New York State Testing Program 2009: English Language Arts, Grades 3–8

Technical Report

Submitted
November 2009

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) 2009 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 2009, 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 in 2009. 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 for Public and Nonpublic Schools* (SAM), available online at http://www.emsc.nysed.gov/osa/sam/gr3-8ela-08.pdf

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 (Not Meeting Learning Standards), Level II (Partially Meeting Learning Standards), Level III (Meeting Learning Standards), and Level IV (Meeting Learning Standards with Distinction). The proficiency cut scores used to distinguish among Levels I, II, III, and IV were established during the process of Standard Setting. There is reason to believe and evidence to support the claim that New York State ELA proficiency cut scores reflect the abilities intended by the New York State Education Department. Performance of students on the Grades 3–8 ELA Tests in relation to proficiency level cut scores is reported in a form of performance level classification. The performances of schools, districts, and the State are reported as percentages of students in each performance level. Detailed information on a process of establishing 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 questions onto scannable answer sheets; and the

teachers transcribe the responses to the constructed-response 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 student 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 January 2009 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/testing/elaei/09exams/home.htm. 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 2009 *Teacher's Directions* available online (http://www.emsc.nysed.gov/osa/elaei/ela-td-3-5.pdf and http://www.emsc.nysed.gov/osa/elaei/ela-td-6-8.pdf) as well as the 2009 *School Administrator's Manual* (http://www.emsc.nysed.gov/osa/sam/ela-sam-09.pdf) provide details on security, scheduling, classroom organization and preparation, test materials, and administration.

Table 1. NYSTP ELA 2009 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 2009 Test Configuration (cont.)

Grade	Day	Book	Number of Items			Allotted Time (minutes)	
			MC	CR*	Total**	Testing	Prep
7	1	1	26	2	28	55	10
	2	2	4	3	7	30	15
	Totals		30	5	35	85	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 constructed-response 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 2009 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 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 2009 OP tests.

Table 3. NYSTP ELA 2009 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	9	33.0	30.0
			S2	14	16	47.0	53.0
			S3	6	5	20.0	17.0
4	39	3	S1	13	11	36.0	31.0
			S2	16	17	44.5	47.0
			S3	7	8	19.5	22.0
5	31	3	S1	12	14	43.0	50.0
			S2	10	9	36.0	32.0
			S3	6	5	21.0	18.0
6	39	3	S1	13	10	36.0	28.0
			S2	16	16	44.5	44.0
			S3	7	10	19.5	28.0
7	41	3	S1	15	17	39.0	45.0
			S2	15	13	39.0	34.0
			S3	8	8	22.0	21.0
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 2009 Operational Test Map, Grade 3

Question	Type	Points	Standard	Performance Indicator	Answer Key
Book 1	**Reading**				
1	Multiple Choice	1	1	Determine the meaning of unfamiliar words by using context clues, dictionaries, and other classroom resources	B
2	Multiple Choice	1	1	Identify main ideas and supporting details in informational texts	C
3	Multiple Choice	1	1	Read unfamiliar texts to collect and interpret data, facts, and ideas	C
4	Multiple Choice	1	3	Evaluate the content by identifying important and unimportant details	D
5	Multiple Choice	1	1	Identify main ideas and supporting details in informational texts	A
6	Multiple Choice	1	2	Summarize main ideas and supporting details from imaginative texts	D
7	Multiple Choice	1	2	Use specific evidence from stories to describe characters, their actions, and their motivations; relate sequences of events	C
8	Multiple Choice	1	2	Use specific evidence from stories to describe characters, their actions, and their motivations; relate sequences of events	B
9	Multiple Choice	1	2	Make predictions, draw conclusions, and make inferences about events and characters	D
10	Multiple Choice	1	1	Read and understand written directions	C
11	Multiple Choice	1	1	Read and understand written directions	D
12	Multiple Choice	1	1	Read and understand written directions	B
13	Multiple Choice	1	1	Determine the meaning of unfamiliar words by using context clues, dictionaries, and other classroom resources	A
14	Multiple Choice	1	1	Use graphic organizers to record significant details from informational texts	B
15	Multiple Choice	1	3	Evaluate the content by identifying important and unimportant details	C
16	Multiple Choice	1	2	Summarize main ideas and supporting details from imaginative texts	B
17	Multiple Choice	1	2	Use knowledge of story structure, story elements, and key vocabulary to interpret stories	D
18	Multiple Choice	1	2	Use specific evidence from stories to describe characters, their actions, and their motivations; relate sequences of events	A
19	Multiple Choice	1	3	Evaluate the content by identifying important and unimportant details	B
20	Multiple Choice	1	3	Evaluate the content by identifying the author's purpose	D

(Continued on next page)

Table 4a. NYSTP ELA 2009 Operational Test Map, Grade 3 (cont.)

