to generate student scores. Among other things, the models require that the items fit the model well and the questions in a test measure a single domain of skill: that they are unidimensional. The item-model fit was assessed using Q1 statistics (Yen, 1981) and the results are described in detail in Section VI, "IRT Scaling and Equating." It was found that except for items 8 and 14 in Grade 3 test, item 8 in Grade 5 test, items 2 and 23 in Grade 7 test, and items 24, 26, and 28 in Grade 8 test, all other items on the 2010 Grades 3–8 ELA Tests displayed good item-model fit, which provided solid evidence for the appropriateness of IRT models used to calibrate and scale the test data. Another evidence for the efficacy of modeling ability was provided by demonstrating that the questions on New York State ELA Tests were related. What relates the questions is most parsimoniously claimed to be the common ability acquired by students studying the subject. Factor analysis of the test data is one way of modeling the common ability. This analysis may show that there is a single or main factor that can account for much of the variability among responses to test questions. A large first component would provide evidence of the latent ability students have in common with respect to the particular questions asked. A large main factor found from a factor analysis of an achievement test would suggest a primary ability construct that may be considered related to what the questions were designed to have in common, i.e., English language arts ability.

To demonstrate the common factor (ability) underlying student responses to ELA test items, a principal component factor analysis was conducted on a correlation matrix of

individual items for each test. Factoring a correlation matrix rather than actual item response data is preferable when dichotomous variables are in the analyzed data set. Because the New York State ELA Tests contain both MC and CR items, the matrix of polychoric correlations was used as input for factor analysis (polychoric correlation is an extension of tetrachoric correlations that are appropriate only for MC items). The study was conducted on the total population of New York State public and charter school students in each grade. A large first principal component was evident in each analysis.

More than one factor with an eigenvalue greater than 1.0 present in each data set would suggest the presence of small additional factors. However, the ratio of the variance accounted for by the first factor to the remaining factors was sufficiently large to support the claim that these tests were essentially unidimensional. These ratios showed that the first eigenvalues were at least four times as large as the second eigenvalues for all the grades. In addition, the total amount of variance accounted for by the main factor was evaluated. According to M. Reckase (1979), "*… the 1PL and the 3PL models estimate different abilities when a test measures independent factors, but … both estimate the first principal component when it is large relative to the other factors. In this latter case, good ability estimates can be obtained from the models, even when the first factor accounts for less than 10 percent of the test variance, although item calibration results will be unstable*." It was found that all the New York State Grades 3–8 ELA Tests exhibited first principal components accounting for more than 10% of the test variance. The results of factor analysis including eigenvalues greater than 1.0 and proportion of variance explained by extracted factors are presented in Table 6.

Table 6. Factor Analysis Results for ELA Tests (Total Population)

Grade	Initial Eigenvalues			
	Component	Total	% of Variance	Cumulative %
3	**1**	**6.18**	**22.06**	**22.06**
	2	1.16	4.15	26.21
	3	1.06	3.80	30.01
	4	1.01	3.60	33.60
4	**1**	**6.49**	**20.93**	**20.93**
	2	1.39	4.50	25.43
	3	1.06	3.43	28.85
	4	1.02	3.29	32.15
5	**1**	**5.77**	**21.38**	**21.38**
	2	1.31	4.85	26.23
	3	1.08	4.00	30.23
6	**1**	**7.52**	**25.94**	**25.94**
	2	1.35	4.64	30.57

(Continued on next page)

Table 6. Factor Analysis Results for ELA Tests (Total Population) (cont.)

Grade	Initial Eigenvalues			
	Component	Total	% of Variance	Cumulative %
7	**1**	**7.71**	**22.02**	**22.02**
	2	1.39	3.98	26.00
	3	1.11	3.18	29.18
8	**1**	**6.14**	**21.18**	**21.18**
	2	1.19	4.10	25.29

This evidence supports the claim that there is a construct ability underlying the items/tasks in each ELA Test and that scores from each test would be representing performance primarily determined by that ability. Construct-irrelevant variance does not appear to create significant nuisance factors.

As additional evidence for construct validity, the same factor analysis procedure was employed to assess dimensionality of ELA construct for selected subgroups of students in each grade: English language learners (ELL), students with disabilities (SWD), and students using test accommodations (SUA). The results were comparable to the results obtained from the total population data. Evaluation of eigenvalue magnitude and proportions of variance explained by the main and secondary factors provide evidence of essential unidimensionality of the construct measured by the ELA Tests for the analyzed subgroups. Factor analysis results for ELL, SWD, SUA, ELL/SUA, and SWD/SUA classifications are provided in Table D1 of Appendix D. ELL/SUA subgroup is defined as examinees whose ELL status are true and use one or more ELL-related accommodation. SWD/SUA subgroup includes examinees who are classified with disabilities and use one or more disability-related accommodations.

Minimization of Bias

Minimizing item bias contributes to minimization of construct-irrelevant variance and contributes to improved test validity. The developers of the NYSTP tests gave careful attention to questions of possible ethnic, gender, and socioeconomic status (SES) bias. All materials were written and reviewed to conform to CTB/McGraw-Hill's editorial policies and guidelines for equitable assessment, as well as NYSED's guidelines for item development. At the same time, all materials were written to NYSED's specifications and carefully checked by groups of trained New York State educators during the item review process.

Four procedures were used to eliminate bias and minimize differential item functioning (DIF) in the New York State ELA Tests.

The first procedure was based on the premise that careful editorial attention to validity is an essential step in keeping bias to a minimum. Bias occurs if the test is differentially valid for a given group of test takers. If the test entails irrelevant skills or knowledge, the possibility of DIF is increased. Thus, preserving content validity is essential.

The second procedure was to follow the item-writing guidelines established by NYSED. Developers reviewed NYSTP materials with these guidelines in mind. These internal editorial reviews were done by at least four separate people: the content editor, who directly supervises the item writers; the project director; a style editor; and a proofreader. The final test built from the field test materials was reviewed by at least these same people.

In the third procedure, New York State educators who reviewed all FT materials were asked to consider and comment on the appropriateness of language, content, and gender and cultural distribution.

It is believed that these three procedures improved the quality of the New York State tests and reduced bias. However, current evidence suggests that expertise in this area is no substitute for data; reviewers are sometimes wrong about which items work to the disadvantage of a group, apparently because some of their ideas about how students will react to items may be faulty (Sandoval and Mille, 1979; Jensen, 1980). Thus, empirical studies were conducted.

In the fourth procedure, statistical methods were used to identify items exhibiting possible DIF. Although items flagged for DIF in the FT stage were closely examined for content bias and avoided during the OP test construction, DIF analyses were conducted again on OP test data. Three methods were employed to evaluate the amount of DIF in all test items: standardized mean difference, Mantel-Haenszel (see Section V "Operational Test Data Collection and Classical Analysis"), and Linn-Harnisch (see Section VI, "IRT Scaling and Equating"). A few items in each grade were flagged for DIF, and typically the amount of DIF present was not large. Very few items were flagged by multiple methods. Items that were flagged for statistically significant DIF were carefully reviewed by multiple reviewers during the OP test item selection. Only those items deemed free of bias were included in the OP tests.

Section IV: Test Administration and Scoring

Listed in this section are brief summaries of New York State test administration and scoring procedures. For further information, refer to the *New York State ELA Scoring Leader Handbook* and *School Administrator's Manual*. In addition, please refer to *the Scoring Site Operations Manual* (2010) located at http://www.p12.nysed.gov/osa/ei/ssom-10.pdf

Test Administration

NYSTP Grades 3–8 ELA Tests were administered at the classroom level during April and May 2010. The testing window for Grades 3–8 was April 26–28. The makeup test administration window for Grades 3–8 was April 27–May 5. The makeup test administration windows allowed students who were ill or otherwise unable to test during the assigned window to take the test.

Scoring Procedures of Operational Tests

The scoring of the OP test was performed at designated sites by qualified teachers and administrators. The number of personnel at a given site varied, as districts have the option of regional, districtwide, or schoolwide scoring (please refer to the next subsection, "Scoring Models," for more detail). Administrators were responsible for the oversight of scoring operations, including the preparation of the test site, the security of test books, and the supervision of the scoring process. At each site, designated trainers taught scoring committee members the basic criteria for scoring each question and monitored the scoring sessions in the room. The trainers were assisted by facilitators or leaders who also helped in monitoring the sessions and enforced scoring accuracy. The titles for administrators, trainers, and facilitators vary by the scoring model that is selected. At the regional level, oversight was conducted by a site coordinator. A scoring leader trained the scoring committee members and monitored the sessions, and a table facilitator assisted in monitoring the sessions. At the districtwide level, a school district administrator oversaw OP scoring. A district ELA leader trained the scoring committee members and monitored the sessions, and a school ELA leader assisted in monitoring the sessions. For schoolwide scoring, oversight was provided by the principal; otherwise, titles for the schoolwide model were the same as those for the districtwide model. The general title "scoring committee members" included scorers at every site.

Scoring Models

For the 2009–10 school year, schools and school districts used local decision-making processes to select the model that best met their needs for the scoring of the Grades 3–8 ELA Tests. Schools were able to score these tests regionally, districtwide, or individually. Schools were required to enter one of the following scoring model codes on student answer sheets:

1. Regional scoring—The scorers for the school's test papers included either staff from three or more school districts or staff from all nonpublic schools in an affiliation group (nonpublic or charter schools may participate in regional scoring with public school districts and may be counted as one district);

2. Schools from two districts—The scorers for the school's test papers included staff from two school districts, nonpublic schools, charter school districts, or a combination thereof;

3. Three or more schools within a district—The scorers for the school's test papers included staff from all schools administering this test in a district, provided at least three schools are represented;

4. Two schools within a district—The scorers for the school's test papers included staff from all schools administering this test in a district, provided that two schools are represented; or

5. One school only (local scoring)—The first readers for the school's test papers included staff from the only school in the district administering this test, staff from one charter school, or staff from one nonpublic school.

Schools and districts were instructed to carefully analyze their individual needs and capacities to determine their appropriate scoring model. BOCES and the Staff and Curriculum Development Network (SCDN) provided districts with technical support and advice in making this decision.

For further information, refer to the following link for a brief comparison between regional, district, and local scoring: http://www.p12.nysed.gov/osa/ei/ssom-10.pdf (see Appendices B and C).

Scoring of Constructed-Response Items

The scoring of CR items was based primarily on the scoring guides, which were created by CTB/McGraw-Hill handscoring and content development specialists with guidance from NYSED and New York State teachers during rangefinding sessions conducted after each FT. The CTB ELA handscoring team was composed of six supervisors, each representing one grade. Supervisors are selected on the basis of their handscoring experiences along with their educational and professional backgrounds.

In March 2009, CTB/McGraw-Hill staff met with groups of teachers from across the state in rangefinding sessions. Sets of actual FT student responses were reviewed and discussed openly, and consensus scores were agreed upon by the teachers based on the teaching methods and criteria across the state, as well as on NYSED policies. In addition, audio files were created to further explain each section of the scoring guides. Trainers used these materials to train scoring committee members on the criteria for scoring CR items. Scoring Leader Handbooks were also distributed to outline the responsibilities of the scoring roles. CTB/McGraw-Hill handscoring staff also conducted training sessions in New York City to better equip these teachers and administrators with enhanced knowledge of scoring principles and criteria.

Scoring was conducted with pen and pencil scoring as opposed to electronic scoring, and each scoring committee member evaluated actual student papers instead of electronically scanned papers. All scoring committee members were trained by previously trained and

approved trainers along with guidance from scoring guides and a CD containing the audio files that highlighted important elements of the scoring guides. Each test book was scored by three separate scoring committee members, who scored three distinct sections of the test book. After test books were completed, the table facilitator or ELA leader conducted a "read-behind" of approximately 12 sets of test books per hour to verify the accuracy of scoring. If a question arose that was not covered in the training materials, facilitators or trainers were to call the New York State ELA Helpline (see the subsection "Quality Control Process").

Scorer Qualifications and Training

The scoring of the OP test was conducted by qualified administrators and teachers. Trainers used the scoring guides and audio files to train scoring committee members on the criteria for scoring CR items. Part of the training process was the administration of a consistency assurance set (CAS) that provided the State's scoring sites with information regarding strengths and weaknesses of their scorers. This tool allowed trainers to retrain their scorers, if necessary. The CAS also acknowledged those scorers who had grasped all aspects of the content area being scored and were well prepared to score student responses.

Quality Control Process

Test books were randomly distributed throughout each scoring room so that books from each region, district, school, or class were evenly dispersed. Teams were divided into groups of three to ensure that a variety of scorers graded each book. If a scorer and facilitator could not reach a decision on a paper after reviewing the scoring guides and audio files, they called the New York State ELA Helpline. This call center was established to help teachers and administrators during OP scoring. The helpline staff consisted of trained CTB/McGraw-Hill handscoring personnel who answered questions by phone, fax, or email. When a member of the staff was unable to resolve an issue, they deferred to NYSED for a scoring decision. A quality check was also performed on each completed box of scored tests to certify that all questions were scored and that the scoring committee members darkened each score on the answer document appropriately. The log of calls received during the scoring Helpline was delivered to NYSED after the scoring window. To affirm that all schools across the state adhered to scoring guidelines and policies, approximately 5% of the schools' results are audited each year by an outside vendor.

Section V: Operational Test Data Collection and Classical Analysis

Data Collection

OP test data were collected in two phases. During phase 1, a sample of approximately 98% of the student test records were received from the data warehouse and delivered to CTB/McGraw-Hill in May 2010. These data were used for all data analysis. Phase 2 involved submitting "straggler files" to CTB/McGraw-Hill in early-June 2010. The straggler files were later merged with the main data sets. The straggler files contained around 2% of the total population cases and due to late submission were excluded from research data analyses. Data from nonpublic schools were excluded from any data analysis.

Data Processing

Data processing refers to the cleaning and screening procedures used to identify errors (such as out-of-range data) and the decisions made to exclude student cases or to suppress particular items in analyses. CTB/McGraw-Hill established a scoring program, EDITCHECKER, to do initial quality assurance on data and identify errors. This program verifies that the data fields are in-range (as defined), that students' identifying information is present, and that the data are acceptable for delivery to CTB/McGraw-Hill research. NYSED and the data repository were provided with the results of the checking, and some edits to the initial data were made; however, CTB/McGraw-Hill research performs data cleaning to the delivered data and excludes some student cases in order to obtain a sample of the utmost integrity. It should be noted that the major groups of cases excluded from the data set were out-of-grade students (students whose grade level did not match the test level) and students from nonpublic schools. Other deleted cases included students with no grade-level data and duplicate record cases. A list of the data cleaning procedures conducted by research and accompanying case counts is presented in Tables 7a–7f.

Table 7a. NYSTP ELA Grade 3 Data Cleaning

Exclusion Rule	# Deleted	# Cases Remain
Initial # of cases		196437
Out of grade	84	196353
No grade	2	196351
Duplicate record	0	196351
Non-public and out-of-district schools	3063	193288
Missing values for ALL items on OP form	0	193288
Out-of-range CR scores	0	193288

Table 7b. NYSTP ELA Grade 4 Data Cleaning

Exclusion Rule	# Deleted	# Cases Remain
Initial # of cases		209404
Out of grade	70	209334
No grade	10	209324
Duplicate record	0	209324
Non-public and out-of-district schools	13141	196183
Missing values for ALL items on OP form	3	196180
Out-of-range CR scores	0	196180

Table 7c. NYSTP ELA Grade 5 Data Cleaning

Exclusion Rule	# Deleted	# Cases Remain
Initial # of cases		197719
Out of grade	33	197686
No grade	5	197681
Duplicate record	0	197681
Non-public and out-of-district schools	2899	194782
Missing values for ALL items on OP form	0	194782
Out-of-range CR scores	0	194782

Table 7d. NYSTP ELA Grade 6 Data Cleaning

Exclusion Rule	# Deleted	# Cases Remain
Initial # of cases		204546
Out of grade	119	204427
No grade	0	204427
Duplicate record	0	204427
Non-public and out-of-district schools	10275	194152
Missing values for ALL items on OP form	0	194152
Out-of-range CR scores	0	194152

Table 7e. NYSTP ELA Grade 7 Data Cleaning

Exclusion Rule	# Deleted	# Cases Remain
Initial # of cases		198461
Out of grade	145	198316
No grade	0	198316
Duplicate record	0	198316
Non-public and out-of-district schools	2912	195404
Missing values for ALL items on OP form	1	195403
Out-of-range CR scores	0	195403

Table 7f. NYSTP ELA Grade 8 Data Cleaning

Exclusion Rule	# Deleted	# Cases Remain
Initial # of cases		213650
Out of grade	166	213484
No grade	0	213484
Duplicate record	0	213484
Non-public and out-of-district schools	12364	201120
Missing values for ALL items on OP form	3	201117
Out-of-range CR scores	0	201117

Classical Analysis and Calibration Sample Characteristics

The demographic characteristics of students in the cleaned calibration and equating data sets are presented in the proceeding tables. The clean data sets included over 95% of New York State students and were used for classical analyses presented in this section and calibrations. The needs resource code (NRC) is assigned at district level and is an indicator of district and school socioeconomic status. The ethnicity and gender designations are assigned at the student level. Please note that the tables do not include data for gender variable as it was found that the New York State population is fairly evenly split by gender categories.

Table 8a. Grade 3 Sample Characteristics (N = 193288)

Demographic Category		N-count	% of Total N-count
NRC	NYC	69525	36.09
	Big cities	8338	4.33
	Urban/Suburban	15550	8.07
	Rural	11381	5.91
	Average needs	56785	29.47
	Low needs	26979	14.00
	Charter	4111	2.13
Ethnicity	Asian	14861	7.69
	Black	36617	18.94
	Hispanic	42987	22.24
	American Indian	949	0.49
	Multi-Racial	1086	0.56
	Unknown	112	0.06
	White	96676	50.02
ELL	No	175091	90.59
	Yes	18197	9.41
SWD	No	165781	85.77
	Yes	27507	14.23
SUA	No	146915	76.01
	Yes	46373	23.99

Table 8b. Grade 4 Sample Characteristics (N = 196180)

Demographic Category		N-count	% of Total N-count
NRC	NYC	70150	35.87
	Big cities	8106	4.15
	Urban/Suburban	15713	8.03
	Rural	11376	5.82
	Average needs	58319	29.82
	Low needs	28506	14.58
	Charter	3390	1.73
Ethnicity	Asian	15898	8.10
	Black	37197	18.96
	Hispanic	42242	21.53
	American Indian	926	0.47
	Multi-Racial	969	0.49
	Unknown	109	0.06
	White	98839	50.38
ELL	No	180067	91.79
	Yes	16113	8.21
SWD	No	167070	85.16
	Yes	29110	14.84
SUA	No	148478	75.68
	Yes	47702	24.32

Table 8c. Grade 5 Sample Characteristics (N = 194782)

Demographic Category		N-count	% of Total N-count
NRC	NYC	67531	34.78
	Big cities	7843	4.04
	Urban/Suburban	15163	7.81
	Rural	11479	5.91
	Average needs	58421	30.09
	Low needs	29275	15.08
	Charter	4434	2.28
Ethnicity	Asian	14928	7.66
	Black	37177	19.09
	Hispanic	41347	21.23
	American Indian	910	0.47
	Multi-Racial	846	0.43
	Unknown	91	0.05
	White	99483	51.07
ELL	No	181886	93.38
	Yes	12896	6.62
SWD	No	165020	84.72
	Yes	29762	15.28
SUA	No	148520	76.25
	Yes	46262	23.75

Table 8d. Grade 6 Sample Characteristics (N = 194152)

Demographic Category		N-count	% of Total N-count
NRC	NYC	66896	34.59
	Big cities	7499	3.88
	Urban/Suburban	14563	7.53
	Rural	11430	5.91
	Average needs	60294	31.17
	Low needs	29099	15.04
	Charter	3640	1.88
Ethnicity	Asian	14589	7.51
	Black	36870	18.99
	Hispanic	40615	20.92
	American Indian	948	0.49
	Multi-Racial	737	0.38
	Unknown	98	0.05
	White	100295	51.66
ELL	No	183566	94.55
	Yes	10586	5.45
SWD	No	164221	84.58
	Yes	29931	15.42
SUA	No	152048	78.31
	Yes	42104	21.69

Table 8e. Grade 7 Sample Characteristics (N = 195403)

Demographic Category		N-count	% of Total N-count
NRC	NYC	68304	35.11
	Big cities	7598	3.91
	Urban/Suburban	14122	7.26
	Rural	11662	5.99
	Average needs	61125	31.42
	Low needs	28881	14.84
	Charter	2862	1.47
Ethnicity	Asian	14834	7.59
	Black	37120	19.00
	Hispanic	40007	20.47
	American Indian	952	0.49
	Multi-Racial	701	0.36
	Unknown	69	0.04
	White	101720	52.06
ELL	No	185305	94.83
	Yes	10098	5.17
SWD	No	165850	84.88
	Yes	29553	15.12
SUA	No	155163	79.41
	Yes	40240	20.59

Table 8f. Grade 8 Sample Characteristics (N = 201117)

Demographic Category		N-count	% of Total N-count
NRC	NYC	70809	35.39
	Big cities	7486	3.74
	Urban/Suburban	14560	7.28
	Rural	11963	5.98
	Average needs	62539	31.26
	Low needs	30380	15.19
	Charter	2328	1.16
Ethnicity	Asian	15291	7.60
	Black	37841	18.82
	Hispanic	41286	20.53
	American Indian	912	0.45
	Multi-Racial	582	0.29
	Unknown	92	0.05
	White	105113	52.26
ELL	No	191112	95.03
	Yes	10005	4.97
SWD	No	171082	85.07
	Yes	30035	14.93
SUA	No	159910	79.51
	Yes	41207	20.49

Classical Data Analysis

Classical data analysis of the Grades 3–8 ELA Tests consists of four primary elements. One element is the analysis of item level statistical information about student performance. It is important to verify that the items and test forms function as intended. Information on item response patterns, item difficulty (p-value), and item-test correlation (point biserial) is examined thoroughly. If any serious error were to occur with an item (i.e., a printing error or potentially correct distractor), item analysis is the stage that errors should be flagged and evaluated for rectification (suppression, credit, or other acceptable solution). Analyses of test level data comprise the second element of classical data analysis. These include examination of the raw score statistics (mean and standard deviation) and test reliability measures (Cronbach's alpha and Feldt-Raju coefficient). Assessment of test speededness is another important element of classical analysis. Additionally, classical differential item functioning (DIF) analysis is conducted at this stage. DIF analysis includes computation of standardized mean differences and Mantel-Haenszel statistics for New York State items to identify potential item bias. All classical data analysis results contribute information on the validity and reliability of the tests (also see Sections III, "Validity," and VIII, "Reliability and Standard Error of Measurement").

Item Difficulty and Response Distribution

Item difficulty and response distribution tables (Table 9a–9f) illustrate student test performance, as observed from both MC and CR item responses. Omit rates signify the percentage of students who did not attempt the item. For MC items, "% at 0" represents the percentage of students who double-bubbled responses, and other "% SEL" categories represent the percentage of students who selected each answer response (without double marking). Proportions of students who selected the correct answer option are denoted with an asterisk (*) and are repeated in the p-value field. For CR items, the "% at 0," "% SEL," and "% at 5" (only in Grades 6 and 8) categories depict the percentage of students who earned a valid score on the item, from zero to the maximum score.

Item difficulty is classically measured by the p-value statistic. It assesses the proportion of students who responded correctly to each MC item or the average proportion of the maximum score that students earned on each CR item. It is important to have a good range of p-values to increase test information and to avoid floor or ceiling effects. Generally, p-values should range between 0.30 and 0.90. P-values represent the overall degree of difficulty, but do not account for demonstrated student performance on other test items. Usually, p-value information is coupled with point biserial (pbis) statistics, to verify that items are functioning as intended (point biserials are discussed in the next subsection). Item difficulties (p-values) on the ELA Tests ranged from 0.29 to 0.97. For Grade 3, the item p-values were between 0.49 and 0.95 with a mean of 0.82. For Grade 4, the item p-values were between 0.29 and 0.97 with a mean of 0.77. For Grade 5, the item p-values were between 0.54 and 0.96 with a mean of 0.82. For Grade 6, the item p-values were between 0.46 and 0.95 with a mean of 0.82. For Grade 7, the item p-values were between 0.53 and 0.96 with a mean of 0.82. For Grade 8, the item p-values were between 0.59 and 0.95 with a mean of 0.80. These p-value statistics are also provided in Tables 9a–9f, along with other classical test summary statistics.

Table 9a. P-values, Scored Response Distributions, and Point Biserials, Grade 3

Item	N-count	P-value	% Omit	% at 0	% Sel Option 1	% Sel Option 2	% Sel Option 3	% Sel Option 4	Pbis Option 1	Pbis Option 2	Pbis Option 3	Pbis Option 4	Pbis Key
1	193203	0.93	0.02	0.00	3.52	2.98	0.86	92.60	-0.26	-0.23	-0.17	0.40*	0.40
2	193167	0.92	0.04	0.00	4.71	91.81	1.71	1.72	-0.40	0.54*	-0.20	-0.29	0.54
3	193030	0.91	0.09	0.00	91.07	2.15	4.29	2.35	0.36*	-0.25	-0.19	-0.16	0.36
4	193020	0.94	0.10	0.00	2.86	93.72	1.58	1.70	-0.20	0.37*	-0.23	-0.19	0.37
5	192870	0.72	0.15	0.00	9.27	6.77	71.84	11.91	-0.16	-0.19	0.33*	-0.16	0.33
6	193022	0.86	0.07	0.00	85.67	10.64	1.89	1.66	0.30*	-0.14	-0.17	-0.30	0.30
7	192941	0.91	0.10	0.00	1.79	1.86	5.05	91.12	-0.29	-0.25	-0.23	0.44*	0.44
8	192824	0.49	0.17	0.00	11.96	7.00	31.49	49.31	-0.29	-0.16	0.00	0.28*	0.28
9	192885	0.92	0.17	0.00	2.02	3.43	91.75	2.59	-0.22	-0.23	0.40*	-0.22	0.40
10	192863	0.86	0.19	0.00	3.80	6.95	85.57	3.46	-0.22	-0.18	0.42*	-0.29	0.42
11	192947	0.75	0.13	0.00	8.22	8.82	74.61	8.17	-0.21	-0.17	0.34*	-0.14	0.34
12	192957	0.83	0.13	0.00	83.13	5.54	6.26	4.89	0.42*	-0.24	-0.17	-0.28	0.42
13	192881	0.91	0.16	0.00	2.23	4.10	90.45	3.01	-0.25	-0.31	0.47*	-0.21	0.47
14	192572	0.57	0.25	0.00	56.42	18.74	11.78	12.68	0.15*	-0.01	-0.15	-0.05	0.15
15	192341	0.73	0.31	0.00	2.56	11.86	12.90	72.19	-0.26	-0.15	-0.24	0.39*	0.39
16	192896	0.90	0.13	0.00	90.26	2.57	4.18	2.79	0.44*	-0.30	-0.22	-0.21	0.44
17	192857	0.84	0.16	0.00	3.81	3.06	83.89	9.01	-0.26	-0.29	0.36*	-0.10	0.36
18	192757	0.77	0.20	0.00	76.35	9.80	6.63	6.95	0.43*	-0.17	-0.31	-0.19	0.43
19	192595	0.67	0.29	0.00	9.19	67.24	15.91	7.30	-0.23	0.37*	-0.17	-0.16	0.37
20	192068	0.77	0.62	0.00	5.96	76.27	6.99	10.15	-0.24	0.40*	-0.18	-0.21	0.40
21	192016	0.83	0.66	8.18	17.01	74.15							
22	193147	0.93	0.06	0.00	5.64	0.38	93.39	0.52	-0.27	-0.11	0.32*	-0.14	0.32
23	193116	0.62	0.07	0.00	29.54	61.69	3.81	4.87	-0.06	0.20*	-0.17	-0.18	0.20
24	193050	0.95	0.08	0.00	94.89	2.09	0.50	2.40	0.35*	-0.19	-0.18	-0.23	0.35
25	193015	0.79	0.13	0.00	7.50	79.17	7.12	6.07	-0.17	0.46*	-0.29	-0.28	0.46
26	192893	0.87	0.20	4.17	16.92	78.71							
27	192824	0.91	0.24	6.51	5.61	87.64							
28	192900	0.92	0.20	3.40	3.02	7.31	86.07						

Table 9b. P-values, Scored Response Distributions, and Point Biserials, Grade 4

Item	N-count	P-value	% Omit	% at 0	% Sel Option 1	% Sel Option 2	% Sel Option 3	% Sel Option 4	Pbis Option 1	Pbis Option 2	Pbis Option 3	Pbis Option 4	Pbis Key
1	196142	0.87	0.01	0.00	87.24	1.30	9.76	1.68	0.30*	-0.22	-0.17	-0.19	0.30
2	196114	0.95	0.02	0.00	1.40	1.46	95.05	2.05	-0.21	-0.20	0.38*	-0.23	0.38
3	195978	0.78	0.06	0.00	78.07	4.38	4.50	12.94	0.32*	-0.21	-0.19	-0.14	0.32
4	196055	0.97	0.04	0.00	1.53	96.47	1.19	0.75	-0.22	0.34*	-0.13	-0.23	0.34
5	195982	0.88	0.06	0.00	2.67	3.50	87.63	6.11	-0.20	-0.16	0.37*	-0.24	0.37
6	196062	0.94	0.04	0.00	93.78	3.18	1.15	1.83	0.38*	-0.25	-0.21	-0.18	0.38
7	196022	0.96	0.04	0.00	95.66	2.42	0.80	1.04	0.30*	-0.17	-0.15	-0.21	0.30
8	195911	0.67	0.09	0.00	14.68	67.34	7.50	10.35	-0.22	0.36*	-0.14	-0.17	0.36
9	195909	0.73	0.09	0.00	6.61	13.98	6.34	72.93	-0.27	-0.15	-0.22	0.39*	0.39
10	195885	0.91	0.08	0.00	2.98	3.66	1.99	91.23	-0.28	-0.25	-0.25	0.46*	0.46
11	195966	0.87	0.08	0.00	5.34	86.91	4.89	2.75	-0.29	0.46*	-0.24	-0.23	0.46
12	195937	0.76	0.09	0.00	76.01	4.98	11.61	7.28	0.32*	-0.20	-0.16	-0.16	0.32
13	195868	0.90	0.12	0.00	6.18	90.33	1.66	1.66	-0.26	0.38*	-0.17	-0.21	0.38
14	195795	0.85	0.14	0.00	1.52	12.06	1.48	84.75	-0.21	-0.26	-0.25	0.40*	0.40
15	195859	0.90	0.13	0.00	89.96	4.91	1.83	3.14	0.46*	-0.28	-0.26	-0.22	0.46
16	195697	0.77	0.18	0.00	7.41	9.05	6.00	77.29	-0.26	-0.23	-0.16	0.42*	0.42
17	195741	0.62	0.19	0.00	61.86	30.71	3.84	3.37	0.39*	-0.21	-0.27	-0.20	0.39
18	195550	0.88	0.28	0.00	3.09	5.84	87.33	3.41	-0.26	-0.16	0.39*	-0.23	0.39
19	195342	0.53	0.35	0.00	53.15	16.05	8.54	21.84	0.16*	-0.14	-0.15	0.04	0.16
20	195339	0.72	0.36	0.00	6.31	71.44	3.81	18.01	-0.24	0.42*	-0.26	-0.19	0.42
21	195255	0.55	0.43	0.00	14.27	55.10	4.83	25.33	-0.23	0.25*	-0.24	0.03	0.25
22	195127	0.84	0.46	0.00	4.36	4.88	6.23	84.00	-0.24	-0.23	-0.19	0.42*	0.42
23	194719	0.63	0.69	0.00	9.86	7.44	19.52	62.43	-0.29	-0.17	-0.11	0.38*	0.38
24	194600	0.88	0.77	0.00	3.41	5.62	86.92	3.25	-0.22	-0.26	0.46*	-0.23	0.46
25	194403	0.91	0.85	0.00	89.81	3.11	3.28	2.89	0.35*	-0.17	-0.18	-0.19	0.35
26	194063	0.29	1.02	0.00	29.10	19.75	26.42	23.66	0.11*	-0.10	0.01	-0.01	0.11
27	193847	0.55	1.15	0.00	5.66	4.80	34.31	54.04	-0.31	-0.23	-0.11	0.37*	0.37
28	193702	0.75	1.25	0.00	7.84	73.60	8.76	8.54	-0.27	0.47*	-0.20	-0.23	0.47
29	196062	0.67	0.06	0.38	5.94	32.28	46.15	15.20					
30	196033	0.69	0.07	0.83	5.80	28.20	46.39	18.70					
31	196032	0.73	0.08	1.06	16.61	45.91	36.33						

Table 9c. P-values, Scored Response Distributions, and Point Biserials, Grade 5

Item	N-count	P-value	% Omit	% at 0	% Sel Option 1	% Sel Option 2	% Sel Option 3	% Sel Option 4	Pbis Option 1	Pbis Option 2	Pbis Option 3	Pbis Option 4	Pbis Key
1	194751	0.92	0.01	0.00	1.85	4.09	92.29	1.75	-0.24	-0.22	0.37*	-0.16	0.37
2	194707	0.88	0.02	0.00	5.95	88.28	1.34	4.39	-0.22	0.36*	-0.17	-0.21	0.36
3	194712	0.94	0.03	0.00	2.91	93.90	1.38	1.76	-0.23	0.39*	-0.21	-0.23	0.39
4	194651	0.95	0.03	0.00	95.18	1.77	1.80	1.18	0.41*	-0.19	-0.28	-0.22	0.41
5	194597	0.68	0.05	0.00	12.48	14.30	4.84	68.28	-0.19	-0.17	-0.19	0.35*	0.35
6	194649	0.88	0.05	0.00	2.72	87.53	8.44	1.24	-0.24	0.38*	-0.24	-0.17	0.38
7	194580	0.93	0.07	0.00	0.85	3.07	3.45	92.52	-0.15	-0.25	-0.13	0.31*	0.31
8	194549	0.54	0.10	0.00	19.83	10.06	54.14	15.86	-0.08	-0.09	0.18*	-0.07	0.18
9	194659	0.92	0.05	0.00	1.23	3.24	91.88	3.59	-0.25	-0.21	0.41*	-0.24	0.41
10	194642	0.91	0.05	0.00	3.65	91.28	2.43	2.57	-0.22	0.39*	-0.24	-0.19	0.39
11	194521	0.81	0.12	0.00	80.98	4.21	5.60	9.08	0.38*	-0.21	-0.19	-0.21	0.38
12	194587	0.82	0.08	0.00	82.01	2.73	13.31	1.86	0.22*	-0.21	-0.08	-0.16	0.22
13	194477	0.58	0.12	0.00	20.33	8.51	12.73	58.27	-0.09	-0.21	-0.31	0.40*	0.40
14	194468	0.88	0.11	0.00	5.23	3.31	3.64	87.66	-0.31	-0.27	-0.26	0.51*	0.51
15	194426	0.88	0.14	0.00	3.50	3.77	4.44	88.11	-0.27	-0.26	-0.23	0.46*	0.46
16	194118	0.71	0.31	0.00	6.49	71.17	6.02	15.98	-0.12	0.32*	-0.26	-0.13	0.32
17	194038	0.76	0.33	0.00	4.90	6.58	76.15	11.99	-0.32	-0.28	0.49*	-0.19	0.49
18	194017	0.94	0.35	0.00	2.05	1.49	2.81	93.26	-0.23	-0.22	-0.19	0.38*	0.38
19	193872	0.72	0.43	0.00	23.59	2.25	72.10	1.59	-0.25	-0.25	0.39*	-0.21	0.39
20	193636	0.78	0.58	0.00	6.52	5.78	77.14	9.97	-0.20	-0.18	0.45*	-0.31	0.45
21	192892	0.66	0.97	17.57	32.22	49.24							
22	194662	0.91	0.05	0.00	3.10	91.25	4.51	1.08	-0.26	0.38*	-0.21	-0.18	0.38
23	194648	0.85	0.06	0.00	7.46	84.45	2.37	5.65	-0.20	0.40*	-0.21	-0.26	0.40
24	194526	0.92	0.07	0.00	91.76	3.56	0.37	4.17	0.32*	-0.22	-0.13	-0.20	0.32
25	194506	0.97	0.13	0.00	0.71	1.39	96.42	1.34	-0.12	-0.14	0.25*	-0.17	0.25
26	194588	0.76	0.10	5.28	37.32	57.30							
27	194448	0.59	0.17	14.75	22.96	34.01	28.11						

Table 9d. P-values, Scored Response Distributions, and Point Biserials, Grade 6

Item	N-count	P-value	% Omit	% at 0	% Sel Option 1	% Sel Option 2	% Sel Option 3	% Sel Option 4	% Sel Option 5	Pbis Option 1	Pbis Option 2	Pbis Option 3	Pbis Option 4	Pbis Key
1	194097	0.89	0.02	0.00	3.01	6.35	2.00	88.62		-0.26	-0.32	-0.16	0.46*	0.46
2	194006	0.76	0.06	0.00	12.27	75.76	6.99	4.90		-0.22	0.41*	-0.18	-0.27	0.41
3	194018	0.89	0.03	0.00	3.20	5.95	1.46	89.32		-0.15	-0.29	-0.13	0.36*	0.36
4	194068	0.90	0.03	0.00	0.69	90.45	5.43	3.39		-0.17	0.46*	-0.26	-0.33	0.46
5	193995	0.94	0.04	0.00	1.97	1.49	3.00	93.46		-0.22	-0.20	-0.17	0.35*	0.35
6	194056	0.90	0.03	0.00	1.31	90.33	0.48	7.83		-0.17	0.37*	-0.10	-0.31	0.37
7	194062	0.94	0.03	0.00	94.03	2.07	2.46	1.39		0.43*	-0.22	-0.25	-0.26	0.43
8	194061	0.95	0.04	0.00	2.47	95.42	1.41	0.65		-0.19	0.32*	-0.18	-0.18	0.32
9	193999	0.90	0.06	0.00	1.40	89.76	5.05	3.71		-0.21	0.47*	-0.31	-0.26	0.47
10	194018	0.87	0.05	0.00	87.29	8.15	2.04	2.46		0.38*	-0.23	-0.25	-0.19	0.38
11	194002	0.87	0.06	0.00	2.18	6.78	87.28	3.68		-0.25	-0.26	0.45*	-0.25	0.45
12	193987	0.87	0.06	0.00	2.37	87.37	6.14	4.03		-0.26	0.43*	-0.20	-0.27	0.43
13	194003	0.90	0.06	0.00	4.21	2.80	2.86	90.06		-0.28	-0.30	-0.25	0.50*	0.50
14	194036	0.92	0.05	0.00	1.79	2.22	91.49	4.44		-0.28	-0.25	0.51*	-0.32	0.51
15	193960	0.84	0.07	0.00	3.15	83.71	2.30	10.74		-0.23	0.46*	-0.24	-0.30	0.46
16	193910	0.91	0.06	0.00	3.06	2.68	2.90	91.23		-0.26	-0.27	-0.29	0.50*	0.50
17	193866	0.83	0.11	0.00	83.24	2.66	7.09	6.86		0.44*	-0.26	-0.18	-0.29	0.44
18	193888	0.89	0.10	0.00	4.24	3.21	3.97	88.45		-0.34	-0.29	-0.27	0.55*	0.55
19	193866	0.73	0.13	0.00	7.58	8.73	73.06	10.49		-0.23	-0.25	0.37*	-0.09	0.37
20	193751	0.46	0.16	0.00	12.60	45.91	9.12	32.17		-0.20	0.26*	-0.24	0.02	0.26
21	193732	0.69	0.17	0.00	68.85	10.77	4.96	15.21		0.41*	-0.21	-0.18	-0.23	0.41
22	193382	0.76	0.36	0.00	4.05	75.71	4.84	15.00		-0.29	0.31	-0.24	-0.06	0.31
23	193422	0.93	0.34	0.00	92.34	1.88	2.30	3.11		0.44*	-0.22	-0.29	-0.22	0.44
24	193300	0.81	0.39	0.00	8.96	5.10	80.21	5.29		-0.22	-0.24	0.45*	-0.26	0.45
25	193193	0.84	0.47	0.00	83.91	5.92	5.04	4.63		0.54*	-0.30	-0.29	-0.28	0.54
26	193130	0.58	0.51	0.00	3.81	3.37	34.85	57.43		-0.27	-0.17	-0.17	0.35*	0.35
27	194001	0.70	0.08	0.55	4.18	12.34	29.28	34.82	18.75					
28	193980	0.60	0.09	1.00	9.61	21.77	32.01	25.01	10.51					
29	193944	0.75	0.11	0.81	14.17	44.39	40.53							

Table 9e. P-values, Scored Response Distributions, and Point Biserials, Grade 7

Item	N-count	P-value	% Omit	% at 0	% Sel Option 1	% Sel Option 2	% Sel Option 3	% Sel Option 4	Pbis Option 1	Pbis Option 2	Pbis Option 3	Pbis Option 4	Pbis Key
1	195317	0.71	0.03	0.00	24.57	2.15	70.59	2.65	-0.27	-0.20	0.39*	-0.19	0.39
2	195211	0.82	0.07	0.00	5.75	6.86	5.71	81.58	-0.03	-0.35	-0.22	0.38*	0.38
3	195111	0.79	0.13	0.00	10.64	78.57	7.23	3.41	-0.23	0.42*	-0.25	-0.20	0.42
4	195189	0.93	0.06	0.00	2.51	2.43	2.23	92.72	-0.24	-0.24	-0.25	0.43*	0.43
5	195266	0.91	0.06	0.00	5.57	91.20	2.54	0.62	-0.32	0.45*	-0.27	-0.14	0.45
6	195273	0.94	0.04	0.00	94.01	2.99	2.10	0.84	0.38*	-0.24	-0.20	-0.20	0.38
7	195220	0.85	0.08	0.00	2.95	3.96	84.72	8.28	-0.24	-0.19	0.34*	-0.16	0.34
8	195207	0.82	0.08	0.00	5.95	1.41	81.50	11.05	-0.20	-0.21	0.43*	-0.30	0.43
9	195265	0.94	0.06	0.00	93.85	2.79	1.19	2.10	0.35*	-0.25	-0.16	-0.18	0.35
10	195272	0.81	0.05	0.00	1.43	81.36	9.54	7.60	-0.19	0.40*	-0.17	-0.31	0.40
11	195219	0.96	0.05	0.00	1.54	0.69	1.67	96.01	-0.21	-0.18	-0.26	0.38*	0.38
12	195220	0.76	0.08	0.00	4.54	76.00	13.40	5.97	-0.19	0.35*	-0.21	-0.14	0.35
13	195234	0.94	0.07	0.00	1.11	1.96	93.99	2.85	-0.23	-0.23	0.41*	-0.24	0.41
14	195278	0.95	0.05	0.00	95.41	2.66	1.27	0.59	0.40*	-0.27	-0.23	-0.17	0.40
15	195158	0.86	0.08	0.00	85.77	7.98	0.89	5.24	0.47*	-0.41	-0.19	-0.14	0.47
16	195195	0.80	0.08	0.00	2.90	4.93	80.35	11.70	-0.28	-0.16	0.43*	-0.27	0.43
17	195238	0.91	0.07	0.00	1.99	91.26	3.81	2.86	-0.25	0.41*	-0.24	-0.20	0.41
18	195219	0.90	0.07	0.00	5.12	90.17	1.22	3.40	-0.22	0.38*	-0.20	-0.23	0.38
19	195113	0.80	0.12	0.00	7.18	80.16	4.88	7.63	-0.19	0.44*	-0.22	-0.29	0.44
20	194976	0.75	0.19	0.00	74.42	2.46	15.99	6.91	0.30*	-0.28	-0.13	-0.14	0.30
21	194982	0.77	0.17	0.00	4.33	76.80	6.50	12.15	-0.22	0.41*	-0.26	-0.19	0.41
22	194902	0.81	0.22	0.00	10.43	5.36	80.81	3.14	-0.30	-0.26	0.50*	-0.25	0.50
23	194884	0.69	0.21	0.00	6.98	13.43	68.55	10.78	-0.32	0.02	0.24*	-0.10	0.24
24	194582	0.88	0.39	0.00	4.58	87.93	3.44	3.63	-0.28	0.41*	-0.22	-0.17	0.41
25	194220	0.54	0.57	0.00	30.99	53.27	10.89	4.25	-0.12	0.34*	-0.22	-0.19	0.34
26	194183	0.89	0.61	0.00	88.80	4.74	3.28	2.57	0.43*	-0.21	-0.28	-0.22	0.43
27	190269	0.63	2.63	20.43	30.23	46.71							
28	188236	0.77	3.67	8.22	27.47	60.64							
29	195121	0.77	0.13	0.00	77.36	8.07	10.95	3.48	0.36*	-0.22	-0.17	-0.18	0.36
30	195113	0.79	0.14	0.00	78.98	0.93	18.39	1.56	0.40*	-0.16	-0.32	-0.16	0.40
31	195094	0.93	0.14	0.00	1.20	1.11	93.06	4.46	-0.20	-0.12	0.39*	-0.30	0.39
32	195085	0.91	0.16	0.00	6.87	91.23	1.19	0.55	-0.24	0.32*	-0.17	-0.13	0.32
33	195127	0.81	0.14	3.13	31.70	65.03							
34	194802	0.73	0.31	5.89	42.42	51.38							
35	194853	0.62	0.28	14.21	19.38	32.91	33.22						

Table 9f. P-values, Scored Response Distributions, and Point Biserials, Grade 8

Item	N-count	P-value	% Omit	% at 0	% Sel Option 1	% Sel Option 2	% Sel Option 3	% Sel Option 4	% Sel Option 5	Pbis Option 1	Pbis Option 2	Pbis Option 3	Pbis Option 4	Pbis Key
1	201073	0.95	0.02	0.00	94.71	1.42	2.60	1.25		0.31*	-0.19	-0.19	-0.14	0.31
2	201022	0.93	0.04	0.00	1.72	2.32	2.63	93.29		-0.12	-0.21	-0.26	0.36*	0.36
3	201042	0.94	0.03	0.00	3.50	1.53	94.12	0.81		-0.24	-0.22	0.37*	-0.17	0.37
4	201026	0.81	0.03	0.00	17.42	81.30	0.90	0.33		-0.09	0.16*	-0.21	-0.14	0.16
5	201001	0.93	0.04	0.00	1.18	4.45	1.59	92.72		-0.23	-0.21	-0.18	0.35*	0.35
6	200988	0.87	0.06	0.00	3.53	5.32	87.40	3.68		-0.24	-0.20	0.38*	-0.19	0.38
7	200911	0.78	0.09	0.00	13.33	77.96	3.34	5.26		-0.20	0.40*	-0.22	-0.25	0.40
8	200948	0.72	0.06	0.00	72.22	24.27	2.33	1.10		0.28*	-0.17	-0.23	-0.16	0.28
9	200932	0.91	0.08	0.00	2.58	2.16	91.04	4.12		-0.26	-0.22	0.42*	-0.23	0.42
10	200974	0.80	0.06	0.00	4.56	11.96	80.43	2.99		-0.28	-0.15	0.33*	-0.14	0.33
11	200940	0.91	0.06	0.00	91.32	2.96	3.61	2.03		0.33*	-0.24	-0.20	-0.11	0.33
12	200904	0.74	0.07	0.00	10.41	0.98	14.55	73.95		-0.23	-0.22	-0.19	0.37*	0.37
13	200903	0.85	0.09	0.00	84.82	8.86	1.36	4.85		0.36*	-0.20	-0.18	-0.22	0.36
14	200997	0.90	0.05	0.00	2.26	5.20	89.45	3.02		-0.23	-0.23	0.42*	-0.25	0.42
15	200884	0.81	0.08	0.00	7.74	7.96	3.72	80.46		-0.22	-0.31	-0.14	0.43*	0.43
16	200884	0.77	0.10	0.00	4.05	76.73	14.37	4.73		-0.23	0.38*	-0.14	-0.29	0.38
17	200871	0.92	0.09	0.00	91.78	1.08	1.53	5.49		0.47*	-0.21	-0.20	-0.35	0.47
18	200812	0.60	0.12	0.00	4.46	59.52	16.56	19.31		-0.18	0.24*	-0.14	-0.06	0.24
19	200843	0.79	0.10	0.00	3.46	5.16	12.61	78.63		-0.22	-0.21	-0.29	0.45*	0.45
20	200833	0.80	0.12	0.00	7.92	79.70	6.74	5.50		-0.22	0.42*	-0.24	-0.20	0.42
21	200794	0.59	0.14	0.00	22.55	11.25	59.11	6.93		-0.05	-0.16	0.25*	-0.19	0.25
22	200728	0.70	0.16	0.00	7.42	70.10	11.52	10.78		-0.18	0.38*	-0.21	-0.18	0.38
23	200721	0.71	0.18	0.00	3.10	71.29	19.38	6.03		-0.23	0.33*	-0.13	-0.22	0.33
24	200637	0.76	0.21	0.00	2.59	5.81	75.86	15.51		-0.27	-0.29	0.29*	-0.02	0.29
25	200619	0.75	0.23	0.00	4.55	17.91	74.92	2.37		-0.28	-0.14	0.34*	-0.20	0.34
26	200599	0.83	0.24	0.00	82.58	2.26	3.22	11.68		0.31*	-0.23	-0.27	-0.10	0.31
27	200614	0.71	0.25	1.26	4.82	11.14	25.64	33.10	23.79					
28	200852	0.73	0.13	0.77	4.72	10.03	23.91	34.44	26.00					
29	200709	0.77	0.20	1.32	12.35	40.48	45.65							

Point-Biserial Correlation Coefficients

Point-biserial (pbis) statistics are used to examine item-test correlations or item discrimination for MC items. In the Tables 9a–9f, point-biserial correlation coefficients were computed for each answer option. Point biserials for the correct answer option are denoted with an asterisk (*) and are repeated in the Pbis Key field. The point-biserial correlation is a measure of internal consistency that ranges between +/-1. It indicates a correlation of students' responses to an item relative to their performance on the rest of the test. The criterion for point biserial for the correct answer option used for New York State test was 0.15. The point biserials for the correct answer option that was equal to or greater than 0.15 indicated that students who responded correctly also tended to do well on the overall test. For incorrect answer options (distractors), the point biserial should be negative, which indicated that students who scored lower on the overall test had a tendency to pick a distractor. The only item that had a low point biserial was item number 26 in Grade 4 test, which had a point biserial of 0.11. Point biserials for correct answer options (pbis*) on the tests ranged 0.11–0.55. For Grade 3, the pbis* were between 0.15 and 0.54. For Grade 4, the pbis* were between 0.11 and 0.47. For Grade 5, the pbis* were between 0.18 and 0.51. For Grade 6, pbis* were between 0.26 and 0.55. For Grade 7, the pbis* were between 0.24 and 0.50. For Grade 8, the pbis* were between 0.16 and 0.47.

Distractor Analysis

Item distractors provide additional information on student performance on test questions. Two types of information on item distractors are available from New York State test data: information on proportion of students selecting incorrect item response options and the point biserial coefficient of distractors (discrimination power of incorrect answer choices). The proportions of students selecting incorrect responses while responding to MC items are provided in Tables 9a–9f of this report. Distribution of student responses across answer choices was evaluated. It was expected that the proportion of students selecting the correct answer would be higher than proportions of students selecting any other answer choice. This was true for all New York State ELA items.

As mentioned in the "Point-Biserial Correlations Coefficients" subsection, items were flagged if the point biserial of any distractor was positive. The items with a distractor that had a non-negative point biserial were item number 21 and item number 26 in Grade 4, item number 20 in Grade 6, and item number 23 in Grade 7, which had a point biserial of 0.03, 0.01, 0.02 and 0.02 respectively. All other point biserials for distractors in each grade were negative.

Test Statistics and Reliability Coefficients

Test statistics including raw-score mean and raw-score standard deviation are presented in Table 10. For both Grades 4 and 8, weighted and unweighted test statistics are provided. Grade 4 and 8 CR items were weighted by a 1.38 factor to increase proportion of score points obtainable from these items. Weighting CR items for these two grades resulted in better alignment of proportions of test raw-score points obtainable from MC and CR items between 2006 and 2010 ELA OP tests for these grades. More information on weighting CR items and the effect on test content is provided in Section VI, "IRT Scaling and Equating." Reliability coefficients provide measures of internal consistency that range from zero to one. Two reliability coefficients, Cronbach's alpha and Feldt-Raju coefficient, were computed for the

Grades 3–8 ELA Tests. Both types of reliability estimates are appropriate to use when a test contains both MC and CR items. Calculated Cronbach's alpha reliabilities ranged 0.84–0.88. Feldt-Raju reliability coefficients ranged 0.86–0.90. The lowest reliability was observed for the Grade 5 test, but since that test had the lowest number of score points, it was reasonable that its reliability would not be as high as the other grades' tests. The highest reliability was observed for the Grade 6 and Grade 7 tests. All reliabilities met or exceeded 0.80, across statistics, which is a good indication that the NYSTP 3–8 ELA Tests are acceptably reliable. High reliability indicates that scores are consistent and not unduly influenced by random error (for more information on test reliability and standard error of measurement, see Section VIII, "Reliability and Standard Error of Measurement").

Table 10. NYSTP ELA 2010 Test Form Statistics and Reliability

Grade	Max RS	RS Mean	RS SD	P-value Mean	Cronbach's Alpha	Feldt-Raju
3	33	27.41	5.12	0.82	0.86	0.87
4	39 (43 WGT)	29.43 (32.30 WGT)	5.94 (6.51 WGT)	0.77	0.86	0.87
5	31	24.64	4.94	0.82	0.84	0.86
6	39	30.51	6.33	0.82	0.87	0.90
7	41	32.75	6.78	0.82	0.88	0.89
8	39 (44 WGT)	30.55 (34.19 WGT)	6.27 (7.24 WGT)	0.80	0.86	0.88

Note: WGT = weighted results

Speededness

Speededness is the term used to refer to interference in test score observation due to insufficient testing time. Test developers considered speededness in the development of the NYSTP tests. NYSED believes that achievement tests should not be speeded; little or no useful instructional information can be obtained from the fact that a student does not finish a test, while a great deal can be learned from student responses to questions. Further, NYSED prefers all scores to be based on actual student performance, because all students should have ample opportunity to demonstrate that performance to enhance the validity of their scores. Test reliability is directly impacted by the number of test questions, so excluding questions that were impacted by a lack of timing would negatively impact reliability. For these reasons, sufficient administration time limits were set for the NYSTP tests. The research department at CTB/McGraw-Hill routinely conducts additional speededness analyses based on actual test data. The general rule of thumb is that omit rates should be less than 5.0%. Tables 9a–9f show the omit rates for items on the Grades 3–8 ELA Tests. These results provide no evidence of speededness on these tests.

Differential Item Functioning

Classical differential item functioning (DIF) was evaluated using two methods. First, the standardized mean difference (SMD) was computed for all items. The SMD statistic (Dorans, Schmitt, and Bleistein, 1992) compares the mean scores of reference and focal groups, after adjusting for ability differences. A moderate amount of significant DIF, for or against the focal group, is represented by an SMD with an absolute value between 0.10 and 0.19, inclusive. A large amount of practically significant DIF is represented by an SMD with an

absolute value of 0.20 or greater. Then, the Mantel-Haenszel method is employed to compute DIF statistics for MC items. This non-parametric DIF method partitions the sample of examinees into categories based on total raw test scores. It then compares the log-odds ratio of keyed responses for the focal and reference groups. The Mantel-Haenszel method has a critical value of 6.63 (degrees of freedom = 1 for MC items; alpha = 0.01) and is compared to its corresponding delta-value (significant when absolute value of delta > 1.50) to factor in effect size (Zwick, Donoghue, and Grima, 1993). It is important to recognize that the two methods differ in assumptions and computation; therefore, the results from both methods may not be in agreement. It should be noted that two methods of classical DIF computation and one method of IRT DIF computation (described in Section VI) were employed because no single method can identify all DIF items on a test (Hambleton, Clauser, Mazer, and Jones, 1993).

Classical DIF analyses were conducted on subgroups of the needs resource category (focal group: High Needs; reference group: Low Needs), gender (focal group: Female; reference group: Male), ethnicity (focal groups: Black, Hispanic, and Asian; reference group: White), and English language learners (focal group: English language learners; reference group: Non-English language learners). The DIF analyses were conducted using all cases from the clean data sets. Table 11 shows the number of cases for subgroups.

Table 11. NYSTP ELA 2010 Classical DIF Sample N-Counts

Grade	Ethnicity				Gender		Needs Resource Category		English Language Learner Status	
	Black/African American	Hispanic/ Latino	Asian	White	Female	Male	High	Low	Yes	No
3	36617	42987	14861	96676	94406	98882	103859	84814	18197	175091
4	37197	42242	15898	98839	95853	100327	104365	87888	16113	180067
5	37177	41347	14928	99483	94956	99826	101126	88650	12896	181886
6	36870	40615	14589	100295	95144	99008	99488	90376	10586	183566
7	37120	40007	14834	101720	95406	99997	100133	91519	10098	185305
8	37841	41286	15291	105113	98052	103065	103409	94202	10005	191112

Table 12 presents the number of items flagged for DIF by either of the classical methods described earlier. It should be noted that items showing statistically significant DIF do not necessarily pose bias. In addition to item bias, DIF may be attributed to item-impact or type-one error. All items that were flagged for significant DIF were carefully examined by multiple reviewers during OP item selection for possible item bias. Only those items that were determined free of bias were included in the OP tests.

Table 12. Number of Items Flagged by SMD and Mantel-Haenszel DIF Methods

Grade	Number of Flagged Items
3	2
4	6
5	2
6	3
7	7
8	6

A detailed list of items flagged by either one or both of these classical DIF methods, including DIF direction and associated DIF statistics, is presented in Appendix E.

Section VI: IRT Scaling and Equating

IRT Models and Rationale for Use

Item response theory (IRT) allows comparisons among items and scale scores, even those from different test forms, by using a common scale for all items and examinees (i.e., as if there were a hypothetical test that contained items from all forms). The three-parameter logistic (3PL) model (Lord and Novick, 1968; Lord, 1980) was used to analyze item responses on the MC items. For analysis of the CR items, the two-parameter partial credit (2PPC) model (Muraki, 1992; Yen, 1993) was used.

IRT is a statistical methodology that takes into account the fact that not all test items are alike and that all items do not provide the same amount of information in determining how much a student knows or can do. Computer programs that implement IRT models use actual student data to estimate the characteristics of the items on a test, called "parameters." The parameter estimation process is called "item calibration."

IRT models typically vary according to the number of parameters estimated. For the New York State tests, three parameters are estimated: the discrimination parameter, the difficulty parameter(s), and, for MC items, the guessing parameter. The discrimination parameter is an index of how well an item differentiates between high-performing and low-performing students. An item that cannot be answered correctly by low-performing students, but can be answered correctly by high-performing students, will have a high-discrimination value. The difficulty parameter is an index of how easy or difficult an item is. The higher the difficulty parameter is, the harder the item. The guessing parameter is the probability that a student with very low ability will answer the item correctly.

Because the characteristics of MC and CR items are different, two IRT models were used in item calibration. The three-parameter logistic (3PL) model was used in the analysis of MC items. In this model, the probability that a student with ability θ responds correctly to item i is

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + \exp[-1.7a_i(\theta - b_i)]},$$

where

a_i is the item discrimination, b_i is the item difficulty, and c_i is the probability of a correct response by a very low-scoring student.

For analysis of the CR items, the 2PPC model was used. The 2PPC model is a special case of Bock's (1972) nominal model. Bock's model states that the probability of an examinee with ability θ having a score $(k-1)$ at the k-th level of the j-th item is

$$P_{jk}(\theta) = P(x_j = k-1 \mid \theta) = \frac{\exp Z_{jk}}{\sum_{i=1}^{m_j} \exp Z_{ji}}, \; k = 1 \ldots m_j,$$

where

$$Z_{jk} = A_{jk}\theta + C_{jk},$$

and

k is the item response category (k = 1, 2, , …. m_j).

The m_j denotes the number of score levels for the j-th item, and typically the highest score level is assigned ($m_j - 1$) score points. For the special case of the 2PPC model used here, the following constraints were used:

$$A_{jk} = \alpha_j (k-1),$$

and

$$C_{jk} = -\sum_{i=0}^{k-1} \gamma_{ji},$$

where

$$\gamma_{j0} = 0,$$

and

α_j and γ_{ji} are the free parameters to be estimated from the data.

Each item has (m_j –1) independent γ_{ji} parameters and one α_j parameter; a total of m_j parameters are estimated for each item.

Calibration Sample

The calibration sample included response data from both the OP form and the two FT anchor forms, each containing 12 items. The data containing student responses to items included in the FT anchor forms, administered approximately two weeks after the OP test to representative samples of NYS students, were collected and used for a purpose of equating 2010 OP tests to NYS OP scales as described in ”Scaling and Equating” sub-section.

The sample representativeness of these FT anchor forms was evaluated and the OP test form and the FT form data were merged together for the calibration.

The cleaned sample data were used for calibration and scaling of New York State ELA Tests. It should be noted that the scaling was done on nearly all (96%–99%, depending on grade level) of the New York State public school student population in each tested grade and that exclusion of some cases during the data cleaning process had minimal effect on parameter estimation. As shown in Tables 13 through 15, the 2010 OP samples were comparable to 2009 populations in terms of needs resource category (NRC), student race and ethnicity, proportions of English language learners, proportions of students with disabilities, and proportions of students using testing accommodations.

Table 13. Grades 3 and 4 Demographic Statistics

Demographics	2009 Grade 3 Population	2010 Grade 3 Sample	2009 Grade 4 Population	2010 Grade 4 Sample
	%	%	%	%
NRC SUBGROUPS				
NYC	34.94	36.09	34.54	35.87
Big cities	4.17	4.33	4.10	4.15
Urban/Suburban	8.26	8.07	8.17	8.03
Rural	5.89	5.91	5.95	5.82
Average needs	30.03	29.47	30.59	29.82
Low needs	15.37	14.00	15.47	14.58
Charter	1.24	2.13	1.05	1.73
ETHNICITY				
Asian	8.04	7.69	7.51	8.10
Black	18.43	18.94	18.65	18.96
Hispanics	21.03	22.24	20.96	21.53
American Indian	0.47	0.49	0.47	0.47
Multi-Racial	0.32	0.56	0.24	0.49
White	51.65	50.02	52.12	50.38
Unknown	0.05	0.06	0.04	0.06
ELL STATUS				
No	91.19	90.59	92.68	91.79
Yes	8.81	9.41	7.32	8.21
DISABILITY				
No	86.84	85.77	85.67	85.16
Yes	13.16	14.23	14.33	14.84
ACCOMMODATIONS				
No	77.56	76.01	77.05	75.68
Yes	22.44	23.99	22.95	24.32

Table 14. Grades 5 and 6 Demographic Statistics

Demographics	2009 Grade 5 Population	2010 Grade 5 Sample	2009 Grade 6 Population	2010 Grade 6 Sample
	%	%	%	%
NRC SUBGROUPS				
NYC	34.29	34.78	34.43	34.59
Big cities	3.87	4.04	3.77	3.88
Urban/Suburban	8.00	7.81	7.63	7.53
Rural	5.93	5.91	5.78	5.91
Average needs	31.02	30.09	30.90	31.17
Low needs	15.82	15.08	15.71	15.04
Charter	0.92	2.28	1.62	1.88
ETHNICITY				
Asian	7.44	7.66	7.46	7.51
Black	18.51	19.09	19.23	18.99
Hispanics	20.70	21.23	20.47	20.92
American Indian	0.49	0.47	0.46	0.49
Multi-Racial	0.24	0.43	0.22	0.38
White	52.58	51.07	52.13	51.66
Unknown	0.05	0.05	0.04	0.05
ELL STATUS				
No	93.8	93.38	94.75	94.55
Yes	6.20	6.62	5.25	5.45
DISABILITY				
No	84.91	84.72	84.73	84.58
Yes	15.09	15.28	15.27	15.42
ACCOMMODATIONS				
No	77.02	76.25	78.93	78.31
Yes	22.98	23.75	21.07	21.69

Table 15. Grades 7 and 8 Demographic Statistics

Demographics	2009 Grade 7 Population	2010 Grade 7 Sample	2009 Grade 8 Population	2010 Grade 8 Sample
	%	%	%	%
NRC SUBGROUPS				
NYC	34.35	35.11	34.53	35.39
Big cities	3.79	3.91	3.71	3.74
Urban/Suburban	7.57	7.26	7.51	7.28
Rural	5.97	5.99	5.97	5.98
Average needs	31.16	31.42	31.46	31.26
Low needs	15.74	14.84	15.51	15.19
Charter	1.20	1.47	1.04	1.16
ETHNICITY				
Asian	7.34	7.59	7.25	7.60
Black	18.91	19.00	19.02	18.82
Hispanics	20.30	20.47	20.20	20.53
American Indian	0.46	0.49	0.49	0.45
Multi-Racial	0.19	0.36	0.14	0.29
White	52.76	52.06	52.87	52.26
Unknown	0.04	0.04	0.03	0.05
ELL STATUS				
No	95.26	94.83	95.30	95.03
Yes	4.74	5.17	4.70	4.97
DISABILITY				
No	84.72	84.88	85.37	85.07
Yes	15.28	15.12	14.63	14.93
ACCOMMODATIONS				
No	80.01	79.41	80.31	79.51
Yes	19.99	20.59	19.69	20.49

The student NRC and ethnicity distributions of the FT anchor form samples were compared with the OP samples in Tables 16 through 18. It is apparent that the FT anchor samples represent the OP student population well.

Table 16. Grades 3 and 4 Demographic Statistics for Field Test Anchor Forms

Demographics	2010 Grade 3 FT Anchor Form 1	2010 Grade 3 FT Anchor Form 2	2010 Grade 3 OP Sample	2010 Grade 4 FT Anchor Form 1	2010 Grade 4 FT Anchor Form 2	2010 Grade 4 OP Sample
	%	%	%	%	%	%
NRC SUBGROUPS						
NYC	36.09	34.88	36.09	34.73	34.19	35.87
Big cities	4.48	4.07	4.33	4.03	3.93	4.15
Urban/Suburban	9.74	7.88	8.07	9.80	7.67	8.03
Rural	5.30	5.68	5.91	5.32	5.80	5.82
Average needs	28.33	30.3	29.47	29.37	31.25	29.82
Low needs	13.97	14.89	14.00	14.97	15.31	14.58
Charter	1.81	2.06	2.13	1.52	1.61	1.73
ETHNICITY						
Asian	7.40	7.77	7.69	7.98	8.05	8.10
Black	16.15	17.67	18.94	15.04	17.83	18.96
Hispanics	27.6	21.67	22.24	26.73	20.59	21.53
American Indian	0.40	0.61	0.49	0.70	0.42	0.47
Multi-Racial	0.43	0.61	0.56	0.35	0.59	0.49
White	47.98	51.62	50.02	49.16	52.45	50.38
Unknown	0.04	0.04	0.06	0.04	0.07	0.06

Table 17. Grades 5 and 6 Demographic Statistics for Field Test Anchor Forms

Demographics	2010 Grade 5 FT Anchor Form 1	2010 Grade 5 FT Anchor Form 2	2010 Grade 5 OP Sample	2010 Grade 6 FT Anchor Form 1	2010 Grade 6 FT Anchor Form 2	2010 Grade 6 OP Sample
	%	%	%	%	%	%
NRC SUBGROUPS						
NYC	34.64	33.58	34.78	34.41	32.62	34.59
Big cities	4.17	3.79	4.04	3.45	3.65	3.88
Urban/Suburban	8.62	7.40	7.81	8.00	7.37	7.53
Rural	5.38	6.00	5.91	5.94	5.83	5.91
Average needs	29.39	31.08	30.09	30.55	32.6	31.17
Low needs	15.51	15.84	15.08	15.89	16.00	15.04
Charter	2.00	2.07	2.28	1.44	1.64	1.88
ETHNICITY						
Asian	7.61	7.85	7.66	7.84	7.80	7.51
Black	16.11	18.11	19.09	16.03	18.49	18.99
Hispanics	25.19	20.81	21.23	23.79	19.44	20.92
American Indian	0.50	0.43	0.47	0.35	0.42	0.49
Multi-Racial	0.38	0.36	0.43	0.53	0.33	0.38
White	50.18	52.41	51.07	51.39	53.47	51.66
Unknown	0.03	0.04	0.05	0.08	0.05	0.05

Table 18. Grades 7 and 8 Demographic Statistics for Field Test Anchor Forms

Demographics	2010 Grade 7 FT Anchor Form 1	2010 Grade 7 FT Anchor Form 2	2010 Grade 7 OP Sample	2010 Grade 8 FT Anchor Form 1	2010 Grade 8 FT Anchor Form 2	2010 Grade 8 OP Sample
	%	%	%	%	%	%
NRC SUBGROUPS						
NYC	35.33	32.89	35.11	35.66	33.38	35.39
Big cities	3.45	3.37	3.91	3.33	3.18	3.74
Urban/Suburban	7.67	7.23	7.26	7.53	7.46	7.28
Rural	5.81	6.07	5.99	5.5	6.21	5.98
Average needs	31.02	33.05	31.42	30.91	32.5	31.26
Low needs	15.35	15.55	14.84	15.75	15.87	15.19
Charter	1.18	1.48	1.47	0.93	1.01	1.16
ETHNICITY						
Asian	8.82	7.83	7.59	8.56	8.03	7.60
Black	14.78	17.79	19.00	15.57	17.97	18.82
Hispanics	24.04	19.2	20.47	23.52	18.99	20.53
American Indian	0.49	0.56	0.49	0.37	0.29	0.45
Multi-Racial	0.26	0.25	0.36	0.26	0.44	0.29
White	51.55	54.33	52.06	51.66	54.22	52.26
Unknown	0.07	0.03	0.04	0.06	0.06	0.05

Calibration Process

The IRT model parameters were estimated using CTB/McGraw-Hill's PARDUX software (Burket, 2002). PARDUX estimates parameters simultaneously for MC and CR items using marginal maximum likelihood procedures implemented via the expectation-maximization (EM) algorithm (Bock and Aitkin, 1981; Thissen, 1982). Simulation studies have compared PARDUX with MULTILOG (Thissen, 1991), PARSCALE (Muraki and Bock, 1991), and BIGSTEPS (Wright and Linacre, 1992). PARSCALE, MULTILOG, and BIGSTEPS are among the most widely known and used IRT programs. PARDUX was found to perform at least as well as these other programs (Fitzpatrick, 1990; Fitzpatrick, 1994; Fitzpatrick and Julian, 1996).