Question	Type	Points	Standard	Performance Indicator	Answer Key
Book 1	**Reading**				
21	Short Response	2	2	Use specific evidence from stories to describe characters, their actions, and their motivations; relate sequences of events	n/a
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	B
23	Multiple Choice	1	3	Distinguish between fact and opinion	A
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	2	Identify elements of character, plot, and setting to understand the author's message or intent	D
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 2009 Operational Test Map, Grade 4

Question	Type	Points	Standard	Performance Indicator	Answer Key
Book 1	**Reading**				
1	Multiple Choice	1	1	Identify a main idea and supporting details in informational texts	D
2	Multiple Choice	1	1	Collect and interpret data, facts, and ideas from unfamiliar texts	A
3	Multiple Choice	1	1	Identify a main idea and supporting details in informational texts	B
4	Multiple Choice	1	1	Collect and interpret data, facts, and ideas from unfamiliar texts	C
5	Multiple Choice	1	3	Evaluate the content by identifying whether events, actions, characters, and/or settings are realistic	D
6	Multiple Choice	1	1	Identify a conclusion that summarizes the main idea	D
7	Multiple Choice	1	3	Evaluate the content by identifying the author's purpose	A
8	Multiple Choice	1	2	Make predictions, draw conclusions, and make inferences about events and characters	A

(Continued on next page)

Table 4b. NYSTP ELA 2009 Operational Test Map, Grade 4 (cont.)

Question	Type	Points	Standard	Performance Indicator	Answer Key
Book 1	**Reading**				
9	Multiple Choice	1	2	Make predictions, draw conclusions, and make inferences about events and characters	D
10	Multiple Choice	1	2	Use graphic organizers to record significant details about characters and events in stories	B
11	Multiple Choice	1	3	Evaluate the content by identifying important and unimportant details	A
12	Multiple Choice	1	3	Evaluate the content by identifying whether events, actions, characters, and/or settings are realistic	B
13	Multiple Choice	1	1	Identify a main idea and supporting details in informational texts	C
14	Multiple Choice	1	1	Understand written directions and procedures	B
15	Multiple Choice	1	1	Understand written directions and procedures	A
16	Multiple Choice	1	1	Understand written directions and procedures	D
17	Multiple Choice	1	1	Collect and interpret data, facts, and ideas from unfamiliar texts	B
18	Multiple Choice	1	1	Identify a conclusion that summarizes the main idea	C
19	Multiple Choice	1	2	Make predictions, draw conclusions, and make inferences about events and characters	C
20	Multiple Choice	1	2	Use specific evidence from stories to identify themes; describe characters, their actions and motivations; relate a sequence of events	A
21	Multiple Choice	1	2	Determine the meaning of unfamiliar words by using context clues, dictionaries, and other classroom resources	A
22	Multiple Choice	1	2	Use specific evidence from stories to identify themes; describe characters, their actions and motivations; relate a sequence of events	B
23	Multiple Choice	1	2	Use knowledge of story structure, story elements, and key vocabulary to interpret stories	D
24	Multiple Choice	1	2	Use specific evidence from stories to identify themes; describe characters, their actions and motivations; relate a sequence of events	B
25	Multiple Choice	1	2	Use knowledge of story structure, story elements, and key vocabulary to interpret stories	A
26	Multiple Choice	1	2	Use specific evidence from stories to identify themes; describe characters, their actions and motivations; relate a sequence of events	B
27	Multiple Choice	1	2	Make predictions, draw conclusions, and make inferences about events and characters	A
28	Multiple Choice	1	2	Make predictions, draw conclusions, and make inferences about events and characters	D

(Continued on next page)

Table 4b. NYSTP ELA 2009 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 2009 Operational Test Map, Grade 5

Question	Type	Points	Standard	Performance Indicator	Answer Key
Book 1	**Reading**				
1	Multiple Choice	1	2	Identify literary elements, such as setting, plot, and character, of different genres	A
2	Multiple Choice	1	2	Identify literary elements, such as setting, plot, and character, of different genres	D
3	Multiple Choice	1	2	Identify literary elements, such as setting, plot, and character, of different genres	B
4	Multiple Choice	1	2	Identify literary elements, such as setting, plot, and character, of different genres	C
5	Multiple Choice	1	2	Determine the meaning of unfamiliar words by using context clues, a dictionary, or a glossary	A
6	Multiple Choice	1	1	Read to collect and interpret data, facts, and ideas from multiple sources	A
7	Multiple Choice	1	1	Read to collect and interpret data, facts, and ideas from multiple sources	D
8	Multiple Choice	1	1	Read to collect and interpret data, facts, and ideas from multiple sources	B
9	Multiple Choice	1	1	Identify information that is implied rather than stated	C
10	Multiple Choice	1	3	Evaluate information, ideas, opinions, and themes in texts by identifying a central idea and supporting details	B
11	Multiple Choice	1	3	Evaluate information, ideas, opinions, and themes in texts by identifying a central idea and supporting details	D