The NYSTP ELA Tests did not incur anything problematic during item calibration. The number of estimation cycles was set to 50 for all grades with convergence criterion of 0.001 for all grades. The maximum value of *a*-parameters was set to 3.4, and the range for *b*-parameters was set to be between -7.5 and 7.5. The maximum *c*-parameter value was set to 0.50. These are default parameters that have been used for calibration of NYS test data since its first administration in 1999. The estimated parameters were in the original theta metric, and all the items were well within the prescribed parameter ranges. A number of items on the OP test are set to the default value of the *c*-parameter. When the PARDUX program encounters difficulty estimating the *c*-parameter (guessing), it assigns a default *c*-parameter

value of 0.200. For the Grades 3–8 ELA Tests, all calibration estimation results are reasonable. The summary of calibration results is presented in Table 19.

Table 19. NYSTP ELA 2010 Calibration Results

Grade	Largest *a*-parameter	*b*-parameter/ Gamma Range		# Items with Default *c*-parameter	Theta Mean	Theta Standard Deviation	# Students
3	2.546	-2.778	1.036	10	0.02	1.492	193288
4	2.331	-2.814	2.037	8	-0.03	1.190	196180
5	2.347	-3.310	0.791	10	0.05	1.412	194782
6	2.667	-2.848	1.162	5	-0.06	1.265	194152
7	2.170	-2.594	0.016	11	-0.02	1.364	195403
8	2.234	-3.321	0.581	8	-0.09	1.301	201117

Item-Model Fit

Item fit statistics discern the appropriateness of using an item in the 3PL or 2PPC model. The *Q1* procedure described by Yen (1981) was used to measure fit to the three-parameter model. Students are rank-ordered on the basis of $\hat{\theta}$ values and sorted into ten cells with 10% of the sample in each cell. For each item, the number of students in cell *k* who answered item *i*, N_{ik}, and the number of students in that cell who answered item *i* correctly, R_{ik}, were determined. The observed proportion in cell *k* passing item *i*, O_{ik}, is R_{ik}/N_{ik}. The fit index for item *i* is

$$Q_{1i} = \sum_{k=1}^{10} \frac{N_{ik}(O_{ik} - E_{ik})^2}{E_{ik}(1 - E_{ik})},$$

with

$$E_{ik} = \frac{1}{N_{ik}} \sum_{j\varepsilon \text{ cell } k}^{N_{ik}} P_i(\hat{\theta}_j).$$

A modification of this procedure was used to measure fit to the two-parameter partial credit model. For the two-parameter partial credit model, Q_{1j} was assumed to have approximately a chi-square distribution with the following degrees of freedom:

$$df = I(m_j - 1) - m_j,$$

where

I is the total number of cells (usually 10) and m_j is the possible number of score levels for item j.

To adjust for differences in degrees of freedom among items, Q_1 was transformed to Z_{Q_1}
where

$$Z_{Q_1} = (Q_1 - df)/(2df)^{1/2}.$$

The value of Z increases with sample size, all else being equal. To use this standardized statistic to flag items for potential poor fit, it has been CTB/McGraw-Hill's practice to vary the critical value for Z as a function of sample size. For the OP tests that have large calibration sample sizes, the criterion $Z_{Q_1}Crit$ used to flag items was calculated using the expression

$$Z_{Q_1}Crit = \left(\frac{N}{1500}\right) * 4,$$

where

N is the calibration sample size.

Items were considered to have poor fit if the value of the obtained Z_{Q_1} was greater than the value of Z_{Q_1} critical. If the obtained Z_{Q_1} was less than Z_{Q_1} critical, the items were rated as having acceptable fit. All items in the NYSTP 2010 ELA Tests for Grades 4 and 6 demonstrated good model fit. Items 8 and 14 in Grade 3, item 8 in Grade 5, items 2 and 23 in Grade 7, and items 24, 26, and 28 in Grade 8 exhibited poor item-model fit statistics. The fact that so few items were flagged for poor fit across all NYSTP 2010 ELA Tests further supports the use of the chosen models. Item fit statistics are presented in Tables 20–25.

Table 20. ELA Grade 3 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	Z-observed	Z-critical	Fit OK?
1	3PL	310.29	7	183455	81.06	489.21	Y
2	3PL	246.52	7	183455	64.01	489.21	Y
3	3PL	739.03	7	183455	195.64	489.21	Y
4	3PL	69.05	7	183455	16.58	489.21	Y
5	3PL	855.11	7	183455	226.67	489.21	Y
6	3PL	298.73	7	183455	77.97	489.21	Y
7	3PL	307.74	7	183455	80.38	489.21	Y
8	3PL	2569.24	7	183455	684.79	489.21	N
9	3PL	152.65	7	183455	38.93	489.21	Y
10	3PL	663.07	7	183455	175.34	489.21	Y
11	3PL	465.00	7	183455	122.41	489.21	Y
12	3PL	374.53	7	183455	98.23	489.21	Y
13	3PL	153.99	7	183455	39.28	489.21	Y
14	3PL	2173.71	7	183455	579.08	489.21	N
15	3PL	489.60	7	183455	128.98	489.21	Y
16	3PL	120.61	7	183455	30.36	489.21	Y
17	3PL	288.24	7	183455	75.16	489.21	Y
18	3PL	351.39	7	183455	92.04	489.21	Y
19	3PL	1568.19	7	183455	417.25	489.21	Y

(Continued on next page)

Table 20. ELA Grade 3 Item Fit Statistics (cont.)

Item	Model	Chi Square	DF	Total N	Z-observed	Z-critical	Fit OK?
20	3PL	353.93	7	183455	92.72	489.21	Y
21	2PPC	745.15	17	183455	124.88	489.21	Y
22	3PL	134.30	7	183455	34.02	489.21	Y
23	3PL	544.06	7	183455	143.53	489.21	Y
24	3PL	61.32	7	183455	14.52	489.21	Y
25	3PL	365.38	7	183455	95.78	489.21	Y
26	2PPC	595.36	17	183455	99.19	489.21	Y
27	2PPC	1006.54	17	183455	169.71	489.21	Y
28	2PPC	674.13	26	183455	89.88	489.21	Y

Table 21. ELA Grade 4 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	Z-observed	Z-critical	Fit OK?
1	3PL	341.98	7	194877	89.53	519.67	Y
2	3PL	119.53	7	194877	30.08	519.67	Y
3	3PL	102.34	7	194877	25.48	519.67	Y
4	3PL	746.27	7	194877	197.58	519.67	Y
5	3PL	22.90	7	194877	4.25	519.67	Y
6	3PL	102.94	7	194877	25.64	519.67	Y
7	3PL	384.60	7	194877	100.92	519.67	Y
8	3PL	102.94	7	194877	25.64	519.67	Y
9	3PL	60.70	7	194877	14.35	519.67	Y
10	3PL	143.17	7	194877	36.39	519.67	Y
11	3PL	138.54	7	194877	35.16	519.67	Y
12	3PL	190.31	7	194877	48.99	519.67	Y
13	3PL	323.73	7	194877	84.65	519.67	Y
14	3PL	108.24	7	194877	27.06	519.67	Y
15	3PL	131.09	7	194877	33.16	519.67	Y
16	3PL	133.90	7	194877	33.91	519.67	Y
17	3PL	114.68	7	194877	28.78	519.67	Y
18	3PL	71.69	7	194877	17.29	519.67	Y
19	3PL	265.09	7	194877	68.98	519.67	Y
20	3PL	176.70	7	194877	45.35	519.67	Y
21	3PL	144.15	7	194877	36.66	519.67	Y
22	3PL	76.14	7	194877	18.48	519.67	Y
23	3PL	142.57	7	194877	36.23	519.67	Y
24	3PL	174.03	7	194877	44.64	519.67	Y
25	3PL	79.24	7	194877	19.31	519.67	Y
26	3PL	623.80	7	194877	164.85	519.67	Y
27	3PL	127.45	7	194877	32.19	519.67	Y

(Continued on next page)

Table 21. ELA Grade 4 Item Fit Statistics (cont.)

Item	Model	Chi Square	DF	Total N	Z-observed	Z-critical	Fit OK?
28	3PL	414.44	7	194877	108.89	519.67	Y
29	2PPC	1281.27	35	194877	148.96	519.67	Y
30	2PPC	2503.54	35	194877	295.05	519.67	Y
31	2PPC	914.36	26	194877	123.19	519.67	Y

Table 22. ELA Grade 5 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	Z-observed	Z-critical	Fit OK?
1	3PL	126.55	7	186486	31.95	497.30	Y
2	3PL	69.02	7	186486	16.58	497.30	Y
3	3PL	126.14	7	186486	31.84	497.30	Y
4	3PL	323.02	7	186486	84.46	497.30	Y
5	3PL	141.39	7	186486	35.92	497.30	Y
6	3PL	37.85	7	186486	8.25	497.30	Y
7	3PL	55.84	7	186486	13.05	497.30	Y
8	3PL	2283.95	7	186486	608.54	497.30	N
9	3PL	86.61	7	186486	21.28	497.30	Y
10	3PL	231.49	7	186486	60.00	497.30	Y
11	3PL	300.84	7	186486	78.53	497.30	Y
12	3PL	215.76	7	186486	55.79	497.30	Y
13	3PL	930.06	7	186486	246.70	497.30	Y
14	3PL	390.67	7	186486	102.54	497.30	Y
15	3PL	203.13	7	186486	52.42	497.30	Y
16	3PL	517.91	7	186486	136.55	497.30	Y
17	3PL	477.26	7	186486	125.68	497.30	Y
18	3PL	361.10	7	186486	94.64	497.30	Y
19	3PL	470.25	7	186486	123.81	497.30	Y
20	3PL	398.40	7	186486	104.61	497.30	Y
21	2PPC	857.69	17	186486	144.18	497.30	Y
22	3PL	108.17	7	186486	27.04	497.30	Y
23	3PL	146.90	7	186486	37.39	497.30	Y
24	3PL	61.53	7	186486	14.57	497.30	Y
25	3PL	48.81	7	186486	11.17	497.30	Y
26	2PPC	988.38	17	186486	166.59	497.30	Y
27	2PPC	2582.15	26	186486	354.47	497.30	Y

Table 23. ELA Grade 6 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	Z-observed	Z-critical	Fit OK?
1	3PL	140.31	7	189913	35.63	506.43	Y
2	3PL	50.62	7	189913	11.66	506.43	Y
3	3PL	98.94	7	189913	24.57	506.43	Y
4	3PL	241.29	7	189913	62.62	506.43	Y
5	3PL	37.46	7	189913	8.14	506.43	Y
6	3PL	272.90	7	189913	71.06	506.43	Y
7	3PL	61.97	7	189913	14.69	506.43	Y
8	3PL	304.02	7	189913	79.38	506.43	Y
9	3PL	98.40	7	189913	24.43	506.43	Y
10	3PL	575.28	7	189913	151.88	506.43	Y
11	3PL	110.00	7	189913	27.53	506.43	Y
12	3PL	146.66	7	189913	37.32	506.43	Y
13	3PL	480.02	7	189913	126.42	506.43	Y
14	3PL	162.93	7	189913	41.68	506.43	Y
15	3PL	183.81	7	189913	47.25	506.43	Y
16	3PL	89.07	7	189913	21.93	506.43	Y
17	3PL	146.04	7	189913	37.16	506.43	Y
18	3PL	223.86	7	189913	57.96	506.43	Y
19	3PL	633.14	7	189913	167.34	506.43	Y
20	3PL	294.85	7	189913	76.93	506.43	Y
21	3PL	538.75	7	189913	142.11	506.43	Y
22	3PL	177.70	7	189913	45.62	506.43	Y
23	3PL	169.36	7	189913	43.39	506.43	Y
24	3PL	423.93	7	189913	111.43	506.43	Y
25	3PL	433.41	7	189913	113.96	506.43	Y
26	3PL	199.11	7	189913	51.34	506.43	Y
27	2PPC	2689.88	44	189913	282.05	506.43	Y
28	2PPC	4689.28	44	189913	495.19	506.43	Y
29	2PPC	1026.33	26	189913	138.72	506.43	Y

Table 24. ELA Grade 7 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	Z-observed	Z-critical	Fit OK?
1	3PL	622.99	7	187873	164.63	500.99	Y
2	3PL	2006.38	7	187873	534.36	500.99	N
3	3PL	51.80	7	187873	11.97	500.99	Y
4	3PL	80.76	7	187873	19.71	500.99	Y
5	3PL	111.73	7	187873	27.99	500.99	Y
6	3PL	48.58	7	187873	11.11	500.99	Y
7	3PL	456.56	7	187873	120.15	500.99	Y
8	3PL	31.17	7	187873	6.46	500.99	Y
9	3PL	57.32	7	187873	13.45	500.99	Y
10	3PL	61.53	7	187873	14.58	500.99	Y
11	3PL	203.17	7	187873	52.43	500.99	Y
12	3PL	111.21	7	187873	27.85	500.99	Y
13	3PL	69.16	7	187873	16.61	500.99	Y
14	3PL	128.26	7	187873	32.41	500.99	Y
15	3PL	318.29	7	187873	83.20	500.99	Y
16	3PL	73.28	7	187873	17.71	500.99	Y
17	3PL	184.13	7	187873	47.34	500.99	Y
18	3PL	26.33	7	187873	5.17	500.99	Y
19	3PL	57.66	7	187873	13.54	500.99	Y
20	3PL	222.49	7	187873	57.59	500.99	Y
21	3PL	28.45	7	187873	5.73	500.99	Y
22	3PL	319.04	7	187873	83.40	500.99	Y
23	3PL	2813.41	7	187873	750.05	500.99	N
24	3PL	84.18	7	187873	20.63	500.99	Y
25	3PL	480.57	7	187873	126.57	500.99	Y
26	3PL	108.39	7	187873	27.10	500.99	Y
27	2PPC	1953.08	17	187873	332.04	500.99	Y
28	2PPC	1538.50	17	187873	260.93	500.99	Y
29	3PL	46.54	7	187873	10.57	500.99	Y
30	3PL	182.34	7	187873	46.86	500.99	Y
31	3PL	44.25	7	187873	9.96	500.99	Y
32	3PL	20.39	7	187873	3.58	500.99	Y
33	2PPC	501.32	17	187873	83.06	500.99	Y
34	2PPC	493.14	17	187873	81.66	500.99	Y
35	2PPC	1357.90	26	187873	184.70	500.99	Y

Table 25. ELA Grade 8 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	Z-observed	Z-critical	Fit OK?
1	3PL	178.01	7	196081	45.70	522.88	Y
2	3PL	63.81	7	196081	15.18	522.88	Y
3	3PL	63.49	7	196081	15.10	522.88	Y
4	3PL	1476.37	7	196081	392.71	522.88	Y
5	3PL	175.38	7	196081	45.00	522.88	Y
6	3PL	644.85	7	196081	170.47	522.88	Y
7	3PL	335.63	7	196081	87.83	522.88	Y
8	3PL	70.90	7	196081	17.08	522.88	Y
9	3PL	68.33	7	196081	16.39	522.88	Y
10	3PL	1445.23	7	196081	384.38	522.88	Y
11	3PL	391.24	7	196081	102.69	522.88	Y
12	3PL	88.09	7	196081	21.67	522.88	Y
13	3PL	31.87	7	196081	6.65	522.88	Y
14	3PL	64.31	7	196081	15.32	522.88	Y
15	3PL	34.54	7	196081	7.36	522.88	Y
16	3PL	9.31	7	196081	0.62	522.88	Y
17	3PL	116.18	7	196081	29.18	522.88	Y
18	3PL	136.88	7	196081	34.71	522.88	Y
19	3PL	64.31	7	196081	15.32	522.88	Y
20	3PL	266.21	7	196081	69.28	522.88	Y
21	3PL	133.16	7	196081	33.72	522.88	Y
22	3PL	125.95	7	196081	31.79	522.88	Y
23	3PL	91.73	7	196081	22.64	522.88	Y
24	3PL	3918.17	7	196081	1045.30	522.88	N
25	3PL	390.90	7	196081	102.60	522.88	Y
26	3PL	2388.93	7	196081	636.60	522.88	N
27	2PPC	4858.18	44	196081	513.19	522.88	Y
28	2PPC	5194.15	44	196081	549.01	522.88	N
29	2PPC	1424.97	26	196081	194.00	522.88	Y

Local Independence

In using IRT models, one of the assumptions made is that the items are locally independent. That is, a student's response on one item is not dependent upon his or her response to another item. In other words, when a student's ability is accounted for, his or her response to each item is statistically independent.

One way to measure the statistical independence of items within a test is via the Q_3 statistic (Yen, 1984). This statistic was obtained by correlating differences between students' observed and expected responses for pairs of items after taking into account overall test performance. The Q_3 statistic for binary items was computed as

$$d_{ja} \equiv u_{ja} - P_{j23}\left(\hat{\theta}_a\right)$$

and

$$Q_{3jj'} = r\left(d_j, d_{j'}\right).$$

The generalization to items with multiple response categories uses

$$d_{ja} \equiv x_{ja} - E_{ja},$$

where

$$E_{ja} \equiv E\left(x\middle|\hat{\theta}_a\right) = \sum_{k=1}^{m_j} kP_{jk2}\left(\hat{\theta}_a\right).$$

If a substantial number of items in the test demonstrate local dependence, these items may need to be calibrated separately. All pairs of items with Q_3 values greater than 0.20 were classified as locally dependent. The maximum value for this index is 1.00. When item pairs are flagged by *Q3*, the content of the flagged items is examined to identify possible sources of the local dependence. The primary concern about locally dependent items is that they contribute less psychometric information about examinee proficiency than do locally independent items and they inflate score reliability estimates.

The Q_3 statistics were examined on all ELA Tests, and no items were found to be locally dependent.

Scaling and Equating

The 2010 Grades 3–8 ELA Tests were calibrated and equated to the operational scales, using two separate equating procedures.

In the first equating procedure, the new 2010 OP forms were pre-equated to the corresponding 2009 assessments. Prior to pre-equating, the FT items administered in 2009 were placed onto the OP scales in each grade. The equating of 2009 FT items to the 2009 OP scales was conducted via common examinees. FT items that were eligible for future OP administrations were then included in the NYS item pool. Other items in the NYS item pool were items field tested in 2008, 2007, 2006, and 2005. All items field tested between 2005 and 2008 were also equated to the NYS OP scales. For more details on equating of FT items to the NYS OP scales, refer to *New York State Testing Program 2006: English Language Arts Grades 3–8*, page 56.

At the pre-equating stage, the pool of FT items administered in years 2005, 2006, 2007, 2008, and 2009 was used to select the 2010 OP test forms using the following criteria:

- Content coverage of each form matched the test blueprint
- Psychometric properties of the items:
  - item fit
  - differential item functioning
  - item difficulty
  - item discrimination
  - omit rates

- Test characteristic curve (TCC) and standard error (SE) curve alignment of the 2010 forms with the target 2009 OP forms. (Note that the 2009 OP TCC and SE curves were based on OP parameters and the 2010 TCC and SE curves were based on FT parameters transformed to the OP scale.)

Although it was not possible to entirely avoid including flagged items in OP tests, the number of flagged items included in OP tests was small and content of all flagged items was carefully reviewed.

In the second equating procedure, the 2010 ELA OP data were re-calibrated after the 2010 OP administration. The equating data file included both the OP data and FT anchor forms data, the FT Anchor records were matched to OP test data in two phases: exact match and fuzzy match. An exact match occurs when the school Bedscode (school unique ID) and student ID in both OP and FT data are the same. Fuzzy match includes all the following conditions:

a) at least 10 characters of last name match (including blank spaces)
b) at least 5 characters of first name match (including blank spaces)
c) gender must be the same or one must be blank
d) school Bedscode must be the same or one must be blank
e) 2 of 3 parts of date of birth (MM or DD or YY) must be the same or one must be blank

New OP test equating design was implemented to equate the 2010 OP test in the second OP test equating step. Instead of using FT parameters of MC items contained in the OP test as anchors in test equating, the baseline (2008 administration) year item parameters for items contained in FT anchor forms were used as anchors to transform the 2010 OP item parameters onto the OP scale. Using FT anchor item parameters as anchors in OP test equating helped reduce impact of differential motivation that students might display while responding to OP items versus FT items administered in a stand-alone administration on subsequent student scores. These changes in OP test equating design were endorsed by the NYS Technical Advisory Group.

The MC items contained in the FT anchor sets were representative of the content of the entire test for each grade. The equating was performed using a test characteristic curve (TCC) method (Stocking and Lord, 1983). TCC methods find the linear transformation ($M1$ and $M2$) that transforms the original item parameter estimates (in theta metric) to the scale score metric and minimizes the difference in the relationship between raw scores and ability estimates (i.e., TCC) defined by the FT anchor item parameter estimates from their baseline year 2008 and that relationship defined by the FT anchor item parameter estimates in new administration year 2010. This places the transformed parameters for the OP test items onto the New York State OP scale. In this procedure, new 2010 OP parameter estimates were obtained for all items. For the FT anchor items, the a-parameters and b-parameters were re-estimated within specified constraints (as described in "Calibration Process" sub-section) while c-parameters of anchor items were fixed to their 2008 values.

The relationships between the new and old linear transformation constants that are applied to the original ability metric parameters to place them onto the NYS scale via the Stocking and Lord method are presented below:

$M1 = A * M1_{Anc,}$

$M2 = A * M2_{Anc} + B,$

where

M1 and *M2* are the OP linear transformation constants from the Stocking and Lord (1983) procedure calculated to place the OP test items onto the NYS scale; and $M1_{Anc}$ and $M2_{Anc}$ are the transformation constants previously used to place the FT anchor item parameter estimates onto the NYS scale.

The *A* and *B* values are derived from the input (2008 FT anchor parameter estimates) and estimate (2010 FT anchor parameter estimates) values of anchor items. Anchor input values are known item parameter estimates entered into equating. Anchor estimate values are parameter estimates for the same anchor items re-estimated during the equating procedure. The input and estimate anchor parameter estimates are expected to have similar values.

The *M*1 and *M*2 transformation parameters obtained in the Stocking and Lord equating process were used to transform item parameters obtained in a calibration process onto the final scale score metric. Table 26 presents the 2010 OP transformation constants for New York State Grades 3–8 ELA Tests.

Table 26. NYSTP ELA 2010 Final Transformation Constants

Grade	*M*1	*M*2
3	16.33	665.41
4	24.05	673.92
5	15.45	669.16
6	14.21	663.95
7	16.73	666.34
8	18.93	659.13

Anchor Item Security

In order for an equating to accurately place the items and forms onto the OP scale, it is important to keep the anchor items secure and to reduce anchor item exposure to students and teachers. Although the FT anchor forms were administered in three consecutive years: 2008, 2009, and 2010, they were administered only to small groups of NYS students each year. The FT anchor forms were developed, administered, collected, and scanned by CTB. Given the 'secure' status of these FT anchor forms, there is reason to believe that the item exposure effect was minimal.

Anchor Item Evaluation

Anchor items were evaluated using several procedures. Procedures 1 and 2 evaluate the overall anchor set, while procedure 3 evaluates individual anchor items.

1. Anchor set input and estimates of TCC alignment. The overall alignment of TCCs for the anchor set input and estimates was evaluated to determine the overall stability of anchor item parameters between 2008 and 2010 FT anchor form administrations.

2. Correlations of anchor input and estimates of *a*- and *b*-parameters. Correlations of anchor input and estimate of *a*- and *b*-parameters were evaluated for magnitude. Ideally, the correlations between anchor input and estimate for *a*-parameter should be at least 0.80 and the correlations for *b*-parameters should be at least 0.90. Items contributing to lower than expected correlations were flagged.

3. Iterative linking using Stocking and Lord's TCC method. This procedure, called the TCC method, minimizes the mean squared difference between the two TCCs: one based on 2008 FT anchor estimates and the other on transformed estimates from the 2010 equating of OP test forms. Differential item performance was evaluated by examining previous (input) and transformed (estimated) item parameters. Items with an absolute difference of parameters greater than two times the root mean square deviation were flagged.

In all cases, the overall TCC alignment for anchor set input and estimate was good. Correlations for *b*-parameter input and estimates ranged from 0.88 for Grade 5 to 0.98 for Grade 8. Correlations for *a*-parameter input and estimate ranged from 0.73 for Grade 3 to 0.92 for Grade 5. Correlations between *a*-parameter input and estimates for Grade 3 and correlations between *b*-parameter input and estimates for Grades 5 were slightly below the NYS criterion.

Overall TCC alignment for anchor set input and estimate was very good. In addition, correlations between parameter input and estimates were satisfactory for Grades 3–8. Therefore, despite the fact that a few individual items were flagged, no anchors were removed from any of the anchor sets.

Item Parameters

The OP test item parameters were estimated by the software PARDUX (Burket, 2002) and are presented in Tables 27–32. The parameter estimates are expressed in scale score metric and are defined below:

- *a*-parameter is a discrimination parameter for MC items;
- *b*-parameter is a difficulty parameter for MC items;
- *c*-parameter is a guessing parameter for MC items;
- *alpha* is a discrimination parameter for CR items; and
- *gamma* is a difficulty parameter for category m_j in scale score metric for CR items.

As described in the Section VI "IRT Scaling and Equating," subsection "IRT Models and Rationale for Use," m_j denotes the number of score levels for the *j*-th item, and typically the highest score level is assigned ($m_j - 1$) score points. Note that for the 2PPC model there are $m_j - 1$ independent gammas and one alpha, for a total of m_j independent parameters estimated for each item while there is one *a*- and one *b*-parameter per item in the 3PL model.

Table 27. 2010 Operational Item Parameter Estimates, Grade 3

Item	Max Pts	*a*-par/alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3
1	1	0.052	626.703	0.188	
2	1	0.091	634.261	0.116	
3	1	0.043	625.924	0.200	
4	1	0.051	622.367	0.126	
5	1	0.032	644.398	0.074	
6	1	0.031	628.772	0.188	
7	1	0.057	631.068	0.188	
8	1	0.035	669.587	0.131	
9	1	0.058	632.243	0.281	
10	1	0.046	636.388	0.188	
11	1	0.034	646.176	0.188	
12	1	0.058	644.103	0.260	
13	1	0.064	633.442	0.157	
14	1	0.015	669.800	0.200	
15	1	0.045	650.092	0.154	
16	1	0.059	632.497	0.154	
17	1	0.038	635.809	0.188	
18	1	0.051	646.965	0.145	
19	1	0.038	650.664	0.067	
20	1	0.041	642.774	0.065	
21	2	0.077	49.032	49.406	
22	1	0.042	620.029	0.200	
23	1	0.025	665.712	0.276	
24	1	0.051	620.572	0.200	
25	1	0.055	644.508	0.120	
26	2	0.067	41.374	42.258	
27	2	0.055	35.659	33.502	
28	3	0.047	30.124	29.589	28.557

Table 28. 2010 Operational Item Parameter Estimates, Grade 4

Item	Max Pts	*a*-par/alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3	gamma4
1	1	0.023	622.111	0.200		
2	1	0.042	613.444	0.083		
3	1	0.023	641.518	0.184		
4	1	0.041	609.545	0.227		
5	1	0.029	623.935	0.076		
6	1	0.039	616.263	0.094		
7	1	0.033	606.235	0.227		
8	1	0.032	666.389	0.256		
9	1	0.028	652.064	0.147		
10	1	0.052	632.663	0.203		
11	1	0.044	637.779	0.167		
12	1	0.021	638.683	0.088		
13	1	0.034	626.687	0.227		
14	1	0.036	640.844	0.249		
15	1	0.049	634.893	0.221		
16	1	0.034	649.646	0.178		
17	1	0.030	665.864	0.126		
18	1	0.031	626.089	0.065		
19	1	0.030	705.026	0.402		
20	1	0.037	659.794	0.217		
21	1	0.015	670.677	0.085		
22	1	0.033	636.968	0.133		
23	1	0.030	666.903	0.160		
24	1	0.043	638.160	0.188		
25	1	0.030	624.575	0.214		
26	1	0.036	722.915	0.221		
27	1	0.030	675.114	0.126		
28	1	0.047	659.447	0.234		
29	4	0.049	28.089	30.229	32.572	34.943
30	4	0.060	35.182	36.913	39.241	41.808
31	3	0.051	29.371	32.260	34.527	

Table 29. 2010 Operational Item Parameter Estimates, Grade 5

Item	Max Pts	*a*-par/alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3
1	1	0.054	633.541	0.145	
2	1	0.050	640.725	0.221	
3	1	0.064	631.960	0.044	
4	1	0.080	633.565	0.145	
5	1	0.045	661.175	0.218	
6	1	0.049	638.706	0.102	
7	1	0.045	629.145	0.145	
8	1	0.024	681.391	0.268	
9	1	0.061	634.934	0.039	
10	1	0.056	634.574	0.054	
11	1	0.067	657.357	0.405	
12	1	0.024	634.559	0.200	
13	1	0.069	669.001	0.203	
14	1	0.089	646.605	0.159	
15	1	0.072	644.540	0.172	
16	1	0.033	653.245	0.145	
17	1	0.072	655.254	0.152	
18	1	0.060	633.653	0.145	
19	1	0.056	659.499	0.240	
20	1	0.070	655.898	0.220	
21	2	0.060	38.632	39.432	
22	1	0.055	635.478	0.120	
23	1	0.056	647.672	0.244	
24	1	0.045	630.942	0.145	
25	1	0.046	618.025	0.145	
26	2	0.039	23.929	25.816	
27	3	0.057	36.960	37.553	38.670

Table 30. 2010 Operational Item Parameter Estimates, Grade 6

Item	Max Pts	*a*-par/ alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3	gamma4	gamma5
1	1	0.075	640.553	0.215			
2	1	0.051	647.605	0.125			
3	1	0.049	632.832	0.161			
4	1	0.072	636.683	0.161			
5	1	0.055	626.321	0.077			
6	1	0.053	632.459	0.161			
7	1	0.078	630.906	0.074			
8	1	0.057	623.466	0.161			
9	1	0.073	636.392	0.061			
10	1	0.050	635.926	0.161			
11	1	0.071	641.232	0.201			
12	1	0.058	636.100	0.048			
13	1	0.080	636.733	0.038			
14	1	0.106	639.633	0.208			
15	1	0.062	641.795	0.094			
16	1	0.090	637.383	0.112			
17	1	0.064	644.723	0.207			
18	1	0.109	642.886	0.180			
19	1	0.042	649.060	0.161			
20	1	0.045	674.178	0.200			
21	1	0.075	658.890	0.282			
22	1	0.034	644.242	0.161			
23	1	0.087	638.374	0.302			
24	1	0.088	652.567	0.349			
25	1	0.110	648.099	0.241			
26	1	0.049	663.636	0.186			
27	5	0.087	52.981	54.727	55.810	57.322	58.969
28	5	0.092	56.129	58.543	59.868	61.330	62.811
29	3	0.094	57.182	60.108	62.545		

Table 31. 2010 Operational Item Parameter Estimates, Grade 7

Item	Max Pts	*a*-par/alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3
1	1	0.038	652.624	0.187	
2	1	0.038	639.822	0.187	
3	1	0.049	646.838	0.187	
4	1	0.070	632.679	0.200	
5	1	0.074	636.969	0.255	
6	1	0.057	624.987	0.080	
7	1	0.035	634.252	0.187	
8	1	0.047	640.919	0.098	
9	1	0.054	626.071	0.200	
10	1	0.048	644.400	0.219	
11	1	0.070	622.948	0.064	
12	1	0.044	653.135	0.321	
13	1	0.065	626.740	0.054	
14	1	0.069	624.678	0.076	
15	1	0.058	640.366	0.187	
16	1	0.054	647.164	0.236	
17	1	0.054	629.007	0.056	
18	1	0.047	628.532	0.070	
19	1	0.051	644.526	0.145	
20	1	0.028	641.907	0.117	
21	1	0.045	646.547	0.145	
22	1	0.076	649.256	0.244	
23	1	0.021	650.236	0.200	
24	1	0.057	639.463	0.275	
25	1	0.037	666.607	0.124	
26	1	0.053	632.852	0.073	
27	2	0.063	40.826	41.466	
28	2	0.055	35.067	35.820	
29	1	0.044	650.044	0.281	
30	1	0.054	651.003	0.317	
31	1	0.058	629.447	0.203	
32	1	0.042	627.056	0.221	
33	2	0.060	36.551	39.007	
34	2	0.067	41.500	44.262	
35	3	0.062	39.957	40.392	41.593

Table 32. 2010 Operational Item Parameter Estimates, Grade 8

Item	Max Pts	*a*-par/ alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3	gamma4	gamma5
1	1	0.040	605.667	0.200			
2	1	0.042	611.861	0.209			
3	1	0.046	609.857	0.121			
4	1	0.012	596.240	0.200			
5	1	0.040	611.525	0.209			
6	1	0.037	621.811	0.209			
7	1	0.035	635.476	0.209			
8	1	0.026	644.942	0.297			
9	1	0.045	615.698	0.066			
10	1	0.026	626.288	0.200			
11	1	0.034	610.752	0.200			
12	1	0.037	643.477	0.261			
13	1	0.034	624.689	0.201			
14	1	0.043	618.191	0.081			
15	1	0.037	628.279	0.047			
16	1	0.032	632.616	0.113			
17	1	0.065	623.945	0.253			
18	1	0.024	662.781	0.277			
19	1	0.040	632.610	0.078			
20	1	0.035	628.179	0.050			
21	1	0.043	669.522	0.388			
22	1	0.031	639.352	0.090			
23	1	0.031	644.817	0.248			
24	1	0.021	629.080	0.200			
25	1	0.028	636.509	0.209			
26	1	0.024	620.630	0.200			
27	5	0.092	55.142	56.728	57.940	59.687	61.539
28	5	0.093	55.227	57.458	58.598	60.305	62.259
29	3	0.087	51.759	54.374	57.031		

Test Characteristic Curves

Test characteristic curves (TCCs) provide an overview of the tests in the IRT scale score metric. The 2009 and 2010 TCCs were generated using final OP item parameters for all test items administered in 2009 and 2010. TCCs are the summation of all the item characteristic curves (ICCs) for items that contribute to the OP scale score. Standard error (SE) curves graphically show the amount of measurement error at different ability levels. The 2009 and 2010 TCCs and SE curves are presented in Figures 1–6. Following the adoption of the chain equating method by New York State, the TCCs for new OP test forms are compared to the previous year's TCCs rather than to the baseline 2006 test form TCCs. Therefore, the 2009 OP curves are considered to be target curves for the 2010 OP test TCCs. This equating process enables the comparisons of impact results (i.e., percentages of examinees at and above each proficiency level) between adjacent test administrations. Note that in all figures

the pink TCCs and SE curves represent 2009 OP test and blue TCCs and SE curves represent 2010 OP test. The *x*-axis is the ability scale expressed in scale score metric with the lower and upper bounds established in Year 1 of test administration and presented in the lower corners of the graphs. The y-axis is the proportion of the test that the students can answer correctly.

Figure 1. Grade 3 ELA 2009 and 2010 OP TCCs and SE curves

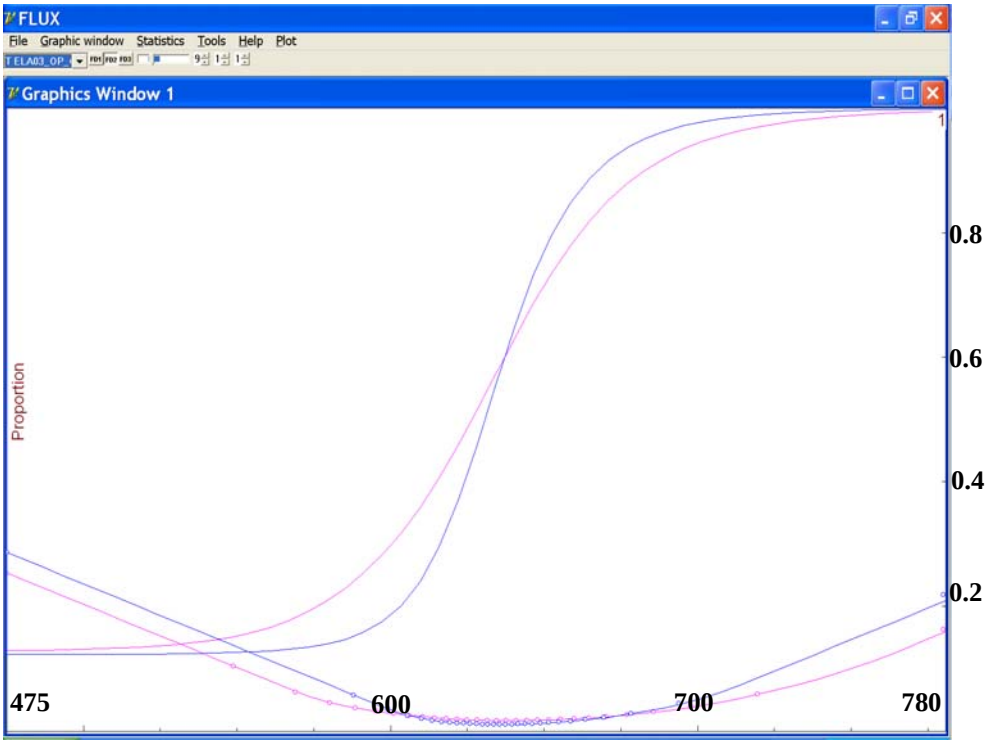


Figure 2. Grade 4 ELA 2009 and 2010 OP TCCs and SE curves

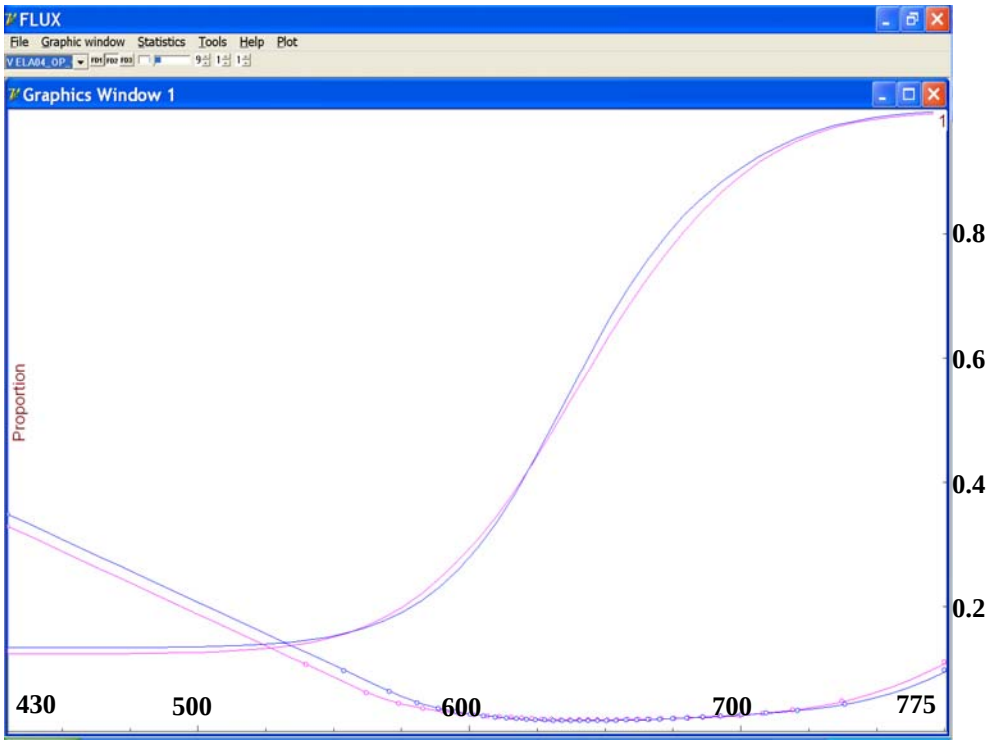


Figure 3. Grade 5 ELA 2009 and 2010 OP TCCs and SE curves

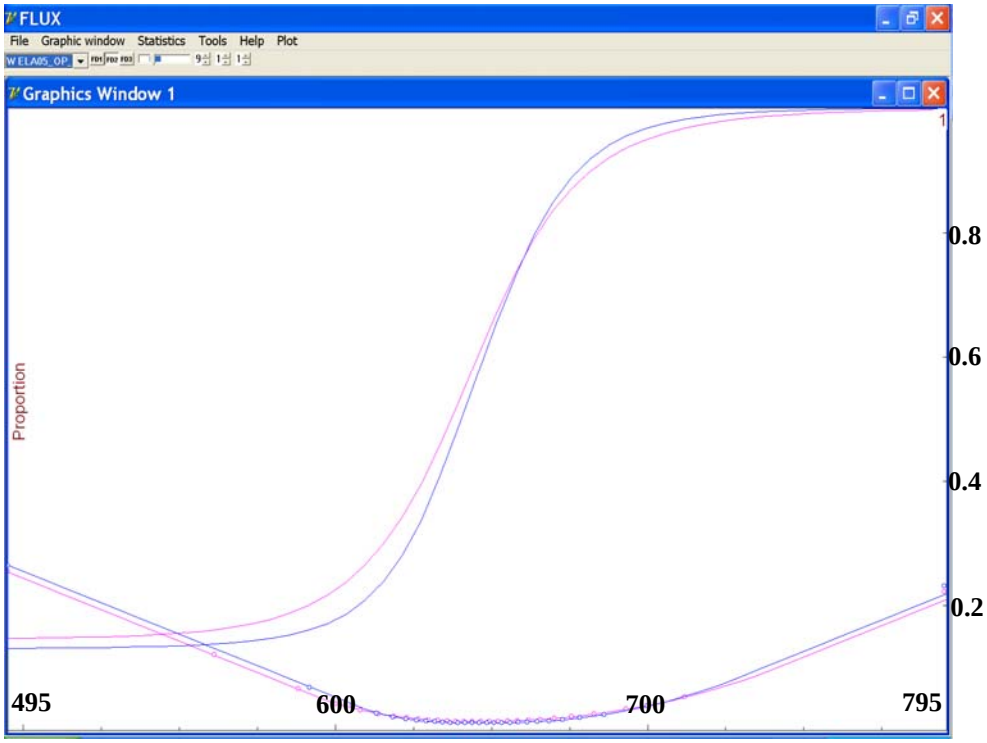


Figure 4. Grade 6 ELA 2009 and 2010 TCCs and SE curves

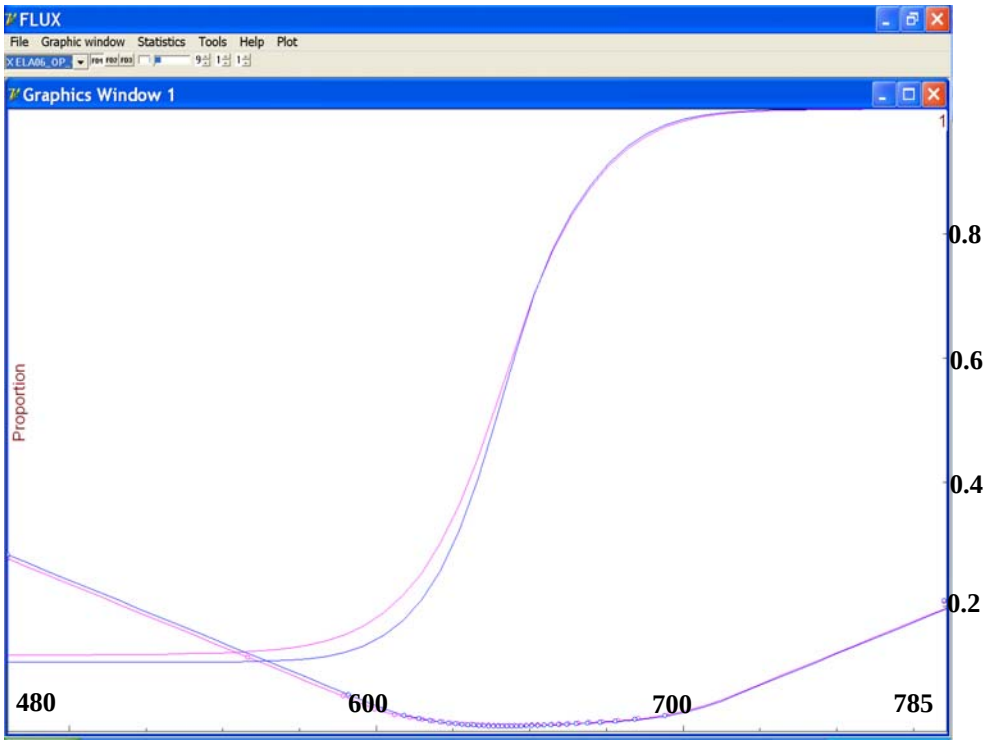


Figure 5. Grade 7 ELA 2009 and 2009 TCCs and SE curves

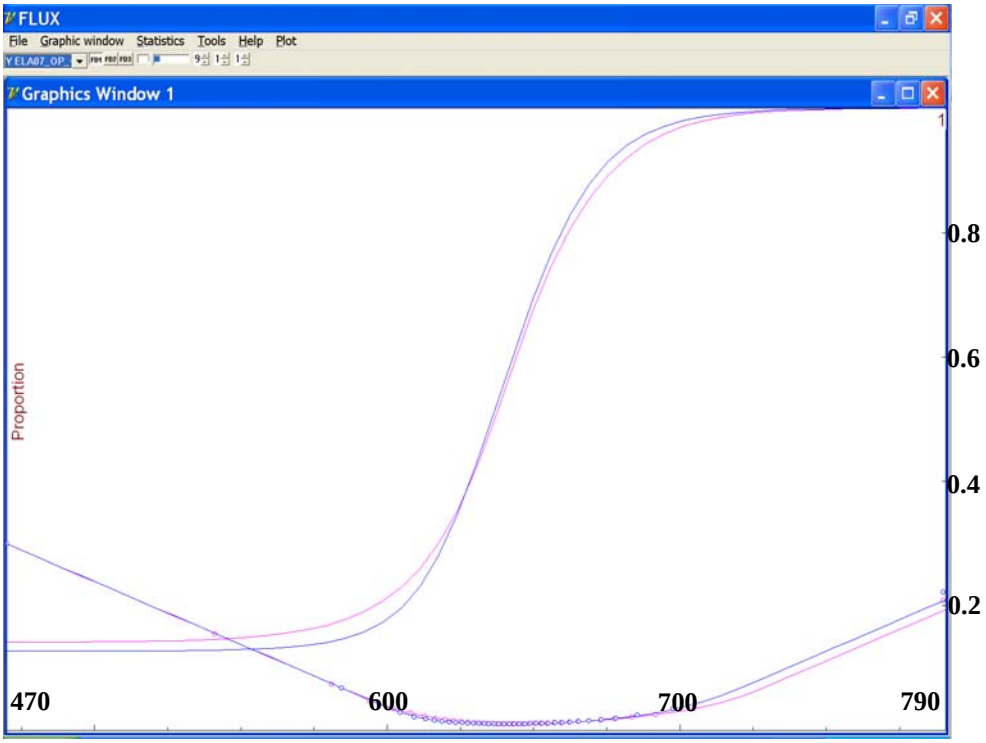
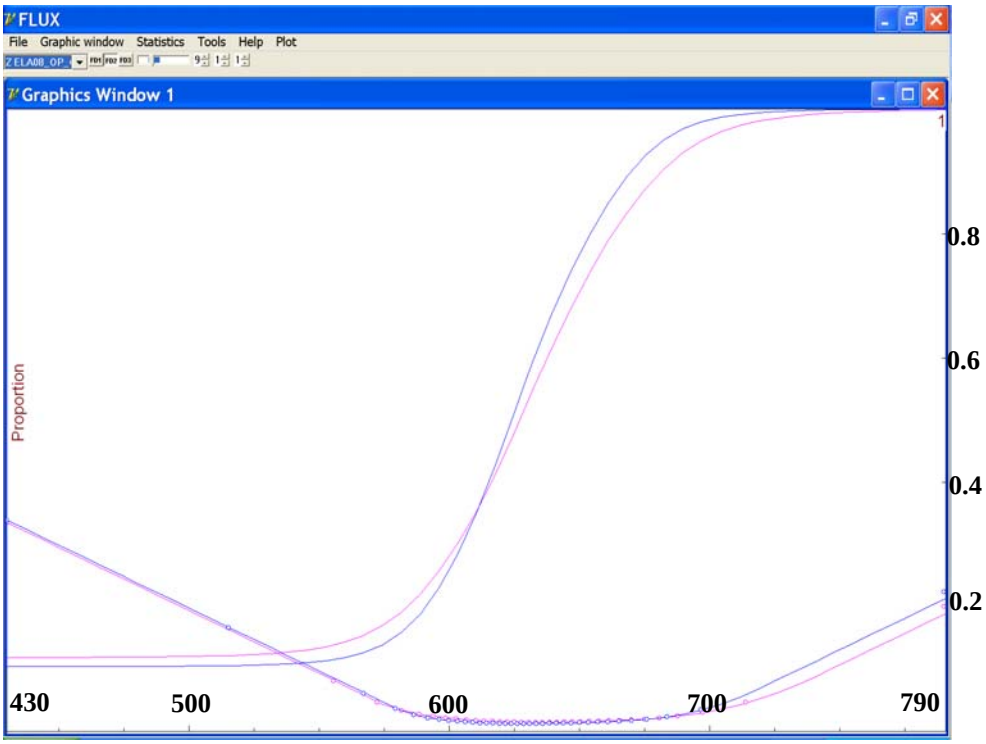


Figure 6. Grade 8 ELA 2009 and 2010 TCCs and SE curves



As seen in Figures 1–6, good alignments of 2009 and 2010 TCCs and SE curves were found for Grades 4, 6, and 7. The TCCs for Grades 3, 5, and 8 were somewhat less well aligned at the lower and upper ends of the scale (indicating that the 2010 form tended to be slightly more difficult for lower-ability students and be slightly easier for the high-ability students). The SE curves were well aligned for all grades. It should be noted that potential differences in test form difficulty at different ability levels are accounted for in the equating and in the resulting raw score-to-scale score conversion tables, so that students of the same ability are expected to obtain the same scale score regardless of which form they took.

Scoring Procedure

New York State students were scored using the number correct (NC) scoring method. This method considers how many score points a student obtained on a test in determining his or her scale score. That is, two students with the same number of score points on the test will receive the same scale score, regardless of which items they answered correctly. In this method, the number correct (or raw) score on the test is converted to a scale score by means of a conversion table. This traditional scoring method is often preferred for its conceptual simplicity and familiarity.

The final item parameters in the scale score metric were used to produce raw score-to-scale score conversion tables for the Grades 3–8 ELA Tests. An inverse TCC method was employed using CTB/McGraw-Hill's proprietary FLUX program. The inverse of the TCC procedure produces trait values based on unweighted raw scores. These estimates show negligible bias for tests with maximum possible raw scores of at least 30 points. All New

York State ELA Tests have a maximum raw score higher than 30 points. In the inverse TCC method, a student's trait estimate is taken to be the trait value that has an expected raw score equal to the student's observed raw score. It was found that for tests containing all MC items, the inverse of the TCC is an excellent first-order approximation to the number correct maximum likelihood estimates (MLE) showing negligible bias for tests of at least 30 items. For tests with a mixture of MC and CR items, the MLE and TCC estimates are even more similar (Yen, 1984).

The inverse of the TCC method relies on the following equation:

$$\sum_{i=1}^{n} v_i x_i = \sum_{i=1}^{n} v_i E(X_i \left| \tilde{\theta} \right.),$$

where

x_i is a student's observed raw score on item *i*.

v_i is a non-optimal weight specified in a scoring process (v_i = 1 if no weights are specified).

$\tilde{\theta}$ is a trait estimate.

Weighting Constructed-Response Items in Grades 4 and 8

Consistent with 2006 scoring procedures, a weight factor of 1.38 was applied to all CR items in Grades 4 and 8. The CR items were weighted in order to align proportions of raw score points obtainable from MC and CR items on 2010 and past ELA Grade 4 and 8 tests. Weighting CR items in Grades 4 and 8 had no substantial effect on the coverage of content standards in the test blueprint.

The inverse TCC scoring method was extended to incorporate weights for CR items for Grades 4 and 8 and weights of 1.38 were specified for these items. It should be noted that when weights are applied, the statistical characteristics of the trait estimates (i.e., bias and standard errors) will depend on the weights that are specified and the statistical characteristics of the items.

Raw Score-to-Scale Score and SEM Conversion Tables

The scale score (SS) is the basic score for the New York State tests. It is used to derive other scores that describe test performance, such as the four performance levels and standards-based performance index scores (SPIs). Number correct raw score-to-scale score conversion tables are presented in this section. Note that the lowest and highest obtainable scale scores for each grade were the same as in 2006 (baseline year).

The standard error (SE) of a scale score indicates the precision with which the ability is estimated, and it inversely is related to the amount of information provided by the test at each ability level. The SE is estimated as follows:

$$SE\left(\hat{\theta}\right) = \frac{1}{\sqrt{I(\theta)}},$$

where

$SE\left(\hat{\theta}\right)$ is the standard error of the scale score (theta), and

$I(\theta)$ is the amount of information provided by the test at a given ability level.

It should be noted that the information is estimated based on thetas in the scale score metric; therefore, the SE is also expressed in the scale score metric. It is also important to note that the SE value varies across ability levels and is the highest at the extreme ends of the scale where the amount of test information is typically the lowest.

Table 33. Grade 3 Raw Score-to-Scale Score (with Standard Error)

Raw Score	Scale Score	Standard Error
0	475	141
1	475	141
2	475	141
3	475	141
4	475	141
5	588	29
6	599	17
7	605	12
8	610	10
9	614	9
10	617	8
11	619	7
12	622	7
13	624	6
14	626	6
15	628	6
16	630	6
17	632	6
18	634	6
19	636	6
20	637	6
21	639	6
22	642	6
23	644	6
24	646	6
25	649	6
26	652	7
27	655	7
28	659	8
29	663	9
30	669	11
31	678	14
32	694	22
33	780	108

Table 34. Grade 4 Raw Score-to-Scale Score (with Standard Error)

Weighted Raw Score	Scale Score	Standard Error
0	430	165
1	430	165
2	430	165
3	430	165
4	430	165
5	430	165
6	549	46
7	565	30
8	576	22
9	583	18
10	590	16
11	595	14
12	600	13
13	604	12
14	608	11
15	612	10
16	616	10
17	619	10
18	622	9
19	625	9
20	628	9
21	631	9
22	634	9
23	637	8
24	640	8
25	643	8
26	646	9
27	649	9
28	652	9
29	656	9
30	659	9
31	663	9
32	667	10
33	671	10
34	675	10
35	679	11
36	685	11
37	690	12
38	696	13
39	704	14
40	712	15
41	722	17
42	738	22
43	775	50

Table 35. Grade 5 Raw Score-to-Scale Score (with Standard Error)

Raw Score	Scale Score	Standard Error
0	495	131
1	495	131
2	495	131
3	495	131
4	495	131
5	592	34
6	606	20
7	613	14
8	618	11
9	623	9
10	626	8
11	629	7
12	632	7
13	634	6
14	637	6
15	639	6
16	642	6
17	644	6
18	646	6
19	648	6
20	651	6
21	653	6
22	656	6
23	659	6
24	661	7
25	665	7
26	668	8
27	673	8
28	678	10
29	686	13
30	700	19
31	795	115

Table 36. Grade 6 Raw Score-to-Scale Score (with Standard Error)

Raw Score	Scale Score	Standard Error
0	480	140
1	480	140
2	480	140
3	480	140
4	480	140
5	591	29
6	603	17
7	609	12
8	614	10
9	618	8
10	621	7
11	623	7
12	626	6
13	628	5
14	630	5
15	632	5
16	633	5
17	635	4
18	637	4
19	638	4
20	640	4
21	641	4
22	643	4
23	644	4
24	646	4
25	647	4
26	649	4
27	651	4
28	653	5
29	655	5
30	657	5
31	659	5
32	662	6
33	665	6
34	669	7
35	673	7
36	678	8
37	684	9
38	694	12
39	785	103

Table 37. Grade 7 Raw Score-to-Scale Score (with Standard Error)

Raw Score	Scale Score	Standard Error
0	470	148
1	470	148
2	470	148
3	470	148
4	470	148
5	470	148
6	584	34
7	597	21
8	604	14
9	609	11
10	613	9
11	616	8
12	619	7
13	621	7
14	623	6
15	626	6
16	628	6
17	629	6
18	631	5
19	633	5
20	635	5
21	637	5
22	638	5
23	640	5
24	642	5
25	643	5
26	645	5
27	647	5
28	649	5
29	651	5
30	653	6
31	655	6
32	657	6
33	660	6
34	662	7
35	665	7
36	669	8
37	673	8
38	678	10
39	685	12
40	698	17
41	790	109

Table 38. Grade 8 Raw Score to Scale Score (with Standard Error)

Weighted Raw Score	Scale Score	Standard Error
0	430	167
1	430	167
2	430	167
3	430	167
4	430	167
5	515	82
6	566	31
7	578	20
8	585	14
9	590	12
10	594	10
11	597	9
12	601	8
13	604	8
14	606	8
15	609	7
16	611	7
17	613	7
18	616	7
19	618	6
20	620	6
21	622	6
22	624	6
23	626	6
24	628	6
25	630	6
26	632	6
27	634	6
28	637	6
29	639	6
30	641	7
31	644	7
32	646	7
33	649	7
34	652	7
35	655	7
36	658	8
37	661	8
38	665	8
39	669	9
40	673	9
41	679	11
42	686	13
43	699	19
44	790	110

Standard Performance Index

The standard performance index (SPI) reported for each objective measured by the Grades 3–8 ELA Tests is an estimate of the percentage of a related set of appropriate items that the student could be expected to answer correctly. An SPI of 75 on an objective measured by a test means, for example, that the student could be expected to respond correctly to 75 out of 100 items that could be considered appropriate measures of that objective. Stated another way, an SPI of 75 indicates that the student would have a 75% chance of responding correctly to any item chosen at random from the hypothetical pool of all possible items that may be used to measure that objective.

Because objectives on all achievement tests are measured by relatively small numbers of items, CTB/McGraw-Hill's scoring system looks not only at how many of those items the student answered correctly, but at additional information as well. In technical terms, the procedure CTB/McGraw-Hill uses to calculate the SPI is based on a combination of item response theory (IRT) and Bayesian methodology. In non-technical terms, the procedure takes into consideration the number of items related to the objective that the student answered correctly, the difficulty level of those items, as well as the student's performance on the rest of the test in which the objective is found. This use of additional information increases the accuracy of the SPI. Details on the SPI derivation procedure are provided in Appendix G.

For the 2010 Grades 3–8 ELA Tests, the performance on objectives was tied to the Level III cut score by computing the SPI target ranges. The expected SPI cuts were computed for the scale scores that are 1 standard error above and 1 standard error below the Level III cut. Table 39 presents the SPI target ranges. The objectives in this table are denoted as follows: 1—Information and Understanding, 2—Literary Response and Expression, and 3—Critical Analysis and Evaluation.

Table 39. SPI Target Ranges

Grade	Objective	# Items	Total Points	Level III Cut SPI Target Range
3	1	9	10	77–91
	2	13	15	84–92
	3	5	5	68–97
4	1	12	12	67–80
	2	14	17	71–82
	3	4	7	73–83
5	1	13	13	85–95
	2	8	8	75–88
	3	5	7	62–82
6	1	14	14	84–93
	2	11	15	77–87
	3	3	7	60–75
7	1	14	16	79–90
	2	13	13	86–94
	3	7	9	71–84
8	1	10	14	73–86
	2	13	13	84–92
	3	5	9	72–84

The SPI is most meaningful in terms of its description of the student's level of skills and knowledge measured by a given objective. The SPI increases the instructional value of test results by breaking down the information provided by the test into smaller, more manageable units. A total test score for a student in Grade 3 who scores below the average on the ELA Test does not provide sufficient information of what specific type of problem the student may be having. On the other hand, this kind of information may be provided by the SPI. For example, evidence that the student has attained an acceptable level of knowledge in the content strand of Information and Understanding but has a low level of knowledge in Literary Response and Expression provides the teacher with a good indication of what type of educational assistance might be most valuable to improve student achievement. Instruction focused on the identified needs of students has the best chance of helping those students increase their skills in the areas measured by the test. SPI reports provide students, parents, and educators the opportunity to identify and target specific areas within the broader content domain for improving student academic performance.

It should be noted that the current NYS test design does not support longitudinal comparison of the SPI scores due to such factors as differences in numbers of items per learning objective from year to year, differences in item difficulties in a given learning objective from year to year, and the fact that the learning objective sub-scores are not equated. The SPI scores are diagnostic scores and are best used at the classroom level to give teachers some insight into their students' strengths and weaknesses.

IRT DIF Statistics

In addition to classical DIF analysis, an IRT-based Linn-Harnisch statistical procedure was used to detect DIF on the Grades 3–8 ELA Tests (Linn and Harnisch, 1981). In this procedure, item parameters (discrimination, location, and guessing) and the scale score (θ) for each examinee were estimated for the 3PL model or the 2PPC model in the case of CR items. The item parameters were based on data from the total population of examinees. Then the population was divided into NRC, gender, or ethnic groups, and the members in each group are sorted into 10 equal score categories (deciles) based upon their location on the scale score (θ) scale. The expected proportion correct for each group, based on the model prediction, is compared to the observed (actual) proportion correct obtained by the group.

The proportion of people in decile *g* who are expected to answer item *i* correctly is

$$P_{ig} = \frac{1}{n_g} \sum_{j \varepsilon g} P_{ij},$$

where

n_g is the number of examinees in decile *g*.

To compute the proportion of students expected to answer item *i* correctly (over all deciles) for a group (e.g., Asian), the formula is

$$P_{i\cdot} = \frac{\sum_{g=1}^{10} n_g P_{ig}}{\sum_{g=1}^{10} n_g}.$$

The corresponding observed proportion correct for examinees in a decile (O_{ig}) is the number of examinees in decile *g* who answered item *i* correctly, divided by the number of students in the decile (n_g). That is,

$$O_{ig} = \frac{\sum_{j \varepsilon g} u_{ij}}{n_g},$$

where

u_{ij} is the dichotomous score for item *i* for examinee *j*.

The corresponding formula to compute the observed proportion answering each item correctly (over all deciles) for a complete ethnic group is

$$O_{i\cdot} = \frac{\sum_{g=1}^{10} n_g O_{ig}}{\sum_{g=1}^{10} n_g}.$$

After the values are calculated for these variables, the difference between the observed proportion correct, for an ethnic group, and expected proportion correct can be computed. The decile group difference (D_{ig}) for observed and expected proportion correctly answering item *i* in decile *g* is

$$D_{ig} = O_{ig} - P_{ig},$$

and the overall group difference (D_i) between observed and expected proportion correct for item *i* in the complete group (over all deciles) is

$$D_{i\cdot} = O_{i\cdot} - P_{i\cdot}.$$

These indices are indicators of the degree to which members of a specific subgroup perform better or worse than expected on each item. Differences for decile groups provide an index for each of the ten regions on the scale score (θ) scale. The decile group difference (D_{ig}) can be either positive or negative. When the difference (D_{ig}) is greater than or equal to 0.100, or less than or equal to -0.100, the item is flagged for potential DIF.

The following groups were analyzed using the IRT-based DIF analysis: Female, Male, Asian, Black, Hispanic, White, High Needs districts (by NRC code), Low Needs districts (by NRC

code), and English language learners. Applying the Linn-Harnisch method revealed that no items were flagged for DIF on the Grades 3, 4, and 8 tests; two items were flagged on the Grades 5, 6, and 7 test, as is shown in Table 40. As indicated in the classical DIF analysis section, items flagged for DIF do not necessarily display bias.

Table 40. Number of Items Flagged for DIF by the Linn-Harnisch Method

Grade	Number of Flagged Items
3	1
4	0
5	1
6	0
7	3
8	3

A detailed list of flagged items including DIF direction and magnitude is presented in Appendix E.

Section VII: Proficiency Level Cut Score Adjustment

This section of the report describes the purpose and methodology of the NYS ELA Grades 3–8 Tests proficiency level cut score adjustment that was conducted after the 2010 OP test administration. Policy decisions that led to changes in the proficiency cut scores were based on two main factors: change in the test administration window between the 2008–2009 and 2009–2010 school years and a decision to align the proficiency standards with Grade 8 student performance on the NYS Regents exam in English.

Proficiency Cut Score Adjustment Process

The NYS ELA scales were maintained between 2009 and 2010 administrations. The 2010 OP tests were equated so that the scale scores from the 2009 and 2010 administrations can be directly compared. That is, a scale score in a given grade level and content area represents the same ability level (comparable knowledge and skills) in 2009 and 2010.

Although the OP test scales did not change, the following steps were taken to set new 2010 cut scores:

1) Grade 8 ELA proficiency Level II score was raised to reflect 75% probability of achieving an ELA Regents score of 65 or above. Grade 8 ELA proficiency Level III score was raised to reflect 75% probability of achieving an ELA Regents score of 75 or above. The alignment of Level II and Level III cut scores with student performance on the Regents exam was conducted by NYSED and the resulting cut scores were provided to CTB. The Grade 8 Level II and Level III cut scores are 625 and 656 respectively. Details on setting Grade 8 Level II and Level III cut scores are available online at http://usny.nysed.gov/scoring_changes/.
2) Grade 8 ELA proficiency Levels II and III cut scores were further adjusted to account for additional instructional time between 2009 and 2010 administration windows, as the 2010 test administration occurred in May instead of January administration in 2009. After the time adjustment the Grade 8 Level II and Level III cut scores are 627 and 658 as shown in Table 42.
3) Grades 3–7 Levels II and III cut scores were established to reflect the corresponding academic rigor applied to the Grade 8 adjusted cut scores by holding the national percentile rank associated with each grade's cut score equal to the national percentile rank associated with the Grade 8 cut score. The national percentile ranks were computed based on NYS student performance on nationally standardized and vertically scaled test items from CTB/McGraw-Hill's *TerraNova* test battery (CTB, 1999, 2000, 2006) that were administered as part of the Secure Anchor/Audit (SAA) test two weeks after OP tests. The percentile ranks for Grade 8 Levels II and III cut scores were 23 and 63 respectively, and the Levels II and III cut scores for the remaining grades were set to correspond to the same percentile ranks. The concordance tables between OP scale scores and *TerraNova* scores were produced to aid the cut score adjustment process and relate the test scores on the NYS scale to the test scores on the *TerraNova* scale. The concordant scores are defined as those having the same percentile rank with respect to the group of students who took the on grade SAA tests. The concordance tables can be found in Appendix I. The national

percentile ranks corresponding to the TerraNova scores are also presented in the concordance tables. Linear interpolation was used to locate the OP cut scores associated with national percentile ranks 23 and 63 if they are not available in the tables.

4) Level IV cut scores for all grades were adjusted only for differences in test administration window between 2008–2009 and 2009–2010 school years.

The above outlined cut score adjustment methodology was endorsed by the NYS Technical Advisory Group and approved by the NYS Board of Regents.

Adjustment of 2009 Cut Scores to Reflect 2010 Administration Window

In order to adjust the 2009 cut scores to reflect the 2010 test administration window, student growth within a school year was estimated using the data from the NYS student performance on the CTB/McGraw-Hill's *TerraNova* Reading items contained in the Secure Anchor/Audit test administered in 2010. An assumption was made that NYS student growth is similar to the growth pattern obtained from a national sample. The estimation was supported by the *TerraNova* norms available for all quarter-months of the school year. Growth between the 17th and 31st quarter-months was computed based on NYS student performance on the *TerraNova* Reading items. The amount of growth on *TerraNova* items was then expressed in standard deviation units and translated back to NYS OP scales. As the last step, the number of scale score points reflecting amount of growth between the two administration windows on the NYS scales was computed and added to the 2009 OP cut scores to derive the time-adjusted cut scores.

The data analysis steps employed in this procedure are described in detail below and the results of each step are presented in Table 41.

1) The 2010 Anchor/Audit item responses were merged at the student level with the 2010 OP data. The NYS ELA OP items and the Reading items in the Anchor/Audit forms were equated to the *TerraNova* Reading scale by using *TerraNova* parameters for Anchor/Audit reading items as anchors in the Stocking and Lord equating method.
2) Item pattern scores were computed for all students who took both the Anchor/Audit forms and OP test, based on their responses to the NYS OP items and Anchor/Audit items.
3) Student scores from step 2 were used to compute mean scale scores on *TerraNova* scale (these scores are presented in column 1).
4) Mean scale scores from step 3 were used to find normative information (national percentile rank) based on the 2007 *TerraNova* national norms. These percentile ranks are presented in column 2 for the quarter-month in which the tests were administered. The NYS ELA Test was administered in the 31st quarter-month of the 2009–2010 school year.
5) *TerraNova* scale scores corresponding to the national percentile rank (from column 2) were found in *TerraNova* norms for the quarter-months in which the NYS ELA Test was administered in 2008–2009 school year. These scores are presented in column 3. The NYS ELA Test was administered in the 17th quarter-month of the 2008–2009 school year.

6) *TerraNova* standard deviations from the nationally representative norming samples (presented in column 4) were used to compute standardized growth (growth in standard deviation units) between the old and new administration windows in the following manner:

$$SG = (TN\_Mean_{\_new} - TN\_Mean_{\_old}) / TN\_SD$$

Standardized growth results are presented in column 5.

7) The standardized growth values (from column 5) were then multiplied by the NYS OP test standard deviations presented in column 6. The resulting values presented in column 7 reflect NYS student growth between the old and new administration windows expressed in scale score metric on NYS ELA scales.

Table 41. Input data for and results of computing NYS student growth in ELA.