(Continued on next page)

Table 4c. NYSTP ELA 2009 Operational Test Map, Grade 5 (cont.)

Question	Type	Points	Standard	Performance Indicator	Answer Key
Book 1	**Reading**				
12	Multiple Choice	1	1	Recognize organizational formats to assist in comprehension of informational text	C
13	Multiple Choice	1	1	Identify information that is implied rather than stated	B
14	Multiple Choice	1	1	Read to collect and interpret data, facts, and ideas from multiple sources	B
15	Multiple Choice	1	1	Distinguish between fact and opinion	D
16	Multiple Choice	1	2	Identify literary elements, such as setting, plot, and character, of different genres	A
17	Multiple Choice	1	2	Read, view, and interpret literary texts from a variety of genres	C
18	Multiple Choice	1	2	Identify literary elements, such as setting, plot, and character, of different genres	D
19	Multiple Choice	1	3	Evaluate information, ideas, opinions, and themes in texts by identifying a central idea and supporting details	C
20	Multiple Choice	1	2	Define the characteristics of different genres	C
21	Multiple Choice	2	3	Evaluate information, ideas, opinions, and themes in texts by identifying a central idea and supporting details	n/a
22	Multiple Choice	1	1	Identify essential details for note taking	A
23	Multiple Choice	1	1	Identify essential details for note taking	A
24	Multiple Choice	1	1	Identify information that is implicit rather than stated	C
25	Multiple Choice	1	1	Identify essential details for note taking	C
26	Short Response	2	1	Identify essential details for note taking	n/a
27	Editing Paragraph	3	n/a	Observe the rules of punctuation, capitalization, and spelling; use correct grammatical construction	n/a

Table 4d. NYSTP ELA 2009 Operational Test Map, Grade 6

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

(Continued on next page)

Table 4d. NYSTP ELA 2009 Operational Test Map, Grade 6 (cont.)

Question	Type	Points	Standard	Performance Indicator	Answer Key
Book 1	**Reading**				
21	Multiple Choice	1	1	Identify missing, conflicting, unclear, and irrelevant information	C
22	Multiple Choice	1	2	Read, view, and interpret literary texts from a variety of genres	A
23	Multiple Choice	1	2	Identify literary elements (e.g., setting, plot, character, rhythm, and rhyme) of different genres	C
24	Multiple Choice	1	2	Determine the meaning of unfamiliar words by using context clues, a dictionary, or a glossary	C
25	Multiple Choice	1	2	Identify literary elements (e.g., setting, plot, character, rhythm, and rhyme) of different genres	A
26	Multiple Choice	1	3	Evaluate information, ideas, opinions, and themes by identifying a central idea and supporting details	A
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 2009 Operational Test Map, Grade 7

Question	Type	Points	Standard	Performance Indicator	Answer Key
Book 1	**Reading**				
1	Multiple Choice	1	1	Interpret data, facts, and ideas from informational texts by applying thinking skills, such as define, classify, and infer	C
2	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	C
3	Multiple Choice	1	1	Identify a purpose for reading	D

(Continued on next page)

Table 4e. NYSTP ELA 2009 Operational Test Map, Grade 7 (cont.)

Question	Type	Points	Standard	Performance Indicator	Answer Key
Book 1	**Reading**				
4	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	C
5	Multiple Choice	1	2	Interpret characters, plot, setting, and theme, using evidence from the text	A
6	Multiple Choice	1	2	Identify the author's point of view, such as first-person narrator and omniscient narrator	B
7	Multiple Choice	1	2	Interpret characters, plot, setting, and theme, using evidence from the text	B
8	Multiple Choice	1	2	Interpret characters, plot, setting, and theme, using evidence from the text	D
9	Multiple Choice	1	3	Identify multiple levels of meaning	D
10	Multiple Choice	1	2	Recognize how the author's use of language creates images or feelings	C
11	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	B
12	Multiple Choice	1	1	Interpret data, facts, and ideas from informational texts by applying thinking skills, such as define, classify, and infer	D
13	Multiple Choice	1	1	Interpret data, facts, and ideas from informational texts by applying thinking skills, such as define, classify, and infer	C
14	Multiple Choice	1	1	Distinguish between relevant and irrelevant information	D
15	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
16	Multiple Choice	1	1	Draw conclusions and make inferences on the basis of explicit and implied information	B
17	Multiple Choice	1	1	Interpret data, facts, and ideas from informational texts by applying thinking skills, such as define, classify, and infer	A
18	Multiple Choice	1	1	Distinguish between relevant and irrelevant information	D
19	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	A

(Continued on next page)

Table 4e. NYSTP ELA 2009 Operational Test Map, Grade 7 (cont.)