	Mean scale scores on *TerraNova* scale from new (2010) administration window ($TN\_Mean_{\_new}$)	National percentile rank (from *TerraNova* norms)	*TerraNova* mean scale scores from old (2009) administration window ($TN\_Mean_{\_old}$)	*TerranNova* standard deviation (TN_SD)	Standardized growth (SG)	NYS Operational Test standard deviation	Growth on NYS scale between old and new administration windows
Column	1	2	3	4	5	6	7
Grade 3	639	62	630	42	0.2143	33.04	7
Grade 4	656	67	650	42	0.1429	29.40	4
Grade 5	670	66	665	42	0.1190	32.07	4
Grade 6	672	61	668	45	0.0889	24.61	2
Grade 7	681	62	678	45	0.0667	31.24	2
Grade 8	687	60	684	47	0.0638	31.00	2

Final 2010 ELA Cut Scores

The resulting 2010 ELA OP proficiency level cut scores are presented in Table 42, columns 4 through 6. The 2009 OP cut scores (in the columns 1–3) are also shown for comparison purposes. The 2010 OP test cut scores were applied to OP test scores for tests administered in the 2009–2010 school year. These cut scores were determined following the procedures outlined in this section of the report.

A "Maximum RS – 1" rule was implemented for a Level IV cut score in cases when it was not possible to adjust this score. The "Maximum RS – 1" rule is used to determine a Level IV

cut score if a perfect test score is required for a student to be classified in proficiency Level IV category. In such situation, a scale score associated with a penultimate raw score (maximum raw score minus 1) is considered a performance Level IV cut score. Information on the cut score adjustment using the "Maximum RS – 1" rule was posted on the NYSED web site at http://www.p12.nysed.gov/irs/ela-math/2008/2008ELAScaleScoretoPerformanceLevels.html. For example, a Level IV cut score for Grade 3 in 2009 was 720. This cut score adjusted for the 2010 OP administration window should be 727 as indicated by the amount of growth on NYS scale between old and new administration windows from Table 41 (column 7). Because there was no scale score of 727 in the 2010 Grade 3 Raw Score-to-Scale Score conversion table and the next higher scale score was the highest obtainable score (780) associated with a perfect raw score, the 2010 Level IV cut score for Grade 3 was set at a penultimate scale score of 694, which associates with a penultimate raw score.

Table 42. NYS 2009 and 2010 ELA proficiency level cut scores

	2009 operational test cut scores			2010 operational test cut scores		
	Proficiency Level			Proficiency Level		
	II	III	IV	II	III	IV
Column	1	2	3	4	5	6
Grade 3	616	650	720	**643**	**662**	**694***
Grade 4	612	650	716	**637**	**668**	**720**
Grade 5	608	650	711	**647**	**666**	**700***
Grade 6	598	650	696	**644**	**662**	**694***
Grade 7	600	650	705	**642**	**664**	**698***
Grade 8	602	650	715	**627**	**658**	**699***

* "Maximum RS – 1" rule was implemented to determine Level IV cut score.

Section VIII: Reliability and Standard Error of Measurement

This section presents specific information on various test reliability statistics (RS) and standard error of measurement (SEM), as well as the results from a study of performance level classification accuracy and consistency. The data set for these studies includes all tested New York State public and charter school students who received valid scores. A study of inter-rater reliability was conducted by a vendor other than CTB/McGraw-Hill and is not included in this *Technical Report*.

Test Reliability

Test reliability is directly related to score stability and standard error and, as such, is an essential element of fairness and validity. Test reliability can be directly measured with an alpha statistic, or the alpha statistic can be used to derive the SEM. For the Grades 3–8 ELA Tests, we calculated two types of reliability statistics: Cronbach's alpha (Cronbach, 1951) and Feldt-Raju coefficient (Qualls, 1995). These two measures are appropriate for assessment of a test's internal consistency when a single test is administered to a group of examinees on one occasion. The reliability of the test is then estimated by considering how well the items that reflect the same construct yield similar results (or how consistent the results are for different items for the same construct measured by the test). Both Cronbach's alpha and Feldt-Raju coefficient measures are appropriate for tests of multiple-item formats (MC and CR items).

Reliability for Total Test

Overall test reliability is a very good indication of each test's internal consistency. Included in Table 43 are the case counts (N-count), number of test items (# Items), Cronbach's alpha and associated SEM, and Feldt-Raju coefficient and associated SEM obtained for the total ELA Tests.

Table 43. ELA 3–8 Tests Reliability and Standard Error of Measurement

Grade	N-count	# Items	# RS points	Cronbach's Alpha	SEM of Cronbach	Feldt-Raju coefficient	SEM of Feldt-Raju
3	196425	28	33	0.85	1.99	0.86	1.95
4	199254	31	39	0.86	2.26	0.87	2.16
5	197200	27	31	0.83	2.06	0.84	1.96
6	197845	29	39	0.87	2.27	0.90	2.04
7	199943	35	41	0.88	2.36	0.89	2.25
8	204080	29	39	0.85	2.44	0.88	2.17

All the coefficients for total test reliability are in the range of 0.83–0.90, which indicates high internal consistency. As expected, the lowest reliabilities were found for the shortest test (i.e., Grade 5), and the highest reliabilities were associated with the longer tests (Grades 4, 6, 7, and 8).

Reliability of MC Items

In addition to overall test reliability, Cronbach's alpha and Feldt-Raju coefficient were computed separately for MC and CR item sets. It is important to recognize that reliability is directly affected by test length; therefore, reliability estimates for tests by item type will always be lower than reliability estimates for the overall test form. Table 44 presents reliabilities for the MC subsets.

Table 44. Reliability and Standard Error of Measurement—MC Items Only

Grade	N-count	# Items	Cronbach's Alpha	SEM of Cronbach	Feldt-Raju	SEM of Feldt-Raju
3	196425	24	0.81	1.68	0.81	1.67
4	199254	28	0.82	1.88	0.83	1.86
5	197200	24	0.81	1.59	0.82	1.57
6	197845	26	0.86	1.61	0.86	1.60
7	199943	30	0.86	1.79	0.86	1.78
8	204080	26	0.80	1.80	0.81	1.79

Reliability of CR Items

Reliability coefficients were also computed for the subsets of CR items. It should be noted that the Grades 3–8 ELA Tests include only three to five CR items, depending on grade level, and the results presented in Table 45 should be interpreted with caution.

Table 45. Reliability and Standard Error of Measurement—CR Items Only

Grade	N-count	# Items	# RS Points	Cronbach's Alpha	SEM of Cronbach	Feldt-Raju	SEM of Feldt-Raju
3	196425	4	9	0.66	0.98	0.67	0.97
4	199254	3	11	0.75	0.98	0.76	0.97
5	197200	3	7	0.51	1.21	0.54	1.17
6	197845	3	13	0.78	1.19	0.81	1.12
7	199943	5	11	0.71	1.40	0.73	1.35
8	204080	3	13	0.83	1.12	0.86	1.04

Note: Results should be interpreted with caution because the number of items is low.

Test Reliability for NCLB Reporting Categories

In this section, reliability coefficients that were estimated for the population and NCLB reporting subgroups are presented. The reporting categories include the following: gender, ethnicity, needs resource code (NRC), English language learners (ELL), all students with disabilities (SWD), all students using test accommodations (SUA), students with disabilities using accommodations falling under 504 Plan (SWD/SUA), and English language learners using accommodations specific to their ELL status (ELL/SUA). Accommodations available to students under the 504 Plan are the following: Flexibility in Scheduling/Timing, Flexibility in Setting, Method of Presentation (excluding Braille), Method of Response, Braille and Large Type, and other. Accommodations available to English language learners

are Time Extension, Separate Location, Third Reading of Listening Selection, and Bilingual Dictionaries and Glossaries.

As shown in Tables 46a–46f, the estimated reliabilities for subgroups were close in magnitude to the test reliability estimates of the population. Cronbach's alpha reliability coefficients were all greater than or equal to 0.80, with the exceptions of Grade 3 Unknown ethnicity group and NRC = Charter schools, Grade 4 NRC = Charter schools, Grade 5 Unknown ethnicity group and NRC = Low Needs districts and Charter schools, and Grade 8 NRC = Low Needs districts and Charter schools. Feldt-Raju reliability coefficients, which tend to be larger than the Cronbach's alpha estimates for the same group, were all larger than or equal to 0.80 with the exceptions of Grade 3 Unknown ethnicity group and NRC = Charter schools, Grade 4 NRC = Charter schools, and Grade 5 Unknown ethnicity group and NRC = Low Needs districts and Charter schools. All other test reliability alpha statistics were in the 0.80–0.90 range, indicating very good test internal consistency (reliability) for analyzed subgroups of examinees.

Table 46a. Grade 3 Test Reliability by Subgroup

Group	Subgroup	N-count	Cronbach's Alpha	SEM of Cronbach	Feldt-Raju	SEM of Feldt-Raju
State	All Students	196425	0.85	1.99	0.86	1.95
Gender	Female	95900	0.83	1.93	0.84	1.89
	Male	100525	0.86	2.05	0.86	2.00
Ethnicity	Asian	15042	0.82	1.79	0.83	1.76
	Black	37101	0.85	2.24	0.86	2.17
	Hispanic	43685	0.85	2.18	0.86	2.13
	American Indian	954	0.86	2.11	0.87	2.06
	Multi-Racial	1107	0.84	1.92	0.85	1.88
	Unknown	121	0.77	1.71	0.78	1.68
	White	98415	0.83	1.82	0.83	1.78
NRC	New York City	69546	0.85	2.12	0.86	2.06
	Big 4 Cites	8364	0.86	2.37	0.86	2.30
	High Needs Urban/Suburban	15386	0.85	2.11	0.86	2.06
	High Needs Rural	11574	0.83	1.98	0.84	1.94
	Average Needs	58701	0.82	1.86	0.83	1.83
	Low Needs	28200	0.80	1.68	0.80	1.66
	Charter	4119	0.76	1.98	0.77	1.95
SWD	All Codes	28021	0.87	2.58	0.88	2.48
SUA	All Codes	47158	0.87	2.45	0.88	2.36
ELL	ELL=Y	18431	0.85	2.43	0.86	2.35
SWD/SUA	SUA=504 plan codes	24502	0.86	2.61	0.87	2.52
ELL/SUA	SUA=ELL codes	16526	0.84	2.40	0.85	2.34

Table 46b. Grade 4 Test Reliability by Subgroup

Group	Subgroup	N-count	Cronbach's Alpha	SEM of Cronbach	Feldt-Raju	SEM of Feldt-Raju
State	All Students	199254	0.86	2.26	0.87	2.16
Gender	Female	97269	0.84	2.22	0.86	2.13
	Male	101985	0.86	2.28	0.88	2.18
Ethnicity	Asian	16193	0.84	2.11	0.86	2.01
	Black	37643	0.84	2.42	0.86	2.33
	Hispanic	42663	0.85	2.39	0.86	2.30
	American Indian	927	0.85	2.35	0.86	2.26
	Multi-Racial	986	0.84	2.19	0.85	2.11
	Unknown	111	0.83	2.01	0.84	1.93
	White	100731	0.83	2.13	0.85	2.03
NRC	New York City	70268	0.86	2.36	0.87	2.25
	Big 4 Cites	8123	0.86	2.50	0.87	2.40
	High Needs Urban/Suburban	15209	0.85	2.34	0.86	2.25
	High Needs Rural	11556	0.85	2.26	0.86	2.17
	Average Needs	60363	0.83	2.15	0.85	2.06
	Low Needs	29726	0.80	2.00	0.81	1.92
	Charter	3449	0.78	2.28	0.79	2.23
SWD	All Codes	29574	0.87	2.57	0.88	2.49
SUA	All Codes	48385	0.86	2.52	0.87	2.43
ELL	ELL=Y	16293	0.84	2.55	0.85	2.47
SWD/SUA	SUA=504 plan codes	26835	0.86	2.59	0.87	2.50
ELL/SUA	SUA=ELL codes	14467	0.84	2.53	0.85	2.46

Table 46c. Grade 5 Test Reliability by Subgroup

Group	Subgroup	N-count	Cronbach's Alpha	SEM of Cronbach	Feldt-Raju	SEM of Feldt-Raju
State	All Students	197200	0.83	2.06	0.84	1.96
Gender	Female	96063	0.82	2.01	0.83	1.92
	Male	101137	0.83	2.10	0.85	2.00
Ethnicity	Asian	15083	0.83	1.86	0.84	1.76
	Black	37742	0.82	2.23	0.83	2.15
	Hispanic	41922	0.83	2.19	0.84	2.11
	American Indian	912	0.81	2.17	0.83	2.09
	Multi-Racial	858	0.80	2.02	0.82	1.93
	Unknown	93	0.78	1.97	0.79	1.89
	White	100590	0.80	1.94	0.82	1.84

(Continued on next page)

Table 46c. Grade 5 Test Reliability by Subgroup (cont.)

Group	Subgroup	N-count	Cronbach's Alpha	SEM of Cronbach	Feldt-Raju	SEM of Feldt-Raju
NRC	New York City	67560	0.83	2.14	0.85	2.04
	Big 4 Cites	7871	0.84	2.31	0.85	2.23
	High Needs Urban/Suburban	14736	0.82	2.16	0.83	2.08
	High Needs Rural	11629	0.82	2.10	0.83	2.01
	Average Needs	60463	0.80	1.98	0.82	1.89
	Low Needs	29773	0.76	1.79	0.78	1.71
	Charter	4583	0.77	2.17	0.79	2.09
SWD	All Codes	30200	0.84	2.39	0.85	2.34
SUA	All Codes	46878	0.84	2.34	0.85	2.28
ELL	ELL=Y	13099	0.83	2.38	0.84	2.34
SWD/SUA	SUA=504 plan codes	27771	0.84	2.4	0.84	2.35
ELL/SUA	SUA=ELL codes	11612	0.83	2.37	0.83	2.33

Table 46d. Grade 6 Test Reliability by Subgroup

Group	Subgroup	N-count	Cronbach's Alpha	SEM of Cronbach	Feldt-Raju	SEM of Feldt-Raju
State	All Students	197845	0.87	2.27	0.90	2.04
Gender	Female	96969	0.86	2.22	0.89	2.00
	Male	100876	0.88	2.30	0.90	2.08
Ethnicity	Asian	14907	0.87	2.13	0.90	1.87
	Black	37941	0.87	2.41	0.88	2.24
	Hispanic	41402	0.87	2.41	0.89	2.23
	American Indian	957	0.87	2.35	0.89	2.14
	Multi-Racial	753	0.87	2.21	0.90	1.98
	Unknown	99	0.83	2.07	0.86	1.89
	White	101786	0.85	2.09	0.88	1.89
NRC	New York City	67488	0.88	2.38	0.90	2.18
	Big 4 Cites	7510	0.88	2.45	0.90	2.27
	High Needs Urban/Suburban	14451	0.87	2.33	0.89	2.16
	High Needs Rural	11632	0.86	2.20	0.89	2.02
	Average Needs	61843	0.85	2.11	0.88	1.92
	Low Needs	30411	0.82	1.93	0.85	1.75
	Charter	3832	0.82	2.27	0.84	2.15

(Continued on next page)

Table 46d. Grade 6 Test Reliability by Subgroup (cont.)

Group	Subgroup	N-count	Cronbach's Alpha	SEM of Cronbach	Feldt-Raju	SEM of Feldt-Raju
SWD	All Codes	30502	0.88	2.60	0.89	2.47
SUA	All Codes	42873	0.88	2.58	0.89	2.44
ELL	ELL=Y	10822	0.86	2.68	0.87	2.56
SWD/SUA	SUA=504 plan codes	27695	0.87	2.61	0.89	2.49
ELL/SUA	SUA=ELL codes	8782	0.86	2.67	0.87	2.55

Table 46e. Grade 7 Test Reliability by Subgroup

Group	Subgroup	N-count	Cronbach's Alpha	SEM of Cronbach	Feldt-Raju	SEM of Feldt-Raju
State	All Students	199943	0.88	2.36	0.89	2.25
Gender	Female	97534	0.87	2.28	0.88	2.17
	Male	102409	0.88	2.43	0.89	2.32
Ethnicity	Asian	15212	0.89	2.17	0.90	2.05
	Black	38306	0.86	2.59	0.87	2.50
	Hispanic	41121	0.88	2.57	0.88	2.49
	American Indian	960	0.87	2.53	0.88	2.44
	Multi-Racial	715	0.85	2.30	0.86	2.19
	Unknown	75	0.81	2.26	0.83	2.17
	White	103554	0.86	2.16	0.87	2.06
NRC	New York City	68297	0.88	2.52	0.89	2.41
	Big 4 Cites	7630	0.88	2.67	0.88	2.59
	High Needs Urban/Suburban	14401	0.87	2.51	0.88	2.42
	High Needs Rural	11857	0.86	2.36	0.87	2.27
	Average Needs	61591	0.85	2.21	0.86	2.11
	Low Needs	32302	0.83	1.97	0.84	1.89
	Charter	2949	0.81	2.47	0.82	2.40
SWD	All Codes	30215	0.87	2.78	0.88	2.73
SUA	All Codes	41110	0.88	2.76	0.88	2.70
ELL	ELL = Y	10374	0.86	2.84	0.87	2.81
SWD/SUA	SUA=504 plan codes	27122	0.87	2.78	0.87	2.74
ELL/SUA	SUA=ELL codes	8335	0.86	2.84	0.87	2.80

Table 46f. Grade 8 Test Reliability by Subgroup

Group	Subgroup	N-count	Cronbach's Alpha	SEM of Cronbach	Feldt-Raju	SEM of Feldt-Raju
State	All Students	204080	0.85	2.44	0.88	2.17
Gender	Female	99450	0.84	2.36	0.87	2.11
	Male	104630	0.85	2.49	0.88	2.22
Ethnicity	Asian	15555	0.87	2.36	0.90	2.02
	Black	38590	0.84	2.60	0.87	2.37
	Hispanic	41976	0.85	2.63	0.88	2.37
	American Indian	921	0.85	2.53	0.88	2.28
	Multi-Racial	590	0.84	2.38	0.87	2.13
	Unknown	94	0.81	2.12	0.84	1.95
	White	106354	0.82	2.21	0.85	2.02
NRC	New York City	70795	0.85	2.60	0.88	2.32
	Big 4 Cites	7513	0.87	2.67	0.89	2.43
	High Needs Urban/Suburban	14341	0.85	2.49	0.88	2.27
NRC	High Needs Rural	11989	0.84	2.35	0.86	2.16
	Average Needs	62845	0.82	2.23	0.85	2.05
	Low Needs	32993	0.79	2.01	0.82	1.87
	Charter	2389	0.78	2.38	0.80	2.24
SWD	All Codes	30580	0.85	2.75	0.87	2.56
SUA	All Codes	41919	0.85	2.78	0.88	2.56
ELL	ELL = Y	10198	0.83	2.86	0.85	2.68
SWD/SUA	SUA=504 plan codes	27741	0.85	2.76	0.87	2.57
ELL/SUA	SUA=ELL codes	8441	0.83	2.86	0.85	2.68

Standard Error of Measurement

The standard error of measurement (SEM), as computed from Cronbach's alpha and the Feldt-Raju reliability statistics, are presented in Table 43. SEMs ranged 1.95–2.44, which is reasonable and small. In other words, the error of measurement from the observed test score ranged from approximately ± 2 to ± 3 raw score points. SEMs are directly related to reliability: the higher the reliability, the lower the standard error. As discussed, the reliability of these tests is relatively high, so it was expected that the SEMs would be very low.

The SEMs for subpopulations, as computed from Cronbach's alpha and the Feldt-Raju reliability statistics, are presented in Tables 46a–46f. The SEMs associated with all reliability estimates for all subpopulations are in the range of 1.66–2.86, which is acceptably close to those for the entire population. This narrow range indicates that across the Grades 3–8 ELA Tests, all students' test scores are reasonably reliable with minimal error.

Performance Level Classification Consistency and Accuracy

This subsection describes the analyses conducted to estimate performance level classification consistency and accuracy for the Grades 3–8 ELA Tests. In other words, this provides statistical information on the classification of students into the four performance categories. Classification consistency refers to the estimated degree of agreement between examinees' performance classification from two independent administrations of the same test (or two parallel forms of the test). Because obtaining test scores from two independent administrations of New York State tests was not feasible due to item release after each administration, a psychometric model was used to obtain the estimated classification consistency indices using test scores from a single administration. Classification accuracy can be defined as the agreement between the actual classifications using observed cut scores and true classifications based on known true cut scores (Livingston and Lewis, 1995).

In conjunction with measures of internal consistency, classification consistency is an important type of reliability and is particularly relevant to high-stakes pass/fail tests. As a form of reliability, classification consistency represents how reliably students can be classified into performance categories.

Classification consistency is most relevant for students whose ability is near the pass/fail cut score. Students whose ability is far above or far below the value established for passing are unlikely to be misclassified because repeated administration of the test will nearly always result in the same classification. Examinees whose true scores are close to the cut score are a more serious concern. These students' true scores will likely lie within the SEM of the cut score. For this reason, the measurement error at the cut scores should be considered when evaluating the classification consistency of a test. Furthermore, the number of students near the cut scores should also be considered when evaluating classification consistency; these numbers show the number of students who are most likely to be misclassified. Scoring tables with SEMs are located in Section VI, "IRT Scaling and Equating," and student scale score frequency distributions are located in Appendix J.

Classification consistency and accuracy were estimated using the IRT procedure suggested by Lee, Hanson, and Brennan (2002) and Wang, Kolen, and Harris (2000). Appendix H includes a description of the calculations and procedure based on the paper by Lee et al. (2002).

Consistency

The results for classifying students into four performance levels are separated from results based solely on the Level III cut. Tables 47 and 48 include case counts (N-count), classification consistency (Agreement), classification inconsistency (Inconsistency), and Cohen's kappa (Kappa). Consistency indicates the rate that a second administration would yield the same performance category designation (or a different designation for the inconsistency rate). The agreement index is a sum of the diagonal element in the contingency table. The inconsistency index is equal to the "1 – agreement index." Kappa is a measure of agreement corrected for chance.

Table 47 depicts the consistency study results based on the range of performance levels for all grades. Overall, between 60% and 72% of students were estimated to be classified

consistently to one of the four performance categories. The coefficient kappa, which indicates the consistency of the placement in the absence of chance, ranged 0.45–0.58.

Table 47. Decision Consistency (All Cuts)

Grade	N-count	Agreement	Inconsistency	Kappa
3	196425	0.5974	0.4026	0.4477
4	199254	0.7160	0.2840	0.5534
5	197200	0.6329	0.3671	0.4813
6	197845	0.7206	0.2794	0.5787
7	199943	0.6881	0.3119	0.5487
8	204080	0.7143	0.2857	0.5693

Table 48 depicts the consistency study results based on two performance levels (passing and not passing) as defined by the Level III cut. Overall, about 81%–85% of the classifications of individual students are estimated to remain stable with a second administration. Kappa coefficients for classification consistency based on one cut ranged 0.62–0.70.

Table 48. Decision Consistency (Level III Cut)

Grade	N-count	Agreement	Inconsistency	Kappa
3	196425	0.8121	0.1879	0.6208
4	199254	0.8408	0.1592	0.6755
5	197200	0.8238	0.1762	0.6460
6	197845	0.8497	0.1503	0.6973
7	199943	0.8485	0.1515	0.6969
8	204080	0.8442	0.1558	0.6884

Accuracy

The results of classification accuracy are presented in Table 49. Included in the table are case counts (N-count) and classification accuracy (Accuracy) for all performance levels (All Cuts) and for the Level III cut score, as well as “false positive” and “false negative” rates for both scenarios. It is always the case that the accuracy of the Level III cut score exceeds the accuracy referring to the entire set of cut scores, because there are only two categories in which the true variable can be located, not four. The accuracy rates indicate that the categorization of a student’s observed performance is in agreement with the location of his or her true ability approximately 67%–79% of the time across all performance levels and approximately 86%–89% of the time in regards to the Level III cut score.

Table 49. Decision Agreement (Accuracy)

Grade	N-count	Accuracy					
		All Cuts	False Positive (All Cuts)	False Negative (All Cuts)	**Level III Cut**	False Positive (Level III Cut)	False Negative (Level III Cut)
3	196425	**0.6737**	0.2364	0.0846	**0.8587**	0.0915	0.0498
4	199254	**0.7894**	0.1425	0.0681	**0.8834**	0.0714	0.0452
5	197200	**0.7124**	0.2087	0.0772	**0.8677**	0.0889	0.0435
6	197845	**0.7850**	0.1544	0.0605	**0.8901**	0.0673	0.0426
7	199943	**0.7534**	0.1855	0.0606	**0.8878**	0.0742	0.0380
8	204080	**0.7806**	0.1606	0.0587	**0.8839**	0.0776	0.0385

Section IX: Summary of Operational Test Results

This section summarizes the distribution of OP scale score results on the New York State 2010 Grades 3–8 ELA Tests. These include the scale score means, standard deviations, percentiles, and performance level distributions for each grade's population and specific subgroups. Gender, ethnic identification, needs resource code (NRC), English language learners (ELL), students with disabilities (SWD), and students using test accommodations (SUA) variables were used to calculate the results of subgroups required for federal reporting and test equity purposes. Especially, the ELL/SUA subgroup is defined as examinees whose ELL status are true and use one or more ELL-related accommodation. The SWD/SUA subgroup includes examinees who are classified with disabilities and use one or more disability-related accommodations. Data include examinees with valid scores from all public and charter schools. Note that complete scale score frequency distribution tables are located in Appendix J.

Scale Score Distribution Summary

Scale score distribution summary tables are presented and discussed in Tables 50–56. In Table 50, scale score statistics for total populations of students from public and charter schools are presented. In Tables 51–56, scale score statistics are presented for selected subgroups in each grade level. Some general observations: Females outperformed Males; Asian and White ethnicities outperformed their peers from other ethnic groups; students from Low Needs and Average Needs districts (as identified by NRC) outperformed students from other districts (New York City, Big 4 Cities, Urban/Suburban, Rural, and Charter); and students with ELL, SWD, and/or SUA achieved below the State aggregate (All Students) in every percentile. This pattern of achievement was consistent across all grades.

Table 50. ELA Grades 3–8 Scale Score Distribution Summary

Grade	N-count	SS Mean	SS Std Dev	10th %tile	25th %tile	50th %tile	75th %tile	90th %tile
3	196425	667.90	33.09	639	652	663	678	694
4	199254	672.82	29.50	637	656	675	690	712
5	197200	672.41	32.09	646	656	668	678	700
6	197845	664.48	24.67	643	653	662	673	684
7	199943	667.91	31.29	640	651	665	678	698
8	204080	659.07	31.11	628	644	658	673	686

Grade 3

Scale score statistics and N-counts of demographic groups for Grade 3 are presented in Table 51. The population scale score mean was 667.90 with a standard deviation of 33.09. By gender subgroup, Females outperformed Males, and the difference was more than four scale score points. Asian, Multi-Racial, and White students' scale score means exceeded the average scale score, as did students from Low Needs and Average Needs districts. Among all ethnic groups, the Asian ethnic group had the highest average scale score mean (674.81). Students from the Big 4 Cities achieved a lower scale score mean than their peers from

schools with other NRC designations and about a half of standard deviation below the population mean. The SWD, SUA, and ELL subgroups scored, on average, approximately a half of one standard deviation below the mean scale score for the population. The SWD/SUA subgroup, which had a scale score mean about 25 scale score units below the population mean, was the lowest performing group analyzed. At the 50th percentile, the following groups exceeded the population score of 663: Asian (669), White (669), Average Needs districts (669), and Low Needs districts (669).

Table 51. Scale Score Distribution Summary, by Subgroup, Grade 3

Demographic Category (Subgroup)		N-count	SS Mean	SS Std Dev	10th %tile	25th %tile	50th %tile	75th %tile	90th %tile
State	All Students	196425	667.90	33.09	639	652	663	678	694
Gender	Female	95900	670.40	33.80	642	652	663	678	694
	Male	100525	665.51	32.22	636	649	663	678	694
Ethnicity	Asian	15042	674.81	35.50	646	655	669	678	694
	Black	37101	658.38	28.31	632	644	655	669	678
	Hispanic	43685	659.72	28.52	634	646	659	669	678
	American Indian	954	662.26	29.85	634	649	659	669	694
	Multi-Racial	1107	670.35	32.93	642	655	663	678	694
	Unknown	121	672.43	23.74	652	659	669	678	694
	White	98415	674.07	34.62	644	655	669	678	694
NRC	New York City	69546	663.37	31.81	636	646	659	669	694
	Big 4 Cites	8364	653.45	27.29	628	639	652	663	678
	High Needs Urban/Suburban	15386	663.50	31.21	636	649	659	669	694
	High Needs Rural	11574	666.64	30.48	639	652	663	678	694
	Average Needs	58701	671.78	32.86	644	655	669	678	694
	Low Needs	28200	679.15	36.23	649	659	669	694	694
	Charter	4119	665.99	27.07	642	652	663	678	694
SWD	All Codes	28021	644.57	27.79	619	632	644	659	669
SUA	All Codes	47158	649.78	27.19	624	636	649	663	678
ELL	ELL=Y	18431	649.71	24.60	626	637	649	659	669
SWD/SUA	SUA=504 plan codes	24502	642.82	26.65	619	630	644	655	669
ELL/SUA	SUA=ELL codes	16526	650.43	23.98	628	639	649	663	669

Grade 4

Scale score statistics and N-counts of demographic groups for Grade 4 are presented in Table 52. The Grade 4 population (All Students) mean was 672.82, with a standard deviation of 29.50. By gender subgroup, Females outperformed Males, but the difference was less than seven scale score points. Asian, Multi-Racial, and White students' scale score means exceeded the average scale score, as did students from Low Needs and Average Needs

districts. Among all ethnic groups, the Asian ethnic group had the highest average scale score mean (684.50). Students from the Big 4 Cities achieved a lower scale score mean than their peers from schools with other NRC designations and about a half of a standard deviation below the population mean. The SWD/SUA subgroup had a scale score mean nearly 30 scale score units below the population mean and was at or below the scale score of any given percentile for any other subgroup. At the 50th percentile, the following groups exceeded the population score of 675: Asian (685), White (679), Average Needs districts (679), and Low Needs districts (685).

Table 52. Scale Score Distribution Summary, by Subgroup, Grade 4

Demographic Category (Subgroup)		N-count	SS Mean	SS Std Dev	10th %tile	25th %tile	50th %tile	75th %tile	90th %tile
State	All Students	199254	672.82	29.50	637	656	675	690	712
Gender	Female	97269	676.05	28.89	643	659	675	690	712
	Male	101985	669.74	29.75	634	652	671	685	704
Ethnicity	Asian	16193	684.50	30.41	649	667	685	704	722
	Black	37643	660.29	27.05	628	643	659	675	690
	Hispanic	42663	662.19	27.56	631	646	663	679	696
	American Indian	927	665.83	28.94	634	649	667	679	696
	Multi-Racial	986	675.25	28.58	643	659	675	690	704
	Unknown	111	685.03	27.70	652	671	685	696	712
	White	100731	680.15	27.90	649	663	679	696	712
NRC	New York City	70268	666.42	29.56	634	649	667	685	704
	Big 4 Cites	8123	657.07	28.68	625	640	656	675	690
	High Needs Urban/Suburban	15209	666.39	28.29	634	649	667	685	696
	High Needs Rural	11556	670.47	28.53	640	656	671	685	704
	Average Needs	60363	678.21	27.20	646	663	679	696	712
	Low Needs	29726	687.25	26.65	659	671	685	704	722
	Charter	3449	666.13	21.58	640	652	667	679	690
SWD	All Codes	29574	644.79	31.00	608	628	646	663	679
SUA	All Codes	48385	651.16	29.87	616	637	652	671	685
ELL	ELL=Y	16293	648.68	27.28	616	634	652	667	679
SWD/SUA	SUA=504 plan codes	26835	643.25	30.67	608	628	646	663	675
ELL/SUA	SUA=ELL codes	14467	649.75	26.41	619	637	652	667	679

Grade 5

Scale score summary statistics for Grade 5 students are in Table 53. Overall, the scale score mean was 672.41, with a standard deviation of 32.09. The difference between mean scale scores by gender groups was about 6 scale score units. Female, Asian, Multi-Racial, and

White students' scale score means exceeded the population mean scale score, as did students from Low Needs and Average Needs districts. The students from the Big 4 Cities scored below their peers from schools with other NRC designations and about two-thirds of standard deviation below the population mean. The SWD, SUA, and ELL subgroups scored approximately two-thirds of a standard deviation below the mean scale score for the population. The SWD/SUA subgroup, which had a scale score mean nearly 23 scale score units below the population mean, was the lowest performing group analyzed. At the 50th percentile, the following groups exceeded the population score of 668: Asian (678), White (673), and Low Needs districts (678).

Table 53. Scale Score Distribution Summary, by Subgroup, Grade 5

Demographic Category (Subgroup)		N-count	SS Mean	SS Std Dev	10th %tile	25th %tile	50th %tile	75th %tile	90th %tile
State	All Students	197200	672.41	32.09	646	656	668	678	700
Gender	Female	96063	675.86	34.06	648	659	668	686	700
	Male	101137	669.13	29.73	644	656	665	678	700
Ethnicity	Asian	15083	685.55	40.67	653	665	678	700	700
	Black	37742	663.02	25.52	639	651	661	673	686
	Hispanic	41922	664.29	26.08	642	651	661	673	686
	American Indian	912	665.14	26.34	644	653	661	673	686
	Multi-Racial	858	674.85	33.19	648	659	668	678	700
	Unknown	93	675.28	27.41	651	661	673	686	700
	White	100590	677.39	33.39	651	661	673	686	700
NRC	New York City	67560	669.55	31.92	644	653	665	678	700
	Big 4 Cites	7871	657.97	24.72	634	646	656	668	678
	High Needs Urban/Suburban	14736	665.91	26.70	644	653	665	673	686
	High Needs Rural	11629	667.26	26.18	644	656	665	678	686
	Average Needs	60463	674.71	31.03	651	659	668	686	700
	Low Needs	29773	684.96	36.82	656	665	678	686	700
	Charter	4583	666.07	24.00	644	653	665	673	686
SWD	All Codes	30200	650.56	22.50	629	639	651	661	673
SUA	All Codes	46878	654.24	23.01	632	644	656	665	678
ELL	ELL=Y	13099	650.66	20.92	629	642	653	661	673
SWD/SUA	SUA=504 plan codes	27771	649.81	22.07	629	639	651	661	673
ELL/SUA	SUA=ELL codes	11612	651.17	20.63	629	642	653	661	673

Grade 6

Scale score summary statistics for Grade 6 students are in Table 54. The scale score mean was 664.48, with a standard deviation of 24.67. Female, Asian, Multi-Racial, and White students' scale score means exceeded the population mean scale score, as did students from Low Needs and Average Needs districts. The students from the Big 4 Cities scored below

their peers from schools with other NRC designations and about two-fifths of a standard deviation below the population mean. The SWD and SUA subgroups scored about four-fifths of a standard deviation below the mean scale score for the population. The ELL subgroup, which had a scale score mean more than 23 scale score units below the population mean, was the lowest performing group analyzed. At the 50th percentile, the following groups exceeded the population score of 662: Female (665), Asian (669), Multi-Racial (665), White (669), Average Needs districts (665), and Low Needs districts (673).

Table 54. Scale Score Distribution Summary, by Subgroup, Grade 6

Demographic Category (Subgroup)		N-count	SS Mean	SS Std Dev	10th %tile	25th %tile	50th %tile	75th %tile	90th %tile
State	All Students	197845	664.48	24.67	643	653	662	673	684
Gender	Female	96969	666.91	25.72	644	653	665	673	684
	Male	100876	662.14	23.38	640	651	662	673	684
Ethnicity	Asian	14907	672.45	30.53	647	657	669	678	694
	Black	37941	655.82	18.28	638	646	655	665	673
	Hispanic	41402	655.87	18.66	637	646	655	665	673
	American Indian	957	660.42	20.71	640	651	659	669	678
	Multi-Racial	753	666.82	28.93	644	655	665	673	684
	Unknown	99	671.18	27.73	649	655	665	678	694
	White	101786	670.05	25.90	649	657	669	678	684
NRC	New York City	67488	658.21	21.01	638	647	657	669	678
	Big 4 Cites	7510	654.93	19.48	635	646	655	665	673
	High Needs Urban/Suburban	14451	659.49	19.81	640	649	659	669	678
	High Needs Rural	11632	663.67	22.24	644	653	662	673	684
	Average Needs	61843	668.59	25.00	647	657	665	678	684
	Low Needs	30411	676.46	29.51	653	662	673	684	694
	Charter	3832	658.47	15.63	643	649	657	665	673
SWD	All Codes	30502	645.37	17.53	628	637	646	655	662
SUA	All Codes	42873	647.23	17.95	628	638	647	657	665
ELL	ELL=Y	10822	641.48	16.99	623	635	643	651	657
SWD/SUA	SUA=504 plan codes	27695	644.78	17.21	626	637	646	655	662
ELL/SUA	SUA=ELL codes	8782	641.88	16.92	623	635	643.5	651	657

Grade 7

Scale score statistics and N-counts of demographic groups for Grade 7 are presented in Table 55. The population scale score mean was 667.91 and the population standard deviation was 31.29. By gender subgroup, Females outperformed Males, the difference was about one-fourth of a standard deviation. Female, Asian, Multi-Racial, and White students' scale score means exceeded the population mean scale score, as did students from Low Needs and

Average Needs districts. Among all ethnic groups the Asian ethnic group had the highest average scale score mean (677.76). The students from the Big 4 Cities scored below their peers from schools with other NRC designations and about a half of a standard deviation below the population mean. The SWD and SUA subgroups scored approximately two-thirds of a standard deviation below the mean scale score for the population. The ELL subgroup, which had a scale score mean more than 29 scale score units below the population mean, was the lowest performing group analyzed. At the 50th percentile, the following groups exceeded the population score of 665: Asian (673), White (669), Average Needs districts (669), and Low Needs districts (673).

Table 55. Scale Score Distribution Summary, by Subgroup, Grade 7

Demographic Category (Subgroup)		N-count	SS Mean	SS Std Dev	10th %tile	25th %tile	50th %tile	75th %tile	90th %tile
State	All Students	199943	667.91	31.29	640	651	665	678	698
Gender	Female	97534	671.77	32.81	643	655	665	678	698
	Male	102409	664.23	29.29	638	649	662	673	685
Ethnicity	Asian	15212	677.76	37.96	645	657	673	685	698
	Black	38306	656.80	22.67	635	645	655	665	678
	Hispanic	41121	657.18	24.05	635	645	657	669	678
	American Indian	960	658.96	23.52	635	647	657	669	685
	Multi-Racial	715	670.10	30.12	643	653	665	678	698
	Unknown	75	673.16	32.06	647	657	669	678	698
	White	103554	674.90	33.03	647	657	669	685	698
NRC	New York City	68297	661.43	28.52	637	647	657	673	685
	Big 4 Cites	7630	653.53	23.09	631	642	653	662	678
	High Needs Urban/Suburban	14401	660.38	25.74	637	647	657	669	685
	High Needs Rural	11857	665.49	27.58	642	651	662	673	685
	Average Needs	61591	672.45	31.06	647	657	669	678	698
	Low Needs	32302	682.20	35.87	653	662	673	685	698
	Charter	2949	661.14	20.27	642	651	660	669	678
SWD	All Codes	30215	644.86	20.42	623	635	645	655	665
SUA	All Codes	41110	646.45	21.28	626	637	647	657	669
ELL	ELL=Y	10374	638.44	20.95	619	629	640	651	657
SWD/SUA	SUA=504 plan codes	41110	646.45	21.28	623	635	645	655	665
ELL/SUA	SUA=ELL codes	8335	638.99	8335	619	629	642	651	660

Grade 8

Scale score statistics and N-counts of demographic groups for Grade 8 are presented in Table 56. The population scale score mean was 659.07 with a standard deviation of 31.11. By gender subgroup, Females outperformed Males, but the difference was less than nine

scale score points. Female, Asian, Multi-Racial, and White students' scale score means exceeded the population mean scale score, as did students from Low Needs and Average Needs districts. The students from the Big 4 Cities scored below their peers from schools with other NRC designations and about a half of a standard deviation below the population mean. The SWD and SUA subgroups scored approximately five-sixths of a standard deviation below the mean scale score for the population. The ELL subgroup, which had a scale score mean more than 35 scale score units below the population mean, was the lowest performing group analyzed. At the 50th percentile, the following groups exceeded the population score of 658: Female (661), Asian (665), White (661), Average Needs districts (661), and Low Needs districts (669).

Table 56. Scale Score Distribution Summary, by Subgroup, Grade 8

Demographic Category (Subgroup)		N-count	SS Mean	SS Std Dev	10th %tile	25th %tile	50th %tile	75th %tile	90th %tile
State	All Students	204080	659.07	31.11	628	644	658	673	686
Gender	Female	99450	663.19	32.19	632	646	661	673	686
	Male	104630	655.15	29.52	626	641	655	669	679
Ethnicity	Asian	15555	668.49	38.29	632	649	665	679	699
	Black	38590	647.13	24.25	622	634	646	658	673
	Hispanic	41976	647.87	25.89	620	634	649	661	673
	American Indian	921	651.19	26.99	624	639	649	665	673
	Multi-Racial	590	662.49	33.22	632	646	658	673	686
	Unknown	94	672.87	36.70	641	649	665	686	699
	White	106354	666.48	31.33	639	649	661	679	699
NRC	New York City	70795	650.94	28.31	622	637	649	665	679
	Big 4 Cites	7513	642.68	26.92	613	630	644	658	669
	High Needs Urban/Suburban	14341	652.42	26.47	626	639	652	665	679
	High Needs Rural	11989	657.61	27.03	630	644	655	669	686
	Average Needs	62845	664.60	30.04	637	649	661	673	686
	Low Needs	32993	674.83	34.34	646	658	669	686	699
	Charter	2389	653.45	19.51	632	641	652	665	673
SWD	All Codes	30580	633.47	23.45	609	622	634	646	658
SUA	All Codes	41919	634.90	24.29	609	622	637	649	661
ELL	ELL = Y	10198	623.62	24.57	597	613	626	639	646
SWD/SUA	SUA=504 plan codes	27741	633.03	22.98	609	622	634	646	658
ELL/SUA	SUA=ELL codes	8441	624.16	24.42	601	613	626	639	649

Performance Level Distribution Summary

Students are classified as Level I (Below Standards), Level II (Meets Basic Standards), Level III (Meets Proficiency Standards), and Level IV (Exceeds Proficiency Standards). The original proficiency cut scores used to distinguish among Levels I, II, III, and IV established during the process of Standard Setting in 2006 were adjusted after the 2010 OP test administration to reflect a change in the test administration window between the 2008–2009 and 2009–2010 school years and the State's policy decision to align the proficiency standards with Grade 8 student performance on the NYS Regents ELA exam. Table 57 shows the ELA cut scores used for classification of students to the four performance level categories in 2010.

Table 57. ELA Grades 3–8 Performance Level Cut Scores

Grade	Level II Cut	Level III Cut	Level IV Cut
3	643	662	694
4	637	668	720
5	647	666	700
6	644	662	694
7	642	664	698
8	627	658	699

Tables 58–64 show the performance level distribution for all examinees from public and charter school with valid scores. Table 58 presents performance level data for total populations of students in Grades 3–8. Tables 59–64 contain performance level data for selected subgroups of students. In general, these distributions reflect the same achievement trends in the scale score summary discussion. More Female students were classified in Level III and above categories as compared to Male students. Similarly more Asian and White students were classified in Level III and above categories as compared to their peers from other ethnic groups. Consistently with the scale score distribution across group pattern, students from Low and Average Needs districts outperformed students from High Needs districts (New York City, Big 4 Cities, Urban/Suburban, and Rural). The Level III and above rates for students in the ELL, SWD, and SUA subgroups were low, compared to the total population of examinees. Across grades, the following subgroups consistently performed above the population average: Asian, White, Average Needs, and Low Needs. Please note that the case counts for the Unknown subgroup are very low and are heavily influenced by very high and/or very low achieving individual students.

Table 58. ELA Grades 3–8 Test Performance Level Distributions

Grade	N-count	Percentage of NYS Student Population in Performance Level				
		Level I	Level II	Level III	Level IV	Levels III & IV
3	196425	13.77	31.47	38.11	16.66	54.77
4	199254	8.34	34.82	50.87	5.97	56.84
5	197200	11.54	35.90	39.71	12.85	52.56
6	197845	11.30	34.44	47.40	6.85	54.26
7	199943	10.35	39.53	38.94	11.18	50.12
8	204080	8.95	39.98	43.37	7.70	51.06

Grade 3

Performance level distributions and N-counts of demographic groups for Grade 3 are presented in Table 59. Statewide, 54.77% of third-graders were Level III or Level IV. 15.91% of Male students were Level I, as compared to only 11.52% of Female students. The percentage of students in Levels III and IV varied widely by ethnicity and NRC subgroups. About 73% of Low Needs district students and about 65% of Asian students were classified in Levels III and IV; whereas the American Indian, Hispanic, Black, Charter, New York City, and/or Big 4 Cities had a range of about 48%–69% of students who were in Level I or Level II. About two-fifths of students with ELL, SWD, or SUA status were in Level I and only about 4% are in Level IV. The following groups had pass rates (percentage of students in Levels III & IV) above the State average: Female, Asian, Multi-Racial, White, Average Needs districts, and Low Needs districts.

Table 59. Performance Level Distribution Summary, by Subgroup, Grade 3

Demographic Category (Subgroup)		N-count	Level I %	Level II %	Level III %	Level IV %	Levels III & IV %
State	All Students	196425	13.77	31.47	38.11	16.66	54.77
Gender	Female	95900	11.52	30.47	39.34	18.67	58.01
	Male	100525	15.91	32.41	36.93	14.74	51.67
Ethnicity	Asian	15042	7.91	27.18	42.35	22.55	64.90
	Black	37101	22.44	38.28	30.25	9.03	39.28
	Hispanic	43685	20.46	38.20	31.84	9.50	41.34
	American Indian	954	17.71	34.70	36.90	10.69	47.59
	Multi-Racial	1107	10.57	29.63	40.56	19.24	59.80
	Unknown	121	5.79	23.97	49.59	20.66	70.25
	White	98415	8.43	26.56	43.17	21.84	65.01
NRC	New York City	69546	18.03	35.35	33.65	12.97	46.62
	Big 4 Cites	8364	30.31	38.79	24.02	6.89	30.91
	High Needs Urban/Suburban	15386	17.35	34.89	34.88	12.89	47.76
	High Needs Rural	11574	12.92	33.50	38.68	14.90	53.59
	Average Needs	58701	9.31	28.78	42.49	19.42	61.91
	Low Needs	28200	5.53	21.85	46.12	26.50	72.62
	Charter	4119	10.44	37.82	38.89	12.84	51.74
SWD	All Codes	28021	46.32	34.47	15.67	3.54	19.22
SUA	All Codes	47158	36.04	37.85	21.12	5.00	26.11
ELL	ELL=Y	18431	33.60	41.99	20.63	3.78	24.41
SWD/SUA	SUA=504 plan codes	24502	49.26	34.20	13.72	2.82	16.55
ELL/SUA	SUA=ELL codes	16526	31.94	43.02	21.14	3.89	25.03

Grade 4

Performance level distributions and N-counts of demographic groups for Grade 4 are presented in Table 60. Across New York, approximately 57% of fourth-grade students were in Levels III and IV. As was seen in Grade 3, the Low Needs subgroup had the highest

percentage of students in Levels III and IV (79.02%), and the SWD/SUA subgroup had the lowest (16.48%). Students in the Black, Hispanic, and American Indian subgroups had percentages classified in Levels III and IV below 45%, which was more than 17% below the other ethnic subgroups. More than twice as many Big 4 City students were in Level I than the population. About a fourth of the students with ELL, SWD, or SUA status were in Level I (over three times the Statewide rate of 8.34%) and fewer than 1% were in Level IV. The following groups had percentages of students classified in Levels III and IV, above the State average: Female, Asian, Multi-Racial, White, Average Needs districts, and Low Needs districts.

Table 60. Performance Level Distribution Summary, by Subgroup, Grade 4

Demographic Category (Subgroup)		N-count	Level I %	Level II %	Level III %	Level IV %	Levels III & IV %
State	All Students	199254	8.34	34.82	50.87	5.97	56.84
Gender	Female	97269	6.52	32.82	53.22	7.44	60.66
	Male	101985	10.08	36.73	48.63	4.56	53.19
Ethnicity	Asian	16193	4.28	23.36	59.53	12.83	72.36
	Black	37643	14.83	48.35	34.99	1.83	36.82
	Hispanic	42663	13.18	46.97	37.63	2.22	39.86
	American Indian	927	10.25	44.88	40.67	4.21	44.88
	Multi-Racial	986	5.17	32.35	56.59	5.88	62.47
	Unknown	111	2.70	20.72	66.67	9.91	76.58
	White	100731	4.54	26.41	61.04	8.01	69.05
NRC	New York City	70268	11.72	42.57	41.31	4.40	45.71
	Big 4 Cites	8123	19.54	47.88	30.59	1.99	32.59
	High Needs Urban/Suburban	15209	10.88	41.73	44.04	3.35	47.39
	High Needs Rural	11556	8.47	37.46	49.60	4.47	54.07
	Average Needs	60363	4.87	28.91	59.35	6.87	66.22
	Low Needs	29726	2.28	18.70	67.53	11.49	79.02
	Charter	3449	7.13	48.48	42.80	1.59	44.39
SWD	All Codes	29574	32.31	49.19	18.02	0.48	18.50
SUA	All Codes	48385	24.48	49.77	24.91	0.84	25.76
ELL	ELL=Y	16293	25.20	54.59	19.95	0.26	20.21
SWD/SUA	SUA=504 plan codes	26835	33.89	49.63	16.15	0.33	16.48
ELL/SUA	SUA=ELL codes	14467	23.52	55.52	20.70	0.27	20.97

Grade 5

Performance level distributions and N-counts of demographic groups for Grade 5 are presented in Table 61. About 83% of the Grade 5 students were in Levels III and IV. As was seen in Grades 3 and 4, the Low Needs subgroup had the highest percentage of students in Levels III and IV (72.91%). Students in the American Indian, Black, and Hispanic subgroups had rates less than 40% of students classified in Levels III and IV, approximately 17% less than other ethnic subgroups. Over two times as many Big 4 City students were in Level I than the population's rate. About 32%–40% of the students with ELL, SWD, or SUA

status were in Level I (approximately three times as many as the Statewide rate of 11.54%), yet only about 20% were in Levels III and IV (combined) and a very low percentage (less than 3%) in Level IV. The following groups had percentages of students classified in Levels III and IV, above the State average: Female, Asian, Multi-Racial, White, Average Needs districts, and Low Needs districts.

Table 61. Performance Level Distribution Summary, by Subgroup, Grade 5

Demographic Category (Subgroup)		N-count	Level I %	Level II %	Level III %	Level IV %	Levels III & IV %
State	All Students	197200	11.54	35.90	39.71	12.85	52.56
Gender	Female	96063	9.33	33.17	41.82	15.68	57.49
	Male	101137	13.64	38.49	37.71	10.17	47.87
Ethnicity	Asian	15083	6.09	23.57	44.99	25.35	70.34
	Black	37742	19.32	44.66	30.11	5.92	36.03
	Hispanic	41922	17.46	43.55	32.29	6.69	38.98
	American Indian	912	14.58	46.05	32.13	7.24	39.36
	Multi-Racial	858	9.09	32.98	43.82	14.10	57.93
	Unknown	93	5.38	32.26	47.31	15.05	62.37
	White	100590	6.97	31.21	45.64	16.18	61.82
NRC	New York City	67560	14.94	38.79	34.96	11.30	46.27
	Big 4 Cites	7871	27.19	44.76	23.95	4.10	28.05
	High Needs Urban/Suburban	14736	15.19	42.70	34.74	7.36	42.11
	High Needs Rural	11629	13.04	41.80	37.50	7.66	45.16
	Average Needs	60463	7.85	34.17	44.31	13.67	57.98
	Low Needs	29773	3.50	23.59	49.96	22.95	72.91
	Charter	4583	13.31	45.08	34.93	6.68	41.61
SWD	All Codes	30200	39.87	43.92	14.64	1.57	16.21
SUA	All Codes	46878	32.36	45.51	19.70	2.43	22.13
ELL	ELL=Y	13099	37.27	47.33	14.32	1.08	15.40
SWD/SUA	SUA=504 plan codes	27771	41.17	43.89	13.64	1.30	14.95
ELL/SUA	SUA=ELL codes	11612	35.77	48.36	14.79	1.08	15.86

Grade 6

Performance level distributions and N-counts of demographic groups for Grade 6 are presented in Table 62. Statewide, 54.26% of Grade 6 students were classified in Levels III and IV. As was seen in other grades, the Low Need subgroup had the most students classified in these two proficiency levels (77.77%), and the ELL, SWD, and SUA subgroups had the fewest. Students in the American Indian, Black, and Hispanic subgroups had around 40% of students classified in Level III and above. Students from Low Needs districts outperformed students in all other subgroups, across demographic categories as in the previous grades. The majority of students with ELL, SWD, and/or SUA status were in Level II, but fewer than 1% were in Level IV. The following groups had percentages of students classified in Levels III and IV, above the State average: Female, Asian, Multi-Racial, White, High Needs Rural, Average Needs districts, and Low Needs districts.

Table 62. Performance Level Distribution Summary, by Subgroup, Grade 6

Demographic Category (Subgroup)		N-count	Level I %	Level II %	Level III %	Level IV %	Levels III & IV %
State	All Students	197845	11.30	34.44	47.40	6.85	54.26
Gender	Female	96969	9.18	32.64	49.67	8.52	58.18
	Male	100876	13.35	36.17	45.22	5.26	50.48
Ethnicity	Asian	14907	6.98	24.40	55.65	12.97	68.63
	Black	37941	18.80	47.00	32.07	2.13	34.20
	Hispanic	41402	19.31	45.26	33.38	2.05	35.43
	American Indian	957	13.90	41.90	40.86	3.34	44.20
	Multi-Racial	753	9.43	31.74	50.33	8.50	58.83
	Unknown	99	4.04	28.28	53.54	14.14	67.68
	White	101786	5.88	26.78	57.65	9.68	67.33
NRC	New York City	67488	17.12	42.67	36.64	3.57	40.21
	Big 4 Cites	7510	22.02	43.91	32.05	2.01	34.06
	High Needs Urban/Suburban	14451	14.50	41.93	40.28	3.29	43.57
	High Needs Rural	11632	9.91	35.63	49.20	5.26	54.46
	Average Needs	61843	6.62	29.07	55.68	8.63	64.31
	Low Needs	30411	3.16	19.07	62.96	14.80	77.77
	Charter	3832	11.66	47.99	38.49	1.85	40.34
SWD	All Codes	30502	41.36	45.26	13.05	0.33	13.37
SUA	All Codes	42873	36.66	46.66	16.09	0.59	16.68
ELL	ELL=Y	10822	51.25	42.40	6.23	0.12	6.35
SWD/SUA	SUA=504 plan codes	27695	42.56	45.26	11.94	0.23	12.18
ELL/SUA	SUA=ELL codes	8782	50.00	43.27	6.62	0.11	6.73

Grade 7

Performance level distributions and N-counts of demographic groups for Grade 7 are presented in Table 63. In Grade 7, 50.12% of the students were in Levels III and IV. Over 11% more Female than Male students were classified in these two proficiency levels. Close to 75% of Big 4 Cities students were in Levels I and II. About 74% of Low Needs students were in Levels III and IV. About 5% of ELL students were in Levels III and IV. The ELL, SWD, and SUA subgroups were well below the performance achievement of the general population, with around 85–95% of those students in Levels I and II. The following subgroups had percentages of students in Levels III and IV, above the general population: Female, Asian, Multi-Racial, White, Average Needs districts, and Low Needs districts.

Table 63. Performance Level Distribution Summary, by Subgroup, Grade 7

Demographic Category (Subgroup)		N-count	Level I %	Level II %	Level III %	Level IV %	Levels III & IV %
State	All Students	199943	10.35	39.53	38.94	11.18	50.12
Gender	Female	97534	7.67	36.34	42.14	13.84	55.98
	Male	102409	12.90	42.57	35.89	8.64	44.53
Ethnicity	Asian	15212	7.08	27.83	45.44	19.65	65.09
	Black	38306	17.21	52.59	26.53	3.67	30.20
	Hispanic	41121	17.55	50.64	27.77	4.04	31.81
	American Indian	960	16.25	49.17	30.21	4.38	34.58
	Multi-Racial	715	6.29	38.04	45.04	10.63	55.66
	Unknown	75	5.33	34.67	45.33	14.67	60.00
	White	103554	5.41	31.94	47.04	15.62	62.66
NRC	New York City	68297	15.30	46.39	31.01	7.30	38.31
	Big 4 Cites	7630	23.68	51.38	21.86	3.08	24.94
	High Needs Urban/Suburban	14401	14.78	48.22	31.25	5.74	37.00
	High Needs Rural	11857	9.62	44.13	37.94	8.30	46.24
	Average Needs	61591	5.75	35.28	45.68	13.29	58.97
	Low Needs	32302	2.73	23.42	52.11	21.74	73.85
	Charter	2949	8.07	55.17	32.93	3.83	36.76
SWD	All Codes	30215	37.49	50.75	11.04	0.71	11.76
SUA	All Codes	41110	34.48	51.32	13.03	1.17	14.20
ELL	ELL=Y	10374	50.54	44.66	4.62	0.18	4.80
SWD/SUA	SUA=504 plan codes	27122	38.30	50.89	10.24	0.57	10.80
ELL/SUA	SUA=ELL codes	8335	48.94	45.76	5.12	0.18	5.30

Grade 8

Performance level distributions and N-counts of demographic groups for Grade 8 are presented in Table 64. In Grade 8, 51.06% of the students were in Levels III and IV. About 12% more Female than Male students were in Levels III or IV. Over 61% of American Indian, Black, and Hispanic students were in Levels I and II. Over 75% of Low Needs students were in Levels III and IV, while fewer than 4% of ELL students were in Levels III and IV. The ELL, SWD, and SUA subgroups were well below the performance achievement of the general population, with over 86% of those students in Levels I and II. The following subgroups had a higher percentage of students in Levels III and IV than the general population: Female, Asian, Multi-Racial, White, Average Needs districts, and Low Needs districts.

Table 64. Performance Level Distribution Summary, by Subgroup, Grade 8

Demographic Category (Subgroup)		N-count	Level I %	Level II %	Level III %	Level IV %	Levels III & IV %
State	All Students	204080	8.95	39.98	43.37	7.70	51.06
Gender	Female	99450	6.84	35.75	47.66	9.75	57.41
	Male	104630	10.97	44.00	39.28	5.75	45.03
Ethnicity	Asian	15555	7.78	26.48	51.23	14.51	65.74
	Black	38590	15.17	54.16	28.40	2.26	30.67
	Hispanic	41976	15.57	51.15	30.61	2.67	33.28
	American Indian	921	12.60	49.29	35.07	3.04	38.11
	Multi-Racial	590	6.27	38.81	45.76	9.15	54.92
	Unknown	94	2.13	35.11	43.62	19.15	62.77
	White	106354	4.25	32.34	52.74	10.68	63.42
NRC	New York City	70795	14.00	48.39	33.23	4.39	37.62
	Big 4 Cites	7513	22.20	52.07	23.79	1.94	25.73
	High Needs Urban/Suburban	14341	11.43	48.19	36.48	3.90	40.38
	High Needs Rural	11989	7.46	43.58	43.42	5.54	48.96
	Average Needs	62845	4.40	34.92	51.42	9.26	60.68
	Low Needs	32993	1.92	22.85	59.02	16.22	75.23
	Charter	2389	6.49	53.16	37.84	2.51	40.35
SWD	All Codes	30580	33.15	55.61	10.89	0.34	11.23
SUA	All Codes	41919	31.63	54.66	13.12	0.59	13.71
ELL	ELL=Y	10198	51.40	45.07	3.47	0.06	3.53
SWD/SUA	SUA=504 plan codes	27741	33.77	55.73	10.24	0.26	10.50
ELL/SUA	SUA=ELL codes	8441	50.16	46.18	3.60	0.06	3.66

Section X: Longitudinal Comparison of Results

This section provides longitudinal comparison of OP scale score results on the New York State 2006–2010 Grades 3–8 ELA Tests. These include the scale score means, standard deviations, and performance level distributions for each grade's public and charter school population. The longitudinal results are presented in Table 65.

Table 65. ELA Grades 3–8 Test Longitudinal Results

Grade	Year	N-Count	Scale Score Mean	Standard Deviation	Percentage of Students in Performance Levels				
					Level I	Level II	Level III	Level IV	Level III & IV
3	2010	196425	667.90	33.09	13.77	31.47	38.11	16.66	54.77
	2009	198123	669.97	35.81	4.75	19.37	65.17	10.72	75.89
	2008	195231	669.00	39.41	5.84	23.92	57.84	12.40	70.24
	2007	198320	666.99	42.23	8.92	23.89	57.29	9.90	67.20
	2006	185533	668.79	40.91	8.53	22.47	61.92	7.07	69.00
4	2010	199254	672.82	29.50	8.34	34.82	50.87	5.97	56.84
	2009	195634	669.93	34.72	4.28	18.76	69.69	7.27	76.96
	2008	196367	666.4	39.90	7.34	21.37	62.85	8.44	71.29
	2007	197306	664.7	39.52	7.79	24.17	59.82	8.22	68.04
	2006	190847	665.73	40.74	8.92	22.40	59.94	8.74	68.68
5	2010	197200	672.41	32.09	11.54	35.90	39.71	12.85	52.56
	2009	197522	675.47	34.58	0.62	17.09	68.72	13.57	82.29
	2008	197318	667.35	30.89	1.78	20.45	71.83	5.94	77.77
	2007	201841	665.39	37.98	4.89	26.88	61.37	6.86	68.24
	2006	201138	662.69	41.17	6.38	26.45	54.86	12.31	67.17
6	2010	197845	664.48	24.67	11.30	34.44	47.40	6.85	54.26
	2009	197674	667.31	27.64	0.13	18.87	71.98	9.02	81.00
	2008	199689	661.45	30.03	1.63	31.20	62.49	4.68	67.17
	2007	204237	661.47	33.98	2.46	34.22	53.93	9.40	63.32
	2006	204104	656.52	40.85	7.28	32.24	48.88	11.60	60.48
7	2010	199943	667.91	31.29	10.35	39.53	38.94	11.18	50.12
	2009	202400	667.19	27.06	0.42	19.15	73.51	6.91	80.42
	2008	205946	662.3	29.29	1.75	27.90	67.79	2.56	70.35
	2007	211545	654.84	38.23	5.90	36.22	51.91	5.98	57.89
	2006	210518	652.29	40.95	8.03	35.55	48.66	7.76	56.42
8	2010	204080	659.07	31.11	8.95	39.98	43.37	7.70	51.06
	2009	207083	661.09	30.82	1.72	29.66	63.75	4.87	68.62
	2008	207646	657.26	37.66	4.95	38.53	50.80	5.73	56.53
	2007	213676	655.39	39.32	6.12	36.75	51.45	5.68	57.13
	2006	212138	650.14	40.78	9.42	41.20	44.53	4.84	49.38

It should be noted, however, that although the ELA scales were maintained between 2009 and 2010 administrations and the scale scores from the 2009 and 2010 administrations can be directly compared, the performance level results between 2009 and 2010 OP tests are not directly comparable because of re-setting the proficiency level cut score values after the 2010 OP test administration.

As seen in Table 65, an increase in scale score means was observed for all ELA grades except Grade 3 between the 2006 and 2010 test administrations. Grade 3 mean scale score dropped 1 scale score point. The least gain was observed for Grades 4 and 6 for which total gain was 7 and 8 scale score points, respectively, between 2006 and 2010 test administrations. The largest gain in scale score points between 2006 and 2010 test administrations was noted for Grades 5 and 7 (10 and 16 scale score points, respectively). Grades 8 gained around 9 scale score points. Relatively steady yearly gain was noticed for Grade 7 with the overall population mean scale score increase of 16 scale points between years 2006 and 2010. For Grades 3 and 4, a slight mean scale score decline (1 to 2 scale score points) was observed between years 2006 and 2007, a small increase (approximately 2 points) was observed between years 2007 and 2008, and again a small increase (approximately 2 points) for Grade 3 and a moderate increase (4 points) for Grade 4 between years 2008 and 2009, a slight mean score decline (2 points) for Grade 3 and a moderate increase (3 points) for Grade 4 between years 2009 and 2010. Relatively steady yearly gain was noticed for Grades 5 and 8 with the overall population mean scale score increase of 13 and 11 scale score points respectively between years 2006 and 2009, and then slight decline (2–3 scale score points) between years 2009 and 2010. For Grade 6, an increase of approximately 5 scale score points was observed between years 2006 and 2007, no score change was noticed between administration years 2007 and 2008, but another 6 scale score points increase was observed between years 2008 and 2009. A moderate mean scale score decline (3 scale score points) was observed between years 2009 and 2010.

The variability of scale score distribution decreased steadily across years for ELA Grade 6. The scale score standard deviation was around 40 scale score points in 2006 and dropped to around 25 scale score points in 2010. For Grades 3 and 4, the variability of scale score distribution decreased in 2009 and 2010. The standard deviations for these grades decreased from about 40 scale score points in 2006, 2007, and 2008 to approximately 35 points in 2009, and then to 33 and 30 scale score points in 2010. The standard deviation for Grade 5 decreased from approximately 40 scale score points in 2006 to about 31 scale score points in 2008 and then increased to approximately 35 scale score points in 2009, and then decreased to 32 scale score points in 2010. The variability of scale score distribution decreased steadily across years for ELA Grades 7 and 8 between years 2006 and 2009. The scale score standard deviation was around 40 scale score points for these grades in 2006 and dropped to around 30 scale score points in 2009 and then increased to approximately 31 scale score points in 2010.

Appendix A—ELA Passage Specifications

General Guidelines

- Each passage must have a clear beginning, middle, and end.
- Passages may be excerpted from larger works, but internal editing must be avoided. No edits may be made to poems.
- Passages should be age- and grade-appropriate and should contain subject matter of interest to the students being tested.
- Informational passages should span a broad range of topics, including history, science, careers, career training, etc.
- Literary passages should span a variety of genres and should include both classic and contemporary literature.
- Material may be selected from books, magazines (such as *Cricket*, *Cobblestone, Odyssey, National Geographic World,* and *Sports Illustrated for Kids),* and newspapers.
- Avoid selecting literature that is widely studied. To that end, do not select passages from basals.
- If the accompanying art is not integral to the passage, and if permissions are granted separately, you may choose not to use that art or to use different art.
- Illustration- or photograph-dependent passages should be avoided whenever possible.
- Passages should bring a range of cultural diversity to the tests. They should be written by, as well as about, people of different cultures and races.
- Passages should be suitable for items to be written that test the performance indicators as outlined in the New York State Learning Standards Core Curricula.
- Passages (excluding poetry) should be analyzed for readability. Readability statistics are useful in helping to determine grade-level appropriateness of text prior to presenting the passages for formal committee review. An overview of readability concept and summary statistics for passages selected for the 2010 OP administration are provided below.

Use of Readability Formulae in New York State Assessments

A variety of readability formulae currently exist that can be used to help determine the readability level of text. The formulae most associated with the K–12 environment are the Dale-Chall, the Fry, and the Spache formulae. Others (such as Flesch-Kincaid) are more associated with general text (such as newspapers and mainstream publications).

Readability formulae provide some useful information about the reading difficulty of a passage or stimulus. However, it should be noted that a readability score is not the most reliable indicator of grade-level appropriateness and, therefore, should not be the sole determinant of whether a particular passage or stimulus should be included in assessment or instructional materials.

Readability formulae are quantitative measures that assess the surface characteristics of text (e.g., the number of letters or syllables in a word, the number of words in a sentence, the number of sentences in a paragraph, the length of the passage). In order to truly measure the

readability of any text, qualitative factors (e.g., density of concepts, organization of text, coherence of ideas, level of student interest, and quality of writing) must also be considered.

One basic drawback to the usability of readability formulae is that not all passage or stimulus formats can be processed. To produce a score, the formulae generally require a minimum of 100 words in a sample (for Flesch Reading Ease and the Flesch-Kincaid, 200-word samples are recommended). This requirement renders the readability formulae essentially unusable for passages such as poems and many functional documents. Another drawback is evident in passages with specialized vocabulary. For example, if a passage contains scientific terminology, the readability score might appear to be above grade-level, even though the terms might be footnoted or explained within the context of the passage.

In light of the drawbacks that exist in the use of readability formulae, rather than relying solely on readability indices, CTB/McGraw-Hill relies on the expertise of the educators in the State of New York to help determine the suitability of passages and stimuli to be used in Statewide assessments. Prospective passages are submitted for review to panels of New York State educators familiar with the abilities of the students to be tested and with the grade-level curricula. The passages are reviewed for readability, appropriateness of content, potential interest level, quality of writing, and other qualitative features that cannot be measured via readability formulae.

Table A1. Readability Summary Information for 2010 Operational Test Passages

Title	Passage Type	Word Count	Avg. Prose Score	Fry Graph	Spache/ Dale-Chall*	Flesch-Kincaid Formula
GRADE 3						
Book 1 (Reading)						
The Tent	Lit-Fiction	160	2.92	1.64	1.87	1.00
Hot Job	Info-Interview	210	4.61	3.82	2.65	3.19
Buffalo Bill and the Pony Express	Lit-Fiction	315	3.16	1.85	3.03	1.26
The Snowman's Gift	Info-Article	230	4.85	4.11	3.55	3.48
Readability Averages			**3.89**	**2.86**	**2.78**	**2.23**
Book 2 (Listening)						
A Fine Day for a Walk	Lit-Fiction	460	2.6	1.32	1.79	1.00

(Continued on next page)

Table A1. Readability Summary Information for 2010 Operational Test Passages (cont.)