Question	Type	Points	Standard	Performance Indicator	Answer Key
Book 1	**Reading**				
20	Multiple Choice	1	1	Identify a purpose for reading	C
21	Multiple Choice	1	2	Interpret characters, plot, setting, and theme, using evidence from the text	B
22	Multiple Choice	1	2	Interpret characters, plot, setting, and theme, using evidence from the text	C
23	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	C
24	Multiple Choice	1	2	Recognize how the author's use of language creates images or feelings	A
25	Multiple Choice	1	2	Determine the meaning of unfamiliar words by using context clues, a dictionary, a glossary, and structural analysis	B
26	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	C
27	Short Response	2	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	n/a
28	Short Response	2	2	Interpret characters, plot, setting, and theme, using evidence from the text	n/a
Book 2	**Listening and Writing**				
29	Multiple Choice	1	1	Recall significant ideas and details, and describe relationships between and among them	D
30	Multiple Choice	1	1	Draw conclusions and make inferences on the basis of explicit information	C
31	Multiple Choice	1	1	Recall significant ideas and details, and describe relationships between and among them	C
32	Multiple Choice	1	1	Make, confirm, or revise predictions	D
33	Short Response	2	1	Recall significant ideas and details, and describe relationships between and among them	n/a
34	Short Response	2	1	Draw conclusions and make inferences on the basis of explicit information	n/a
35	Editing Paragraph	3	n/a	Observe rules of punctuation, capitalization, and spelling; use correct grammatical construction	n/a

Table 4f. NYSTP ELA 2009 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	D
2	Multiple Choice	1	2	Interpret characters, plot, setting, theme, and dialogue, using evidence from the text	A
3	Multiple Choice	1	2	Interpret characters, plot, setting, theme, and dialogue, using evidence from the text	A
4	Multiple Choice	1	2	Determine how the use and meaning of literary devices, such as symbolism, metaphor and simile, alliteration, personification, flashback, and foreshadowing, convey the author's message or intent	C
5	Multiple Choice	1	2	Recognize how the author's use of language creates images or feelings	D
6	Multiple Choice	1	1	Identify missing, conflicting, or unclear information	D
7	Multiple Choice	1	1	Draw conclusions and make inferences on the basis of explicit and implied information	D
8	Multiple Choice	1	1	Apply thinking skills, such as define, classify, and infer, to interpret data, facts, and ideas from informational texts	B
9	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)	B
10	Multiple Choice	1	1	Make, confirm, or revise predictions	A
11	Multiple Choice	1	2	Interpret characters, plot, setting, theme, and dialogue, using evidence from the text	C
12	Multiple Choice	1	2	Determine how the use and meaning of literary devices, such as symbolism, metaphor and simile, alliteration, personification, flashback, and foreshadowing, convey the author's message or intent	A
13	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
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	D
15	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)	B
16	Multiple Choice	1	2	Interpret characters, plot, setting, theme, and dialogue, using evidence from the text	C

(Continued on next page)

Table 4f. NYSTP ELA 2009 Operational Test Map, Grade 8 (cont.)

Question	Type	Points	Standard	Performance Indicator	Answer Key
Book 1	**Reading**				
17	Multiple Choice	1	2	Determine how the use and meaning of literary devices, such as symbolism, metaphor and simile, alliteration, personification, flashback, and foreshadowing, convey the author's message or intent	A
18	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	B
19	Multiple Choice	1	2	Determine how the use and meaning of literary devices, such as symbolism, metaphor and simile, alliteration, personification, flashback, and foreshadowing, convey the author's message or intent	A
20	Multiple Choice	1	2	Recognize how the author's use of language creates images or feelings	C
21	Multiple Choice	1	1	Apply thinking skills, such as define, classify, and infer, to interpret data, facts, and ideas from informational texts	D
22	Multiple Choice	1	1	Draw conclusions and make inferences on the basis of explicit and implied information	D
23	Multiple Choice	1	1	Identify missing, conflicting, or unclear information	C
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	3	Evaluate the validity and accuracy of information, ideas, themes, opinions, and experiences in texts to identify cultural and ethnic values and their impact on content	A
26	Multiple Choice	1	1	Identify a purpose for reading	B
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