Title	Passage Type	Word Count	Avg. Prose Score	Fry Graph	Spache/ Dale-Chall*	Flesch-Kincaid Formula
GRADE 4						
Book 2 (Listening)						
Belly Flops and Gutter Balls	Lit-Fiction	375	4.24	3.02	3.65	2.50
The Dragon Hunter	Info-Article	455	5.33	4.14	3.07	3.46
Naming Our Puppy	Lit-Fiction	465	4.64	3.72	2.98	3.11
About Abigail	Lit-Poem	180	n/a	n/a	n/a	n/a
Where Does the Water Go?	Info-How-to	180	4.59	3.33	3.29	2.75
Book 2 (Listening)						
Mr. Hacker	Lit-Fiction	535	3.82	2.96	2.81	2.32
Book 3 (Reading pair)						
From Tadpole to Frog	Info-Article	360	4.16	3.15	3.30	2.51
Butterfly House	Lit-Fiction	535	4.60	3.51	3.27	2.85
Readability Averages			**4.59**	**3.48**	**3.26**	**2.86**
GRADE 5						
Book 1 (Reading)						
Talking Birds	Info-Article	570	7.11	6.78	5-6	5.68
I've Got Fire!	Lit-Fiction	520	4.81	4.11	5-6	3.39
The Art of Silhouette	Info-Article	560	6.53	6.17	9-10	4.90
Tell Me Again!	Lit-Fiction	465	4.44	3.74	1-4	3.10
Readability Averages			**5.72**	**5.20**	**5-6**	**4.27**
Book 2 (Listening)						
Peach-Basket Ball Game	Info-Article	410	5.02	3.86	5-6	3.24

(Continued on next page)

Table A1. Readability Summary Information for 2010 Operational Test Passages (cont.)

Title	Passage Type	Word Count	Avg. Prose Score	Fry Graph	Spache/ Dale-Chall*	Flesch-Kincaid Formula
GRADE 6						
Book 1 (Reading)						
Zhu Li's Gentle Giant	Info-Article	575	6.76	6.47	7-8	5.29
The Pool Visitor	Lit-Fiction	510	5.95	5.43	7-8	4.44
Olykoeks	Info-Article	375	8.27	7.79	7-8	6.87
The Owl and the Painted Bird	Lit-Fiction	530	6.82	6.74	5-6	5.66
A House of Cards	Info-Article	770	7.38	7.04	9-10	5.97
Book 2 (Listening)						
A Winning Heart	Lit-Fiction	600	6.10	5.28	5-6	4.41
Book 3 (Reading pair)						
Gold Fever	Info-Article	580	7.40	7.31	7-8	6.22
A Gold Miner's Tale	Lit-Poem	290	n/a	n/a	n/a	n/a
Readability Averages			**7.10**	**6.80**	**7-8**	**5.74**

Title	Passage Type	Word Count	Avg. Prose Score	Fry Graph	Spache/ Dale-Chall*	Flesch-Kincaid Formula
GRADE 7						
Book 1 (Reading)						
The Dolphin Mystery	Lit-Fiction	645	6.29	6.17	7-8	4.97
The Amazing Mr. Gilbert	Info-Article	490	7.14	7.02	7-8	5.99
Walking Weather	Lit-Poem	160	n/a	n/a	n/a	n/a
Green Apples	Lit-Fiction	795	5.74	5.61	5-6	4.54
Seen and Heard	Info-Essay	565	9.63	9.06	9-10	9.43
Kids CAN!	Info-Article	650	8.13	7.69	7-8	6.77
Readability Averages			**7.39**	**7.11**	**7-8**	**6.34**
Book 2 (Listening)						
Cooking with the Sun	Info-Article	400	8.77	8.22	7-8	7.69

(Continued on next page)

Table A1. Readability Summary Information for 2010 Operational Test Passages (cont.)

Title	Passage Type	Word Count	Avg. Prose Score	Fry Graph	Spache/ Dale-Chall*	Flesch-Kincaid Formula
GRADE 8						
Book 1 (Reading)						
The Hero	Lit-Fiction	815	6.05	5.54	7-8	4.51
Bindi!	Info-Article	510	9.02	8.57	9-10	8.20
Building Bridges	Lit-Fiction	425	7.81	7.42	7-8	6.99
Wilderness Rivers	Lit-Poem	90	n/a	n/a	n/a	n/a
Video Racing Games	Info-Article	445	11.96	11.92	11-12	10.35
Book 2 (Listening)						
Folly or Fortune?	Info-Article	515	9.42	8.65	9-10	8.08
Book 3 (Reading pair)						
Rufus	Lit-Fiction	400	7.06	6.94	7-8	5.64
The Gift of Reason	Info-Essay	635	7.61	7.29	7-8	6.71
Readability Averages			**8.25**	**7.95**	**7-8**	**7.07**

Table A2. Number, Type, and Length of Passages

Grade	# of Listening Passages	Approximate Word Length	# of Reading Passages	Passage Types	Approximate Word Length	Passage Types
3	8	200–400	20 (includes 5 sets of short paired-passages)	Literary	200–600	50% Literary; 50% Informational
4	5	250–450	20 (includes 8 sets of short paired-passages)	Literary	250–600	50% Literary; 50% Informational
5	12	300–500	20 (includes 5 sets of short paired-passages)	Literary	250–600	50% Literary; 50% Informational

(Continued on next page)

Table A2. Number, Type, and Length of Passages (cont.)

Grade	# of Listening Passages	Approximate Word Length	# of Reading Passages	Passage Types	Approximate Word Length	Passage Types
6	8	350–550	24 (includes 5 sets of short paired-passages)	Informational	300–650	50% Literary; 50% Informational
7	8	400–600	24 (includes 5 sets of short paired-passages)	Informational (May be 2 short paired-pieces)	350–700	50% Literary; 50% Informational
8	5	450–650	20 (includes 8 sets of short paired-passages)	Informational (May be 2 short paired-pieces)	350–800	50% Literary; 50% Informational

Appendix B—Criteria for Item Acceptability

For Multiple-Choice Items:

Check that the content of each item

- is targeted to assess only one objective or skill (unless specifications indicate otherwise)
- deals with material that is important in testing the targeted performance indicator
- uses grade-appropriate content and thinking skills
- is presented at a reading level suitable for the grade level being tested
- has a stem that facilitates answering the question or completing the statement without looking at the answer choices
- has a stem that does not present clues to the correct answer choice
- has answer choices that are plausible and attractive to the student who has not mastered the objective or skill
- has mutually exclusive distractors
- has one and only one correct answer choice
- is free of cultural, racial, ethnic, age, gender, disability, regional, or other apparent bias

Check that the format of each item

- is worded in the positive unless it is absolutely necessary to use the negative form
- is free of extraneous words or expressions in both the stem and the answer choices (e.g., the same word or phrase does not begin each answer choice)
- indicates emphasis on key words, such as best, first, least, not, and others, that are important and might be overlooked
- places the interrogative word at the beginning of a stem in the form of a question or places the omitted portion of an incomplete statement at the end of the statement
- indicates the correct answer choice
- provides the rationale for all distractors
- is conceptually, grammatically, and syntactically consistent—between the stem and answer choices, and among the answer choices
- has answer choices balanced in length or contains two long and two short choices
- clearly identifies the passage or other stimulus material associated with the item
- clearly identifies a need of art, if applicable, and the art is conceptualized and sketched, with important considerations explicated

Also check that

- one item does not present clues to the correct answer choice for any other item
- any item based on a passage is answerable from the information given in the passage and is not dependent on skills related to other content areas
- any item based on a passage is truly passage-dependent; that is, not answerable without reference to the passage
- there is a balance of reasonable, non-stereotypic representation of economic classes, races, cultures, ages, genders, and persons with disabilities in context and art

For Constructed-Response Items:

Check that the content of each item is

- designed to assess the targeted performance indicator
- appropriate for the grade level being tested
- presented at a reading level suitable for the grade level being tested
- appropriate in context
- written so that a student possessing knowledge or skill being tested can construct a response that can be scored with the specified rubric or scoring tool; that is, the range of possible correct responses must be wide enough to allow for diversity of responses, but narrow enough so that students who do not clearly show their grasp of the objective or skill being assessed cannot obtain the maximum score
- presented without clue to the correct response
- checked for accuracy and documented against reliable, up-to-date sources (including rubrics)
- free of cultural, racial, ethnic, age, gender, disability, or other apparent bias

Check that the format of each item is

- appropriate for the question being asked and the intended response
- worded clearly and concisely, using simple vocabulary and sentence structure
- precise and unambiguous in its directions for the desired response
- free of extraneous words or expressions
- worded in the positive rather than in the negative form
- conceptually, grammatically, and syntactically consistent
- marked with emphasis on key words, such as best, first, least, and others, that are important and might be overlooked
- clearly identified as needing art, if applicable, and the art is conceptualized and sketched, with important considerations explicated

Also check that

- one item does not present clues to the correct response to any other item
- there is balance of reasonable, non-stereotypic representation of economic classes, races, cultures, ages, genders, and persons with disabilities in context and art
- for each set of items related to a reading passage, each item is designed to elicit a unique and independent response
- items designed to assess reading do not depend on prior knowledge of the subject matter used in the prompt/question

Appendix C—Psychometric Guidelines for Operational Item Selection

It is primarily up to the content development department to select items for the 2010 OP test. Research will provide support, as necessary, and will review the final item selection. Research will provide data files with parameters for all FT items eligible for item pool. The pools of items eligible for 2010 item selection will include 2005, 2006, 2007, and 2009 FT items. All items for each grade will be on the same (grade specific) scale.

Here are general guidelines for item selection:

- Satisfy the content specifications in terms of objective coverage and the number and percentage of MC and CR items on the test. An often used criterion for objective coverage is within 5% of the percentages of score points and items per objective.
- Avoid selecting poor-fitting items, items with too high/low p-values, items with flagged point biserials (the research department will provide a list of such items).
- Avoid items flagged for local dependency if the flagged items come from different passages. If the flagged items come from the same passage, they are expected to be dependent on each other to some degree and they are not a problem.
- Minimize the number of items flagged for DIF (gender, ethnic, and High/Low Needs schools). Flagged items should be reviewed for content again. It needs to be remembered that some items may be flagged for DIF by chance only and their content may not necessarily be biased against any of the analyzed groups. Research will provide DIF information for each item. It is also possible to get "significant" DIF yet not bias if the content is a necessary part of the construct that is measured. That is, some items may be flagged for DIF not out of chance and still not represent bias.
- Verify that the items will be administered in the same relative positions in both the FT and OP forms (e.g., the first item in a FT form should also be the first item in an OP form). When that is impossible, please ensure that they are in the same one-third section of the forms.
- Evaluate the alignment of TCC and SE curves of the proposed 2010 OP forms and the 2009 OP forms.
- From the ITEMWIN output evaluate expected percentage of maximum raw score at each scale score and difference between reference set (2008) and working set (2009)—we want the difference to be no more than 0.01, which is unfortunately sometimes hard to achieve, but please try your best.
    - It is especially important to get a good curve alignment at and around proficiency level cut scores. Good alignment will help preserve the impact data from the previous year of testing.
- Try to get the best scale coverage—make sure that MC items cover a wide range of the scale.
- Provide the research department with the following item selection information:
    - Percentage of score points per learning standard (target, 2010 full selection, 2010 MC items only)
    - Item number in 2010 OP book
    - Item unique identification number, item type, FT year, FT form, and FT item number

- Item classical statistics (p-values, point biserials, etc.)
- ITEMWIN output (including TCCs)
- Summary file with IRT item parameters for selected items

Appendix D—Factor Analysis Results

As described in Section III, "Validity," a principal component factor analysis was conducted on the Grades 3–8 ELA Tests data. The analyses were conducted for the total population of students and selected subpopulations: English language learners (ELL), students with disabilities (SWD), students using accommodations (SUA), SWD students using disability accommodation (SWD/SUA) and ELL students using ELL-related accommodations (ELL/SUA). Table D1 contains the results of factor analysis on subpopulation data.

Table D1. Factor Analysis Results for ELA Tests (Selected Subpopulations)

Grade	Subgroup	Initial Eigenvalues			
		Component	Total	% of Variance	Cumulative %
3	ELL	**1**	6.07	21.67	21.67
		2	1.20	4.27	25.94
		3	1.03	3.66	29.60
	SWD	**1**	6.78	24.21	24.21
		2	1.28	4.59	28.80
		3	1.07	3.82	32.62
		4	1.00	3.59	36.20
	SUA	6.69	23.89	23.89	6.60
		1.22	4.35	28.24	1.20
		1.05	3.76	32.00	1.00
	SWD /SUA	**1**	**6.59**	**23.54**	**23.54**
		2	1.31	4.67	28.21
		3	1.07	3.82	32.04
		4	1.00	3.59	35.62
	ELL /SUA	**1**	**5.87**	**20.97**	**20.97**
		2	1.19	4.26	25.23
		3	1.03	3.68	28.90
		4	1.00	3.57	32.47
4	ELL	**1**	**6.05**	**19.51**	**19.51**
		2	1.26	4.05	23.56
		3	1.08	3.47	27.04
		4	1.04	3.35	30.39
	SWD	**1**	6.85	22.10	22.10
		2	1.26	4.07	26.17
		3	1.08	3.49	29.66
		4	1.03	3.34	33.00
	SUA	1	6.77	21.83	21.83
		2	1.27	4.11	25.94
		3	1.08	3.47	29.40
		4	1.02	3.28	32.68

(Continued on next page)

Table D1. Factor Analysis Results for ELA Tests (Selected Subpopulations) (cont.)

Grade	Subgroup	Initial Eigenvalues			
		Component	Total	% of Variance	Cumulative %
4	SWD /SUA	**1**	**6.71**	**21.65**	**21.65**
		2	1.25	4.04	25.68
		3	1.09	3.51	29.20
		4	1.04	3.34	32.54
	ELL /SUA	**1**	**5.86**	**18.89**	**18.89**
		2	1.26	4.06	22.95
		3	1.07	3.46	26.41
		4	1.04	3.36	29.77
5	ELL	**1**	**5.48**	**20.30**	**20.30**
		2	1.18	4.37	24.66
		3	1.13	4.19	28.85
	SWD	**1**	**5.76**	**21.35**	**21.35**
		2	1.24	4.59	25.93
		3	1.16	4.30	30.24
	SUA	**1**	**5.88**	**21.79**	**21.79**
		2	1.22	4.52	26.31
		3	1.13	4.19	30.50
	SWD /SUA	**1**	**5.67**	**20.99**	**20.99**
		2	1.24	4.60	25.59
		3	1.16	4.29	29.88
	ELL /SUA	**1**	**5.43**	**20.10**	**20.10**
		2	1.19	4.39	24.48
		3	1.11	4.12	28.60
6	ELL	**1**	**6.31**	**21.75**	**21.75**
		2	1.14	3.93	25.68
		3	1.06	3.65	29.34
		4	1.00	3.46	32.79
	SWD	**1**	**7.20**	**24.84**	**24.84**
		2	1.19	4.09	28.93
		3	1.11	3.83	32.76
	SUA	**1**	**7.29**	**25.14**	**25.14**
		2	1.19	4.10	29.24
		3	1.07	3.69	32.93

(Continued on next page)

Table D1. Factor Analysis Results for ELA Tests (Selected Subpopulations) (cont.)

Grade	Subgroup	Initial Eigenvalues			
		Component	Total	% of Variance	Cumulative %
6	SWD /SUA	**1**	**7.09**	**24.44**	**24.44**
		2	1.18	4.06	28.50
		3	1.12	3.88	32.38
	ELL /SUA	**1**	**6.30**	**21.73**	**21.73**
		2	1.16	3.98	25.71
		3	1.06	3.64	29.35
		4	1.01	3.47	32.82
7	ELL	**1**	**6.54**	**18.68**	**18.68**
		2	1.19	3.39	22.06
		3	1.18	3.36	25.42
		4	1.11	3.17	28.59
		5	1.02	2.93	31.52
	SWD	**1**	7.07	20.19	20.19
		2	1.28	3.65	23.84
		3	1.24	3.55	27.39
		4	1.07	3.07	30.46
	SUA	**1**	**7.29**	**20.84**	**20.84**
		2	1.25	3.58	24.42
		3	1.20	3.42	27.84
		4	1.05	3.01	30.85
	SWD /SUA	**1**	**6.94**	**19.84**	**19.84**
		2	1.28	3.65	23.49
		3	1.23	3.51	27.00
		4	1.08	3.09	30.09
	ELL /SUA	**1**	**6.63**	**18.93**	**18.93**
		2	1.18	3.37	22.30
		3	1.17	3.34	25.63
		4	1.11	3.17	28.80
		5	1.03	2.95	31.74
8	ELL	**1**	**5.56**	**19.16**	**19.16**
		2	1.16	3.98	23.14
		3	1.07	3.68	26.82
		4	1.05	3.62	30.44
		5	1.03	3.55	33.99
		6	1.00	3.46	37.45

(Continued on next page)

Table D1. Factor Analysis Results for ELA Tests (Selected Subpopulations) (cont.)

Grade	Subgroup	Initial Eigenvalues			
		Component	Total	% of Variance	Cumulative %
8	SWD	**1**	**6.08**	**20.95**	**20.95**
		2	1.09	3.76	24.71
		3	1.07	3.67	28.38
	SUA	**1**	6.22	21.44	21.44
		2	1.10	3.80	25.24
		3	1.05	3.62	28.85
		4	1.01	3.47	32.32
	SWD /SUA	**1**	**6.00**	**20.70**	**20.70**
		2	1.09	3.75	24.45
		3	1.07	3.68	28.13
		4	1.00	3.45	31.58
	ELL /SUA	**1**	**5.54**	**19.10**	**19.10**
		2	1.16	3.99	23.09
		3	1.08	3.72	26.81
		4	1.06	3.65	30.46
		5	1.03	3.54	33.99

Appendix E—Items Flagged for DIF

These tables support the DIF information in Section V, "Operational Test Data Collection and Classical Analysis," and Section VI, "IRT Scaling and Equating." They include item numbers, focal group, and directions of DIF and DIF statistics. Table E1 shows items flagged by the SMD and Mantel-Haenszel methods, and Table E2 presents items flagged by the Linn-Harnisch method. Note that positive values of SMD and Delta in Table E1 indicate DIF in favor of a focal group and negative values of SMD and Delta indicate DIF against a focal group.

Table E1. NYSTP ELA 2010 Classical DIF Item Flags

Grade	Item #	Subgroup	DIF	SMD	Mantel-Haenszel	Delta
3	5	Asian	Against	-0.119	1237.469	-1.714
3	16	ELL	Against	No Flag	934.169	-1.540
4	6	ELL	Against	No Flag	1216.630	-2.041
4	6	Asian	Against	No Flag	337.993	-1.826
4	16	Asian	Against	No Flag	1378.500	-2.009
4	20	ELL	Against	-0.104	No Flag	No Flag
4	29	Female	In Favor	0.114	No Flag	No Flag
4	30	Female	In Favor	0.149	No Flag	No Flag
4	31	Female	In Favor	0.119	No Flag	No Flag
4	31	ELL	Against	-0.129	No Flag	No Flag
5	21	ELL	In Favor	0.102	No Flag	No Flag
5	27	ELL	Against	-0.255	No Flag	No Flag
5	27	Hispanic	Against	-0.128	No Flag	No Flag
6	27	Female	In Favor	0.132	No Flag	No Flag
6	27	ELL	Against	-0.11	No Flag	No Flag
6	28	High Need	Against	-0.151	No Flag	No Flag
6	29	Female	In Favor	0.137	No Flag	No Flag
6	29	ELL	Against	-0.109	No Flag	No Flag
7	5	ELL	Against	-0.124	895.918	-1.715
7	5	Asian	Against	No Flag	281.803	-1.529
7	13	Asian	Against	No Flag	314.521	-1.812
7	14	ELL	Against	No Flag	654.195	-1.780
7	20	Hispanic	In Favor	0.103	No Flag	No Flag
7	27	ELL	In Favor	0.101	No Flag	No Flag
7	28	Female	In Favor	0.104	No Flag	No Flag
7	28	ELL	In Favor	0.193	No Flag	No Flag
7	30	Female	Against	No Flag	2585.61	-1.517
7	35	High Need	Against	-0.184	No Flag	No Flag
7	35	ELL	Against	-0.328	No Flag	No Flag
8	3	Hispanic	Against	No Flag	1074.310	-2.040
8	3	ELL	Against	-0.166	1756.880	-2.519

(Continued on next page)

Table E1. NYSTP ELA 2010 Classical DIF Item Flags (cont.)

Grade	Item #	Subgroup	DIF	SMD	Mantel-Haenszel	Delta
8	6	Asian	Against	No Flag	684.848	-1.748
8	6	ELL	Against	-0.12	No Flag	No Flag
8	8	ELL	In Favor	0.168	965.292	1.767
8	27	High Need	Against	-0.216	No Flag	No Flag
8	27	Black	Against	-0.187	No Flag	No Flag
8	27	Hispanic	Against	-0.205	No Flag	No Flag
8	27	ELL	Against	-0.338	No Flag	No Flag
8	27	Female	In Favor	0.108	No Flag	No Flag
8	28	High Need	Against	-0.15	No Flag	No Flag
8	28	Black	Against	-0.125	No Flag	No Flag
8	28	Hispanic	Against	-0.108	No Flag	No Flag
8	28	ELL	Against	-0.129	No Flag	No Flag
8	28	Female	In Favor	0.12	No Flag	No Flag
8	29	High Need	Against	-0.15	No Flag	No Flag
8	29	Black	Against	-0.131	No Flag	No Flag
8	29	Hispanic	Against	-0.124	No Flag	No Flag
8	29	ELL	Against	-0.199	No Flag	No Flag
8	29	Female	In Favor	0.11	No Flag	No Flag

In Table E2, note that positive values of D_{ig} indicate DIF in favor of a focal group and negative values of D_{ig} indicate DIF against a focal group.

Table E2. Items Flagged for DIF by the Linn-Harnisch Method

Grade	Item	Focal Group	Direction	Magnitude (D_{ig})
3	5	Asian	Against	-0.104
5	27	ELL	Against	-0.217
7	27	ELL	In Favor	0.123
7	28	ELL	In Favor	0.199
7	35	Black	Against	-0.130
7	35	Hispanic	Against	-0.100
7	35	ELL	Against	-0.259
8	3	ELL	Against	-0.102
8	6	ELL	Against	-0.105
8	8	ELL	In Favor	0.125

Appendix F—Item-Model Fit Statistics

These tables support the item-model fit information in Section VI, "IRT Scaling and Equating." The item number, calibration model, chi-square, degrees of freedom (DF), N-count, obtained-*Z* fit statistic, and critical-*Z* fit statistic are presented for each item. Fit for all items in the Grades 3–8 ELA Tests was acceptable (critical *Z* > obtained *Z*).

Table F1. ELA Item Fit Statistics, Grade 3

Item Number	Model	Chi-Square	DF	N-count	Obtained *Z*	Critical *Z*	Fit Ok?
1	3PL	310.29	7	183455	81.06	489.21	Y
2	3PL	246.52	7	183455	64.01	489.21	Y
3	3PL	739.03	7	183455	195.64	489.21	Y
4	3PL	69.05	7	183455	16.58	489.21	Y
5	3PL	855.11	7	183455	226.67	489.21	Y
6	3PL	298.73	7	183455	77.97	489.21	Y
7	3PL	307.74	7	183455	80.38	489.21	Y
8	3PL	2569.24	7	183455	684.79	489.21	N
9	3PL	152.65	7	183455	38.93	489.21	Y
10	3PL	663.07	7	183455	175.34	489.21	Y
11	3PL	465.00	7	183455	122.41	489.21	Y
12	3PL	374.53	7	183455	98.23	489.21	Y
13	3PL	153.99	7	183455	39.28	489.21	Y
14	3PL	2173.71	7	183455	579.08	489.21	N
15	3PL	489.60	7	183455	128.98	489.21	Y
16	3PL	120.61	7	183455	30.36	489.21	Y
17	3PL	288.24	7	183455	75.16	489.21	Y
18	3PL	351.39	7	183455	92.04	489.21	Y
19	3PL	1568.19	7	183455	417.25	489.21	Y
20	3PL	353.93	7	183455	92.72	489.21	Y
21	2PPC	745.15	17	183455	124.88	489.21	Y
22	3PL	134.30	7	183455	34.02	489.21	Y
23	3PL	544.06	7	183455	143.53	489.21	Y
24	3PL	61.32	7	183455	14.52	489.21	Y
25	3PL	365.38	7	183455	95.78	489.21	Y
26	2PPC	595.36	17	183455	99.19	489.21	Y
27	2PPC	1006.54	17	183455	169.71	489.21	Y
28	2PPC	674.13	26	183455	89.88	489.21	Y

Table F2. ELA Item Fit Statistics, Grade 4

Item Number	Model	Chi-Square	DF	N-count	Obtained *Z*	Critical *Z*	Fit Ok?
1	3PL	341.98	7	194877	89.53	519.67	Y
2	3PL	119.53	7	194877	30.08	519.67	Y
3	3PL	102.34	7	194877	25.48	519.67	Y
4	3PL	746.27	7	194877	197.58	519.67	Y
5	3PL	22.90	7	194877	4.25	519.67	Y
6	3PL	102.94	7	194877	25.64	519.67	Y
7	3PL	384.60	7	194877	100.92	519.67	Y
8	3PL	102.94	7	194877	25.64	519.67	Y
9	3PL	60.70	7	194877	14.35	519.67	Y
10	3PL	143.17	7	194877	36.39	519.67	Y
11	3PL	138.54	7	194877	35.16	519.67	Y
12	3PL	190.31	7	194877	48.99	519.67	Y
13	3PL	323.73	7	194877	84.65	519.67	Y
14	3PL	108.24	7	194877	27.06	519.67	Y
15	3PL	131.09	7	194877	33.16	519.67	Y
16	3PL	133.90	7	194877	33.91	519.67	Y
17	3PL	114.68	7	194877	28.78	519.67	Y
18	3PL	71.69	7	194877	17.29	519.67	Y
19	3PL	265.09	7	194877	68.98	519.67	Y
20	3PL	176.70	7	194877	45.35	519.67	Y
21	3PL	144.15	7	194877	36.66	519.67	Y
22	3PL	76.14	7	194877	18.48	519.67	Y
23	3PL	142.57	7	194877	36.23	519.67	Y
24	3PL	174.03	7	194877	44.64	519.67	Y
25	3PL	79.24	7	194877	19.31	519.67	Y
26	3PL	623.80	7	194877	164.85	519.67	Y
27	3PL	127.45	7	194877	32.19	519.67	Y
28	3PL	414.44	7	194877	108.89	519.67	Y
29	2PPC	1281.27	35	194877	148.96	519.67	Y
30	2PPC	2503.54	35	194877	295.05	519.67	Y
31	2PPC	914.36	26	194877	123.19	519.67	Y

Table F3. ELA Item Fit Statistics, Grade 5

Item Number	Model	Chi-Square	DF	N-count	Obtained *Z*	Critical *Z*	Fit Ok?
1	3PL	126.55	7	186486	31.95	497.30	Y
2	3PL	69.02	7	186486	16.58	497.30	Y
3	3PL	126.14	7	186486	31.84	497.30	Y
4	3PL	323.02	7	186486	84.46	497.30	Y
5	3PL	141.39	7	186486	35.92	497.30	Y
6	3PL	37.85	7	186486	8.25	497.30	Y
7	3PL	55.84	7	186486	13.05	497.30	Y
8	3PL	2283.95	7	186486	608.54	497.30	N
9	3PL	86.61	7	186486	21.28	497.30	Y
10	3PL	231.49	7	186486	60.00	497.30	Y
11	3PL	300.84	7	186486	78.53	497.30	Y
12	3PL	215.76	7	186486	55.79	497.30	Y
13	3PL	930.06	7	186486	246.70	497.30	Y
14	3PL	390.67	7	186486	102.54	497.30	Y
15	3PL	203.13	7	186486	52.42	497.30	Y
16	3PL	517.91	7	186486	136.55	497.30	Y
17	3PL	477.26	7	186486	125.68	497.30	Y
18	3PL	361.10	7	186486	94.64	497.30	Y
19	3PL	470.25	7	186486	123.81	497.30	Y
20	3PL	398.40	7	186486	104.61	497.30	Y
21	2PPC	857.69	17	186486	144.18	497.30	Y
22	3PL	108.17	7	186486	27.04	497.30	Y
23	3PL	146.90	7	186486	37.39	497.30	Y
24	3PL	61.53	7	186486	14.57	497.30	Y
25	3PL	48.81	7	186486	11.17	497.30	Y
26	2PPC	988.38	17	186486	166.59	497.30	Y
27	2PPC	2582.15	26	186486	354.47	497.30	Y

Table F4. ELA Item Fit Statistics, Grade 6

Item Number	Model	Chi-Square	DF	N-count	Obtained *Z*	Critical *Z*	Fit Ok?
1	3PL	140.31	7	189913	35.63	506.43	Y
2	3PL	50.62	7	189913	11.66	506.43	Y
3	3PL	98.94	7	189913	24.57	506.43	Y
4	3PL	241.29	7	189913	62.62	506.43	Y
5	3PL	37.46	7	189913	8.14	506.43	Y
6	3PL	272.90	7	189913	71.06	506.43	Y
7	3PL	61.97	7	189913	14.69	506.43	Y
8	3PL	304.02	7	189913	79.38	506.43	Y
9	3PL	98.40	7	189913	24.43	506.43	Y
10	3PL	575.28	7	189913	151.88	506.43	Y
11	3PL	110.00	7	189913	27.53	506.43	Y
12	3PL	146.66	7	189913	37.32	506.43	Y
13	3PL	480.02	7	189913	126.42	506.43	Y
14	3PL	162.93	7	189913	41.68	506.43	Y
15	3PL	183.81	7	189913	47.25	506.43	Y
16	3PL	89.07	7	189913	21.93	506.43	Y
17	3PL	146.04	7	189913	37.16	506.43	Y
18	3PL	223.86	7	189913	57.96	506.43	Y
19	3PL	633.14	7	189913	167.34	506.43	Y
20	3PL	294.85	7	189913	76.93	506.43	Y
21	3PL	538.75	7	189913	142.11	506.43	Y
22	3PL	177.70	7	189913	45.62	506.43	Y
23	3PL	169.36	7	189913	43.39	506.43	Y
24	3PL	423.93	7	189913	111.43	506.43	Y
25	3PL	433.41	7	189913	113.96	506.43	Y
26	3PL	199.11	7	189913	51.34	506.43	Y
27	2PPC	2689.88	44	189913	282.05	506.43	Y
28	2PPC	4689.28	44	189913	495.19	506.43	Y
29	2PPC	1026.33	26	189913	138.72	506.43	Y

Table F5. ELA Item Fit Statistics, Grade 7

Item Number	Model	Chi-Square	DF	N-count	Obtained *Z*	Critical *Z*	Fit Ok?
1	3PL	622.99	7	187873	164.63	500.99	Y
2	3PL	2006.38	7	187873	534.36	500.99	N
3	3PL	51.8	7	187873	11.97	500.99	Y
4	3PL	80.76	7	187873	19.71	500.99	Y
5	3PL	111.73	7	187873	27.99	500.99	Y
6	3PL	48.58	7	187873	11.11	500.99	Y
7	3PL	456.56	7	187873	120.15	500.99	Y
8	3PL	31.17	7	187873	6.46	500.99	Y
9	3PL	57.32	7	187873	13.45	500.99	Y
10	3PL	61.53	7	187873	14.58	500.99	Y
11	3PL	203.17	7	187873	52.43	500.99	Y
12	3PL	111.21	7	187873	27.85	500.99	Y
13	3PL	69.16	7	187873	16.61	500.99	Y
14	3PL	128.26	7	187873	32.41	500.99	Y
15	3PL	318.29	7	187873	83.20	500.99	Y
16	3PL	73.28	7	187873	17.71	500.99	Y
17	3PL	184.13	7	187873	47.34	500.99	Y
18	3PL	26.33	7	187873	5.17	500.99	Y
19	3PL	57.66	7	187873	13.54	500.99	Y
20	3PL	222.49	7	187873	57.59	500.99	Y
21	3PL	28.45	7	187873	5.73	500.99	Y
22	3PL	319.04	7	187873	83.40	500.99	Y
23	3PL	2813.41	7	187873	750.05	500.99	N
24	3PL	84.18	7	187873	20.63	500.99	Y
25	3PL	480.57	7	187873	126.57	500.99	Y
26	3PL	108.39	7	187873	27.10	500.99	Y
27	2PPC	1953.08	17	187873	332.04	500.99	Y
28	2PPC	1538.5	17	187873	260.93	500.99	Y
29	3PL	46.54	7	187873	10.57	500.99	Y
30	3PL	182.34	7	187873	46.86	500.99	Y
31	3PL	44.25	7	187873	9.96	500.99	Y
32	3PL	20.39	7	187873	3.58	500.99	Y
33	2PPC	501.32	17	187873	83.06	500.99	Y
34	2PPC	493.14	17	187873	81.66	500.99	Y
35	2PPC	1357.9	26	187873	184.70	500.99	Y

Table F6. ELA Item Fit Statistics, Grade 8

Item Number	Model	Chi-Square	DF	N-count	Obtained *Z*	Critical *Z*	Fit Ok?
1	3PL	178.01	7	196081	45.70	522.88	Y
2	3PL	63.81	7	196081	15.18	522.88	Y
3	3PL	63.49	7	196081	15.10	522.88	Y
4	3PL	1476.37	7	196081	392.71	522.88	Y
5	3PL	175.38	7	196081	45.00	522.88	Y
6	3PL	644.85	7	196081	170.47	522.88	Y
7	3PL	335.63	7	196081	87.83	522.88	Y
8	3PL	70.90	7	196081	17.08	522.88	Y
9	3PL	68.33	7	196081	16.39	522.88	Y
10	3PL	1445.23	7	196081	384.38	522.88	Y
11	3PL	391.24	7	196081	102.69	522.88	Y
12	3PL	88.09	7	196081	21.67	522.88	Y
13	3PL	31.87	7	196081	6.65	522.88	Y
14	3PL	64.31	7	196081	15.32	522.88	Y
15	3PL	34.54	7	196081	7.36	522.88	Y
16	3PL	9.31	7	196081	0.62	522.88	Y
17	3PL	116.18	7	196081	29.18	522.88	Y
18	3PL	136.88	7	196081	34.71	522.88	Y
19	3PL	64.31	7	196081	15.32	522.88	Y
20	3PL	266.21	7	196081	69.28	522.88	Y
21	3PL	133.16	7	196081	33.72	522.88	Y
22	3PL	125.95	7	196081	31.79	522.88	Y
23	3PL	91.73	7	196081	22.64	522.88	Y
24	3PL	3918.17	7	196081	1045.30	522.88	N
25	3PL	390.90	7	196081	102.60	522.88	Y
26	3PL	2388.93	7	196081	636.60	522.88	N
27	2PPC	4858.18	44	196081	513.19	522.88	Y
28	2PPC	5194.15	44	196081	549.01	522.88	N
29	2PPC	1424.97	26	196081	194.00	522.88	Y

Appendix G—Derivation of the Generalized SPI Procedure

The standard performance index (SPI) is an estimated true score (estimated proportion of total or maximum points obtained) based on the performance of a given examinee for the items in a given learning standard. Assume a *k*-item test composed of *j* standards with a maximum possible raw score of *n*. Also assume that each item contributes to at most one standard, and the k_j items in standard *j* contribute a maximum of n_j points. Define X_j as the observed raw score on standard *j*. The true score is

$$T_j \equiv E(X_j / n_j).$$

It is assumed that there is information available about the examinee in addition to the standard score, and this information provides a prior distribution for T_j. This prior distribution of T_j for a given examinee is assumed to be $\beta(r_j, s_j)$:

$$g(T_j) = \frac{(r_j + s_j - 1)! T_j^{r_j - 1}(1 - T_j)^{s_j - 1}}{(r_j - 1)!(s_j - 1)!} \qquad (1)$$

for $0 \le T_j \le 1$; $r_j, s_j > 0$. Estimates of r_j and s_j are derived from IRT (Lord, 1980).

It is assumed that X_j follows a binomial distribution, given T_j:

$$p(X_j = x_j \mid T_j) = Binomial\,(n_j, T_j = \sum_{i=1}^{k_j} T_i / n_j),$$

where

T_i is the expected value of the score for item *i* in standard *j* for a given θ.

Given these assumptions, the posterior distribution of T_j, given x_j, is

$$g(T_j \mid X_j = x_j) = \beta(p_j, q_j), \qquad (2)$$

with

$$p_j = r_j + x_j \qquad (3)$$

and

$$q_j = s_j + n_j - x_j. \qquad (4)$$

The SPI is defined to be the mean of this posterior distribution:

$$\tilde{T}_j = \frac{p_j}{p_j + q_j}.$$

Following Novick and Jackson (1974, p. 119), a mastery band is created to be the *C*% central credibility interval for T_j. It is obtained by identifying the values that place $\frac{1}{2}(100 - C)\%$ of the $\beta(p_j, q_j)$ density in each tail of the distribution.

Estimation of the Prior Distribution of T_j

The *k* items in each test are scaled together using a generalized IRT model (3PL/2PPC) that fits a three-parameter logistic model (3PL) to the MC items and a generalized partial-credit model (2PPC) to the CR items (Yen, 1993).

The 3PL model is

$$P_i(\theta) = P(X_i = 1 | \theta) = c_i + \frac{1 - c_i}{1 + \exp\left[-1.7A_i\left(\theta - B_i\right)\right]} \qquad (5)$$

where

A_i is the discrimination, B_i is the location, and c_i is the guessing parameter for item *i*.

A generalization of Master's (1982) partial-credit (2PPC) model was used for the CR items. The 2PPC model, the same as Muraki's (1992) "generalized partial credit model," has been shown to fit response data obtained from a wide variety of mixed-item type achievement tests (Fitzpatrick, Link, Yen, Burket, Ito, and Sykes, 1996). For a CR item with 1_i score levels, integer scores are assigned that ranged from 0 to $1_i - 1$:

$$P_{im}(\theta) = P(X_i = m-1|\theta) = \frac{\exp(z_{im})}{\sum_{g=1}^{1_i}\exp(z_{ig})}, \qquad m = 1, \ldots 1_i \qquad (6)$$

where

$$z_{ig} = \alpha_i\left(m-1\right)\theta - \sum_{h=0}^{m-1}\gamma_{ih}, \qquad (7)$$

and

$$\gamma_{i0} = 0$$

Alpha (α_i) is the item discrimination and gamma (γ_{ih}) is related to the difficulty of the item levels: the trace lines for adjacent score levels intersect at γ_{ih}/α_i .

Item parameters estimated from the national standardization sample are used to obtain SPI values. $T_{ij}(\theta)$ is the expected score for item *i* in standard *j*, and θ is the common trait value to which the items are scaled:

$$T_{ij}(\theta) = \sum_{m=1}^{1_i}(m-1)P_{ijm}(\theta)$$

where

1_i is the number of score levels in item *i*, including 0.

T_j, the expected proportion of maximum score for standard *j*, is

$$T_j = \frac{1}{n_j}\left[\sum_{i=1}^{k_j}T_{ij}(\theta)\right]. \qquad (8)$$

The expected score for item *i* and estimated proportion-correct of maximum score for standard *j* are obtained by substituting the estimate of the trait $(\hat{\theta})$ for the actual trait value.

The theoretical random variation in item response vectors and resulting $(\hat{\theta})$ values for a given examinee produces the distribution $g(\hat{T}_j \mid \hat{\theta})$ with mean $\mu(\hat{T}_j \mid \theta)$ and variance $\sigma^2(\hat{T}_j \mid \theta)$. This distribution is used to estimate a prior distribution of T_j. Given that T_j is assumed to be distributed as a beta distribution (equation 1), the mean $[\mu(\hat{T}_j \mid \theta)]$ and variance $[\sigma^2(\hat{T}_j \mid \theta)]$ of this distribution can be expressed in terms of its parameters, r_j and s_j.

Expressing the mean and variance of the prior distribution in terms of the parameters of the beta distribution produces (Novick and Jackson, 1974, p. 113)

$$\mu(\hat{T}_j \mid \theta) = \frac{r_j}{r_j + s_j} \tag{9}$$

and

$$\sigma^2(\hat{T}_j \mid \theta) = \frac{r_j s_j}{(r_j + s_j)^2 (r_j + s_j + 1)} \ . \tag{10}$$

Solving these equations for r_j and s_j produces

$$r_j = \mu(\hat{T}_j \mid \theta) n_j^* \tag{11}$$

and

$$s_j = [1 - \mu(\hat{T}_j \mid \theta)] n_j^*, \tag{12}$$

where

$$n_j^* = \frac{\mu(\hat{T}_j \mid \theta)\left[1 - \mu(\hat{T}_j \mid \theta)\right]}{\sigma^2(\hat{T}_j \mid \theta)} - 1. \tag{13}$$

Using IRT, $\sigma^2(\hat{T}_j \mid \theta)$ can be expressed in terms of item parameters (Lord, 1983):

$$\mu(\hat{T}_j \mid \theta) \approx \frac{1}{n_j} \sum_{i=1}^{k_j} \hat{T}_{ij}(\theta). \tag{14}$$

Because T_j is a monotonic transformation of θ (Lord, 1980, p.71),

$$\sigma^2(\hat{T}_j \mid \theta) = \sigma^2(\hat{T}_j \mid T_j) \approx I(T_j, \hat{T}_j)^{-1} \tag{15}$$

where

$I(T_j, \hat{T}_j)$ is the information that $\hat{T}_j$ contributes about T_j.

Given these results, Lord (1980, p. 79 and 85) produces

$$I(T_j, \hat{T}_j) = \frac{I(\theta, \hat{T}_j)}{\left(\partial T_j / \partial \theta\right)^2}, \tag{16}$$

and

$$I(\theta, \hat{T}_j) \approx I(\theta, \hat{\theta}). \tag{17}$$

Thus,

$$\sigma^2(\hat{T}_j \mid \theta) \approx \frac{\left[\frac{1}{n_j}\sum_{i=1}^{k_j} \hat{T}_{ij}(\theta)\right]^2}{I(\theta, \hat{\theta})}$$

and the parameters of the prior beta distribution for T_j can be expressed in terms of the parameters of the three-parameter IRT and two-parameter partial-credit models. Furthermore, the parameters of the posterior distribution of T_j also can be expressed in terms of the IRT parameters:

$$p_j = \hat{T}_j n_j^* + x_j, \tag{18}$$

and

$$q_j = \left[1 - \hat{T}_j\right] n_j^* + n_j - x_j. \tag{19}$$

The OPI is

$$\tilde{T}_j = \frac{p_j}{p_j + q_j} \tag{20}$$

$$= \frac{\hat{T}_j n_j^* + x_j}{n^*_j + n_j}. \tag{21}$$

The SPI can also be written in terms of the relative contribution of the prior estimate $\hat{T}_j$, and the observed proportion of maximum raw (correct score) (OPM), x_j / n_j, as

$$\tilde{T}_j = w_j \hat{T}_j + (1 - w_j)\left[x_j / n_j\right]. \tag{22}$$

w_j, a function of the mean and variance of the prior distribution, is the relative weight given to the prior estimate:

$$w_j = \frac{n_j^*}{n_j^* + n_j}. \tag{23}$$

The term n_j^* may be interpreted as the contribution of the prior in terms of theoretical numbers of items.

Check on Consistency and Adjustment of Weight Given to Prior Estimate

The item responses are assumed to be described by $P_i(\hat{\theta})$ or $P_{im}(\hat{\theta})$, depending on the type of item. Even if the IRT model accurately described item performance over examinees, their item responses grouped by standard may be multidimensional. For example, a particular examinee may be able to perform difficult addition but not easy subtraction. Under these circumstances, it is not appropriate to pool the prior estimate, $\hat{T}_j$, with x_j / n_j. In calculating the SPI, the following statistic was used to identify examinees with unexpected performance on the standards in a test:

$$Q = \sum_{j=1}^{J} n_j \left(\frac{x_j}{n_j} - \hat{T}_j\right)^2 / (\hat{T}_j(1-\hat{T}_j)) . \qquad (24)$$

If $Q \le \chi^2(J, .10)$, the weight, w_j, is computed and the SPI is produced. If $Q > \chi^2(J, .10)$, n_j^* and subsequently w_j is set equal to 0 and the OPM is used as the estimate of standard performance.

As previously noted, the prior is estimated using an ability estimate based on responses to all the items (including the items of standard *j*) and hence is not independent of X_j. An adjustment for the overlapping information that requires minimal computation is to multiply the test information in equation 5 by the factor $(n - n_j)/n$. The application of this factor produces an "adjusted" SPI estimate that can be compared to the "unadjusted" estimate.

Possible Violations of the Assumptions

Even if the IRT model fits the test items, the responses for a given examinee, grouped by standard, may be multidimensional. In these cases, it would not be appropriate to pool the prior estimate, $\hat{T}_j$, with x_j / n_j. A chi-square fit statistic is used to evaluate the observed proportion of maximum raw score (OPM) relative to that predicted for the items in the standard on the basis of the student's overall trait estimate. If the chi-square is significant, the prior estimate is not used and the OPM obtained becomes the student's standard score.

If the items in the standard do not permit guessing, it is reasonable to assume $\hat{T}_j$, the expected proportion correct of maximum score for a standard, will be greater or equal to zero. If correct guessing is possible, as it is with MC items, there will be a non-zero lower limit to $\hat{T}_j$, and a three-parameter beta distribution in which $\hat{T}_j$ is greater than or equal to this lower limit (Johnson and Kotz, 1979, p. 37) would be more appropriate. The use of the two-parameter beta distribution would tend to underestimate T_j among very low-performing examinees. Working with tests containing exclusively MC items, Yen found that there does not appear to be a practical importance to this underestimation (Yen, 1987). The impact of any such effect would be reduced as the proportion of CR items in the test increases. The size of this effect, nonetheless, was evaluated using simulations (Yen, Sykes, Ito, and Julian 1997).

The SPI procedure assumes that $p(X_j T_j)$ is a binomial distribution. This assumption is appropriate only when all the items in a standard have the same Bernoulli item response function. Not only do real items differ in difficulty, but when there are mixed-item types, X_j is not the sum of n_j independent Bernoulli variables. It is instead the total raw score. In essence, the simplifying assumption has been made that each CR item with a maximum score of $1_j - 1$ is the sum of $1_j - 1$ independent Bernoulli variables. Thus, a complex compound distribution is theoretically more applicable than the binomial. Given the complexity of working with such a model, it appears valuable to determine if the simpler model described here is sufficiently accurate to be useful.

Finally, because the prior estimate of $T_j, \hat{T}_j$, is based on performance on the entire test, including standard *j*, the prior estimate is not independent of X_j. The smaller the ratio n_j / n, the less impact this dependence will have. The effect of the overlapping information would be to understate the width of the credibility interval. The extent to which the size of the credibility interval is too small was examined (Yen et al., 1997) by simulating standards that contained varying proportions of the total test points.

Appendix H—Derivation of Classification Consistency and Accuracy

Classification Consistency

Assume that θ is a single latent trait measured by a test and denote Φ as a latent random variable. When a test X consists of *K* items and its maximum number correct score is *N*, the marginal probability of the number correct (NC) score *x* is

$$P(X = x) = \int P(X = x \mid \Phi = \theta) g(\theta) d\theta, \quad x = 0,1,..., N$$

where

$g(\theta)$ is the density of θ.

In this report, the marginal distribution $P(X = x)$ is denoted as $f(x)$, and the conditional error distribution $P(X = x \mid \Phi = \theta)$ is denoted as $f(x \mid \theta)$. It is assumed that examinees are classified into one of H mutually exclusive categories on the basis of predetermined H-1 observed score cutoffs, C_1, C_2, …, C_{H-1}. Let L_h represent the h^{th} category into which examinees with $C_{h-1} \le X \le C_h$ are classified. $C_0 = 0$ and $C_H =$ the maximum number-correct score. Then, the conditional and marginal probabilities of each category classification are as follows:

$$P(X \in L_h \mid \theta) = \sum_{x=C_{h-1}}^{C_h} f(x \mid \theta), \; h = 1, 2, \ldots, \text{H}.$$

$$P(X \in L_h) = \int \sum_{x=C_{h-1}}^{C_h} f(x \mid \theta) g(\theta) d\theta, \; h = 1, 2, \ldots, \text{H}.$$

Because obtaining test scores from two independent administrations of New York State tests was not feasible due to item release after each OP administration, a psychometric model was used to obtain the estimated classification consistency indices using test scores from a single administration. Based on the psychometric model, a symmetric H*H contingency table can be constructed. The elements of H*H contingency table consist of the joint probabilities of the row and column observed category classifications.

That two administrations are independent implies that if X_1 and X_2 represent the raw score random variables on the two administrations, then, conditioned on θ, X_1 and X_2 are independent and identically distributed. Consequently, the conditional bivariate distribution of X_1 and X_2 is

$$f(x_1, \; x_2 \mid \theta) = f(x_1 \mid \theta) f(x_2 \mid \theta).$$

The marginal bivariate distribution of X_1 and X_2 can be expressed as follows:

$$f(x_1, \; x_2) = \int f(x_1, x_2 \mid \theta) f(\theta) d\theta.$$

Consistent classification means that both X_1 and X_2 fall in the same category. The conditional probability of falling in the same category on the two administrations is

$$P(X_1 \in L_h,\ X_2 \in L_h \mid \theta) = \left[\sum_{x_1 = C_{h-1}}^{C_h - 1} f(x_1 \mid \theta) \right]^2, \qquad h = 1, 2, \ldots, \text{H}.$$

The agreement index P, conditional on theta, is obtained by

$$P(\theta) = \sum_{h=1}^{H} P(X_1 \in L_h,\ X_2 \in L_h \mid \theta).$$

The agreement index (classification consistency) can be computed as

$$P = \int P(\theta) g(\theta) d(\theta).$$

The probability of consistent classification by chance, P_C, is the sum of squared marginal probabilities of each category classification.

$$P_C = \sum_{h=1}^{H} P(X_1 \in L_h) P(X_2 \in L_h) = \sum_{h=1}^{H} \left[P(X_1 \in L_h) \right]^2.$$

Then, the coefficient kappa (Cohen, 1960) is

$$k = \frac{P - P_C}{1 - P_C}.$$

Classification Accuracy

Let Γ_w denote true category. When an examinee has an observed score, $x \in L_h$ (h =1, 2,…, H), and a latent score , $\theta \in \Gamma_w$ (w=1, 2,…, H), an accurate classification is made when $h = w$. The conditional probability of accurate classification is

$$\gamma(\theta) = P(X \in L_w \mid \theta),$$

where

w is the category such that $\theta \in \Gamma_w$.

Appendix I—Concordance Tables

Table I1. Grade 3 ELA 2010 and TerraNova Scale Score Concordance Table

Raw Score OP	Scale Score OP	Scale Score TERRANOVA	NP	NCE
4	475	482	1	2
5	588	515	2	7
6	599	532	3	10
7	605	539	3	10
8	610	547	3	12
9	614	552	4	13
10	617	557	5	14
11	619	561	5	16
12	622	565	6	17
13	624	569	7	19
14	626	572	8	20
15	628	576	9	21
16	630	579	10	22
17	632	583	11	24
18	634	587	13	26
19	636	590	14	28
20	637	593	16	29
21	639	597	19	31
22	642	601	21	33
23	644	605	24	35
24	646	610	29	38
25	649	614	33	41
26	652	620	39	44
27	655	626	46	48
28	659	633	55	53
29	663	642	66	59
30	669	654	79	67
31	678	670	90	77
32	694	701	98	94
33	780	750	99	99

Table I2. Grade 4 ELA 2010 and TerraNova Scale Score Concordance Table

Raw Score OP	Scale Score OP	Scale Score TERRANOVA	NP	NCE
4	430	433	1	1
5	430	459	1	1
6	549	522	2	7
7	565	539	3	9
8	576	549	3	11
9	583	556	4	12
10	590	563	4	13
11	595	569	4	14
12	600	574	5	15
13	604	579	6	17
14	608	583	7	19
15	612	588	8	21
16	616	592	9	22
17	619	596	11	24
18	622	599	12	25
19	625	603	14	27
20	628	606	16	29
21	631	610	18	31
22	634	613	20	32
23	637	617	23	35
24	640	620	26	36
25	643	624	30	39
26	646	627	33	41
27	649	631	37	43
28	652	635	42	46
29	656	638	46	48
30	659	642	51	50
31	663	647	57	53
32	667	651	61	56
33	671	656	67	59
34	675	661	72	62
35	679	667	78	66
36	685	673	83	70
37	690	679	88	74
38	696	687	92	80
39	704	695	95	85
40	712	704	97	90
41	722	717	99	96
42	738	739	99	99
43	775	780	99	99

Table I3. Grade 5 ELA 2010 and TerraNova Scale Score Concordance Table

Raw Score OP	Scale Score OP	Scale Score TERRANOVA	NP	NCE
4	495	475	1	1
5	592	547	3	10
6	606	563	4	12
7	613	573	4	13
8	618	580	5	15
9	623	587	5	16
10	626	592	6	17
11	629	598	7	19
12	632	602	8	21
13	634	607	10	23
14	637	612	12	25
15	639	616	13	27
16	642	621	16	29
17	644	625	18	31
18	646	629	21	33
19	648	634	25	36
20	651	638	29	38
21	653	643	34	41
22	656	648	39	44
23	659	654	47	48
24	661	659	53	51
25	665	666	61	56
26	668	673	69	60
27	673	682	78	66
28	678	693	87	73
29	686	709	94	83
30	700	741	99	99
31	795	790	99	99

Table I4. Grade 6 ELA 2010 and TerraNova Scale Score Concordance Table

Raw Score	Scale Score	Scale Score	NP	NCE
OP	OP	TERRANOVA		
3	480	486	1	1
4	480	486	1	1
5	591	539	2	8
6	603	558	3	11
7	609	567	4	12
8	614	574	4	13
9	618	582	5	15
10	621	586	5	16
11	623	592	6	17
12	626	596	6	18
13	628	600	7	19
14	630	604	8	21
15	632	608	10	23
16	633	611	11	24
17	635	615	13	26
18	637	618	14	27
19	638	621	16	29
20	640	625	18	31
21	641	628	20	32
22	643	631	22	34
23	644	634	24	35
24	646	637	27	37
25	647	641	30	39
26	649	645	33	41
27	651	648	36	43
28	653	652	40	45
29	655	657	45	47
30	657	662	50	50
31	659	668	57	54
32	662	674	63	57
33	665	681	70	61
34	669	689	77	66
35	673	698	84	71
36	678	709	90	78
37	684	725	96	86
38	694	750	99	99
39	785	800	99	99

Table I5. Grade 7 ELA 2010 and TerraNova Scale Score Concordance Table

Raw Score OP	Scale Score OP	Scale Score TERRANOVA	NP	NCE
4	470	498	1	1
5	470	537	2	6
6	584	562	3	11
7	597	571	4	12
8	604	578	4	14
9	609	584	5	15
10	613	589	5	16
11	616	594	6	17
12	619	598	6	18
13	621	602	7	19
14	623	606	8	21
15	626	610	9	22
16	628	613	10	23
17	629	616	11	24
18	631	619	12	26
19	633	623	14	27
20	635	626	16	29
21	637	629	17	30
22	638	632	19	31
23	640	635	21	33
24	642	638	23	34
25	643	642	25	36
26	645	645	27	37
27	647	648	30	39
28	649	652	33	41
29	651	656	36	43
30	653	660	40	45
31	655	664	44	47
32	657	668	48	49
33	660	673	54	52
34	662	679	60	55
35	665	684	65	58
36	669	691	72	62
37	673	700	80	67
38	678	710	87	73
39	685	726	94	83
40	698	753	98	96
41	790	810	99	99

Table I6. Grade 8 ELA 2010 and *TerraNova* Scale Score Concordance Table

Raw Score OP	Scale Score OP	Scale Score TERRANOVA	NP	NCE
4	430	546	2	6
5	515	564	3	9
6	566	573	3	11
7	578	582	4	13
8	585	588	5	15
9	590	593	5	16
10	594	598	6	17
11	597	602	7	18
12	601	606	7	20
13	604	611	9	21
14	606	614	9	22
15	609	617	10	23
16	611	621	12	25
17	613	624	13	26
18	616	627	14	27
19	618	630	16	29
20	620	633	17	30
21	622	636	19	31
22	624	638	20	32
23	626	641	22	34
24	628	645	24	35
25	630	648	26	37
26	632	651	28	38
27	634	654	31	39
28	637	657	33	41
29	639	661	36	43
30	641	664	39	44
31	644	668	42	46
32	646	672	46	48
33	649	676	50	50
34	652	681	55	52
35	655	685	58	54
36	658	690	63	57
37	661	695	68	60
38	665	700	72	62
39	669	707	77	66
40	673	714	82	70
41	679	722	87	74
42	686	734	93	81
43	699	761	99	96
44	790	820	99	99

Appendix J—Scale Score Frequency Distributions

Tables I1–I6 depict the scale score (SS) distributions, by frequency (N-count), percent, cumulative frequency, and cumulative percent. This data includes all public and charter school students with valid scale scores.

Table J1. Grade 3 ELA 2010 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
475	256	0.13	256	0.13
588	226	0.12	482	0.25
599	353	0.18	835	0.43
605	446	0.23	1281	0.65
610	555	0.28	1836	0.93
614	610	0.31	2446	1.25
617	787	0.40	3233	1.65
619	794	0.40	4027	2.05
622	939	0.48	4966	2.53
624	1076	0.55	6042	3.08
626	1161	0.59	7203	3.67
628	1382	0.70	8585	4.37
630	1461	0.74	10046	5.11
632	1848	0.94	11894	6.06
634	2175	1.11	14069	7.16
636	2500	1.27	16569	8.44
637	2679	1.36	19248	9.80
639	3478	1.77	22726	11.57
642	4317	2.20	27043	13.77
644	5380	2.74	32423	16.51
646	6508	3.31	38931	19.82
649	8359	4.26	47290	24.08
652	10584	5.39	57874	29.46
655	13628	6.94	71502	36.40
659	17349	8.83	88851	45.23

(Continued on next page)

TableJ1. Grade 3 ELA 2010 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
663	21994	11.20	110845	56.43
669	26211	13.34	137056	69.78
678	26649	13.57	163705	83.34
694	22132	11.27	185837	94.61
780	10588	5.39	196425	100.00

Table J2. Grade 4 ELA 2010 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
430	168	0.08	168	0.08
549	109	0.05	277	0.14
565	174	0.09	451	0.23
576	269	0.14	720	0.36
583	316	0.16	1036	0.52
590	418	0.21	1454	0.73
595	496	0.25	1950	0.98
600	576	0.29	2526	1.27
604	595	0.30	3121	1.57
608	782	0.39	3903	1.96
612	835	0.42	4738	2.38
616	991	0.50	5729	2.88
619	1136	0.57	6865	3.45
622	1343	0.67	8208	4.12
625	1640	0.82	9848	4.94
628	1902	0.95	11750	5.90
631	2211	1.11	13961	7.01
634	2663	1.34	16624	8.34
637	3302	1.66	19926	10.00
640	3701	1.86	23627	11.86
643	4487	2.25	28114	14.11
646	5145	2.58	33259	16.69
649	6093	3.06	39352	19.75

(Continued on next page)

Table J2. Grade 4 ELA 2010 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
652	7088	3.56	46440	23.31
656	8088	4.06	54528	27.37
659	9235	4.63	63763	32.00
663	10463	5.25	74226	37.25
667	11776	5.91	86002	43.16
671	12880	6.46	98882	49.63
675	14050	7.05	112932	56.68
679	14590	7.32	127522	64.00
685	14887	7.47	142409	71.47
690	14182	7.12	156591	78.59
696	12645	6.35	169236	84.93
704	9992	5.01	179228	89.95
712	8134	4.08	187362	94.03
722	6413	3.22	193775	97.25
738	4107	2.06	197882	99.31
775	1372	0.69	199254	100.00

Table J3. Grade 5 ELA 2010 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
495	210	0.11	210	0.11
592	246	0.12	456	0.23
606	344	0.17	800	0.41
613	480	0.24	1280	0.65
618	673	0.34	1953	0.99
623	784	0.40	2737	1.39
626	910	0.46	3647	1.85
629	1133	0.57	4780	2.42
632	1323	0.67	6103	3.09
634	1557	0.79	7660	3.88
637	1964	1.00	9624	4.88
639	2298	1.17	11922	6.05

(Continued on next page)

Table J3. Grade 5 ELA 2010 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
642	2863	1.45	14785	7.50
644	3518	1.78	18303	9.28
646	4460	2.26	22763	11.54
648	5240	2.66	28003	14.20
651	6581	3.34	34584	17.54
653	7937	4.02	42521	21.56
656	9721	4.93	52242	26.49
659	11728	5.95	63970	32.44
661	13802	7.00	77772	39.44
665	15781	8.00	93553	47.44
668	17754	9.00	111307	56.44
673	19775	10.03	131082	66.47
678	20780	10.54	151862	77.01
686	19996	10.14	171858	87.15
700	16577	8.41	188435	95.56
795	8765	4.44	197200	100.00

Table J4. Grade 6 ELA 2010 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
480	122	0.06	122	0.06
591	126	0.06	248	0.13
603	210	0.11	458	0.23
609	329	0.17	787	0.40
614	437	0.22	1224	0.62
618	594	0.30	1818	0.92
621	661	0.33	2479	1.25
623	760	0.38	3239	1.64
626	830	0.42	4069	2.06
628	963	0.49	5032	2.54
630	1078	0.54	6110	3.09
632	1168	0.59	7278	3.68
633	1437	0.73	8715	4.40

(Continued on next page)

Table J4. Grade 6 ELA 2010 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
635	1534	0.78	10249	5.18
637	1804	0.91	12053	6.09
638	2022	1.02	14075	7.11
640	2403	1.21	16478	8.33
641	2732	1.38	19210	9.71
643	3154	1.59	22364	11.30
644	3701	1.87	26065	13.17
646	4413	2.23	30478	15.40
647	5116	2.59	35594	17.99
649	5961	3.01	41555	21.00
651	6774	3.42	48329	24.43
653	8157	4.12	56486	28.55
655	9745	4.93	66231	33.48
657	11410	5.77	77641	39.24
659	12862	6.50	90503	45.74
662	14861	7.51	105364	53.26
665	16342	8.26	121706	61.52
669	17440	8.81	139146	70.33
673	16706	8.44	155852	78.77
678	15519	7.84	171371	86.62
684	12915	6.53	184286	93.15
694	8990	4.54	193276	97.69
785	4569	2.31	197845	100.00

Table J5. Grade 7 ELA 2010 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
470	154	0.08	154	0.08
584	133	0.07	287	0.14
597	240	0.12	527	0.26
604	320	0.16	847	0.42
609	457	0.23	1304	0.65

(Continued on next page)

Table J5. Grade 7 ELA 2010 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
613	496	0.25	1800	0.90
616	628	0.31	2428	1.21
619	700	0.35	3128	1.56
621	762	0.38	3890	1.95
623	890	0.45	4780	2.39
626	947	0.47	5727	2.86
628	1147	0.57	6874	3.44
629	1229	0.61	8103	4.05
631	1443	0.72	9546	4.77
633	1602	0.80	11148	5.58
635	1878	0.94	13026	6.51
637	2247	1.12	15273	7.64
638	2500	1.25	17773	8.89
640	2916	1.46	20689	10.35
642	3488	1.74	24177	12.09
643	4013	2.01	28190	14.10
645	4523	2.26	32713	16.36
647	5228	2.61	37941	18.98
649	6015	3.01	43956	21.98
651	6722	3.36	50678	25.35
653	7603	3.80	58281	29.15
655	8704	4.35	66985	33.50
657	9734	4.87	76719	38.37
660	10865	5.43	87584	43.80
662	12150	6.08	99734	49.88
665	13247	6.63	112981	56.51
669	14747	7.38	127728	63.88
673	16047	8.03	143775	71.91
678	17009	8.51	160784	80.41
685	16804	8.40	177588	88.82
698	14359	7.18	191947	96.00
790	7996	4.00	199943	100.00

Table J6. Grade 8 ELA 2010 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
430	130	0.06	130	0.06
515	112	0.05	242	0.12
566	125	0.06	367	0.18
578	215	0.11	582	0.29
585	332	0.16	914	0.45
590	386	0.19	1300	0.64
594	485	0.24	1785	0.87
597	567	0.28	2352	1.15
601	645	0.32	2997	1.47
604	757	0.37	3754	1.84
606	818	0.40	4572	2.24
609	945	0.46	5517	2.70
611	1018	0.50	6535	3.20
613	1107	0.54	7642	3.74
616	1303	0.64	8945	4.38
618	1411	0.69	10356	5.07
620	1593	0.78	11949	5.86
622	1850	0.91	13799	6.76
624	2100	1.03	15899	7.79
626	2373	1.16	18272	8.95
628	2826	1.38	21098	10.34
630	3222	1.58	24320	11.92
632	3636	1.78	27956	13.70
634	4282	2.10	32238	15.80
637	4946	2.42	37184	18.22
639	5871	2.88	43055	21.10
641	6693	3.28	49748	24.38
644	7916	3.88	57664	28.26
646	9086	4.45	66750	32.71
649	9997	4.90	76747	37.61
652	11292	5.53	88039	43.14
655	11830	5.80	99869	48.94
658	12674	6.21	112543	55.15
661	13118	6.43	125661	61.57

(Continued on next page)

Table J6. Grade 8 ELA 2010 SS Frequency Distribution, State (cont.)

SS	N-count	Percent	Cumulative Frequency	Cumulative Percent
665	13348	6.54	139009	68.11
669	13440	6.59	152449	74.70
673	12810	6.28	165259	80.98
679	11881	5.82	177140	86.80
686	11231	5.50	188371	92.30
699	10304	5.05	198675	97.35
790	5405	2.65	204080	100.00

References

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

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika* 37:29–51.

Bock, R.D. and M. Aitkin. 1981. Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika* 46:443–459.

Burket, G.R. (1988). *ITEMWIN* [Computer program].

Burket, G.R. (2002). *PARDUX* [Computer program].

Cattell, R.B. (1966). The Scree Test for the Number of Factors. *Multivariate Behavioral Research* 1:245–276.

CTB/McGraw-Hill (1996). TerraNova™ Assessment Series (1st Ed.). Monterey, CA: CTB/McGraw-Hill.

CTB/McGraw-Hill (2000). TerraNova™ Assessment Series (2nd Ed.) Monterey, CA: CTB/McGraw-Hill.

CTB/McGraw-Hill (2006). TerraNova™ Assessment Series (3rd Ed.) Monterey, CA: CTB/McGraw-Hill.

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

Dorans, N.J., A.P. Schmitt & C.A. Bleistein (1992). The standardization approach to assessing comprehensive differential item functioning. *Journal of Educational Measurement* 29:309–319.

Fitzpatrick, A.R. (1990). *Status report on the results of preliminary analysis of dichotomous and multi-level items using the PARMATE program.*

Fitzpatrick, A.R. (1994). *Two studies comparing parameter estimates produced by PARDUX and BIGSTEPS.*

Fitzpatrick, A.R. & M.W. Julian (1996). *Two studies comparing the parameter estimates produced by PARDUX and PARSCLE*. Monterey, CA: CTB/McGraw-Hill**.**

Fitzpatrick, A.R., V. Link, W. M. Yen, G. Burket, K. Ito & R. Sykes (1996). Scaling performance assessments: A comparison between one-parameter and two-parameter partial credit models. *Journal of Educational Measurement* 33:291–314.

Green, D.R., W.M. Yen & G.R. Burket (1989). Experiences in the application of item response theory in test construction. *Applied Measurement in Education* 2:297–312.

Huynh, H. & C. Schneider (2004). *Vertically moderated standards as an alternative to vertical scaling: assumptions, practices, and an odyssey through NAEP*. Paper presented at the National Conference on Large-Scale Assessment. Boston, MA, June 21.

Jensen, A.R. (1980). *Bias in mental testing*. New York: Free Press.

Johnson, N.L. & S. Kotz (1970). *Distributions in Statistics: Continuous Univariate Distributions,* Vol. 2. New York: John Wiley.

Kim, D. (2004). WLCLASS [Computer program].

Kolen, M.J. & R.L. Brennan (1995). *Test Equating: Methods and Practices.* New York: Springer-Verlag.

Lee, W., B.A. Hanson & R.L. Brennan (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement* 26:412–432.

Linn, R.L. (1991). Linking results of distinct assessments. *Applied Measurement in Education* 6 (1):83–102.

Linn, R.L. & D. Harnisch (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measuremen*t 18:109–118.

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

Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: Lawrence Erlbaum.

Lord, F.M. & M.R. Novick (1968). *Statistical Theories of Mental Test Scores.* Menlo Park, CA: Addison-Wesley.

Mehrens, W.A. & I.J. Lehmann (1991). *Measurement and Evaluation in Education and Psychology, 3rd ed*. New York: Holt, Rinehart, and Winston.

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

Muraki, E. & R.D. Bock (1991). *PARSCALE*: *Parameter Scaling of Rating Data* [Computer program]. Chicago, IL: Scientific Software, Inc.

Novick, M.R. & P.H. Jackson (1974). *Statistical Methods for Educational and Psychological Research.* New York: McGraw-Hill.

Qualls, A.L. (1995). Estimating the reliability of a test containing multiple-item formats. *Applied Measurement in Education* 8:111–120.

Reckase, M.D. (1979). Unifactor latent trait models applied to multifactor tests: results and implications. *Journal of Educational Statistics* 4:207–230.

Sandoval, J.H. & M.P. Mille (1979) *Accuracy of judgments of WISC-R item difficulty for minority groups*. Paper presented at the annual meeting of the American Psychological Association, New York. August.

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

Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika* 47:175–186.

Wang, T.M., J. Kolen & D.J. Harris (2000). Psychometric properties of scale scores and performance levels for performance assessment using polytomous IRT. *Journal of Educational Measurement* 37:141–162.

Wright, B.D. & J. M. Linacre. (1992). *BIGSTEPS Rasch Analysis* [Computer program]. Chicago, IL: MESA Press.

Yen, W.M. (1997). The technical quality of performance assessments: Standard errors of percents of students reaching standards. *Educational Measurement: Issues and Practice*: 5–15.

Yen, W.M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement* 30:187–213.

Yen, W. M. (1984). Obtaining maximum likelihood trait estimates from number correct scores for the three-parameter logistic model. *Journal of Educational Measurement* 21: 93–111.

Yen, W.M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement* 5:245–262.

Yen, W.M., R.C. Sykes, K. Ito & M. Julian (1997). *A Bayesian/IRT index of objective performance for tests with mixed-item types*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL, March.

Zwick, R., J.R. Donoghue & A. Grima (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement* 36:225–33