2009 Item Mapping by New York State Standards

Table 5. NYSTP ELA 2009 Standard Coverage

Grade	Standard	MC Item #s	CR Item #s	Total Items	Total Points
3	S1	1, 2, 3, 5, 10, 11, 12, 13, 14	n/a	9	9
3	S2	6, 7, 8, 9, 16, 17, 18, 22, 24, 25	21, 26, 27	13	16
3	S3	4, 15, 19, 20, 23	n/a	5	5
4	S1	1, 2, 3, 4, 6, 13, 14, 15, 16, 17, 18	n/a	11	11
4	S2	8, 9, 10, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28	29	14	17
4	S3	5, 7, 11, 12	30	5	8
5	S1	6, 7, 8, 9, 12, 13, 14, 15, 22, 23, 24, 25	26	13	14
5	S2	1, 2, 3, 4, 5, 16, 17, 18, 20	n/a	9	9
5	S3	10, 11, 19	21	4	5
6	S1	6, 7, 8, 9, 16, 17, 18, 19, 20, 21	n/a	10	10
6	S2	1, 2, 3, 4, 12, 13, 14, 22, 23, 24, 25	27	12	16
6	S3	5, 10, 11, 15, 26	28	6	10
7	S1	1, 3, 12, 13, 14, 16, 17, 18, 20, 29, 30, 31, 32	33, 34	15	17
7	S2	5, 6, 7, 8, 10, 11, 21, 22, 23, 24, 25	28	12	13
7	S3	2, 4, 9, 15, 19, 26	27	7	8
8	S1	6, 7, 8, 9, 10, 21, 22, 23, 26	27	10	14
8	S2	1, 2, 3, 4, 5, 11, 12, 13, 15, 16, 17, 19, 20	n/a	13	13
8	S3	14, 18, 24, 25	28	5	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 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 constructed-response 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 2009 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 2007. 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 22–25, 2008, and the Grades 6-8 ELA FTs were administered during January 28–31, 2008, 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 2008 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 fourth year of OP NYSTP Grades 3–8 ELA Tests were administered in January 2009. The test items were selected from the pool of items primarily field-tested in 2006, 2007, and 2008, 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 when 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 2008 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 January 2009.

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 sites:

- http://www.emsc.nysed.gov/osa/sam/ela-sam-09.pdf
- http://www.nysedregents.org/testing/elaei/09exams/home.htm

Proficiency and Performance Standards

Proficiency cut score recommendations and the drafting of performance standards occurred at the NYSTP ELA standard setting review held in Albany in June 2006. The results were reviewed by a measurement review committee and were approved in August 2006. For each grade level, there are four proficiency levels. Three cut points demarcate the performance standards needed to demonstrate each ascending level of proficiency. For details on the standard setting method, participants, achievement levels, and results (impact), refer to the *Bookmark Standard Setting Technical Report 2006 for Grades 3, 4, 5, 6, 7, and 8 English Language Arts* and *New York State ELA Measurement Review Technical Report 2006 for English Language Arts*.

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 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 VII, "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 3 students from districts classified as charter and Grade 5 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 item 27 and 28 in Grade 6 test and item 27 in Grade 8 test, all other items on the 2009 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 principle components accounting for more than 10 percent 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**	**6.48**	**23.13**	**23.13**
	2	1.23	4.39	27.52
	3	1.09	3.89	31.40
4	**1**	**7.04**	**22.71**	**22.71**
	2	1.31	4.21	26.93
	3	1.03	3.32	30.24
5	**1**	**5.65**	**20.93**	**20.93**
	2	1.15	4.27	25.20
	3	1.07	3.96	29.16
	4	1.02	3.76	32.92
6	**1**	**6.64**	**22.89**	**22.89**
	2	1.26	4.34	27.22
	3	1.09	3.75	30.97
	4	1.04	3.57	34.54

(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**	**7.92**	**4.14**	**26.76**
	2	1.45	3.08	29.83
	3	1.08	2.94	32.78
	4	1.03	2.94	32.78
8	**1**	**6.33**	**21.81**	**21.81**
	2	1.11	3.84	25.65
	3	1.03	3.56	29.21

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 an 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 as disability 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 field test 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 Scoring Leader Handbooks* and *School Administrator's Manual* (SAM). In addition, please refer to the *Scoring Site Operations Manual* (2009) located at http://www.emsc.nysed.gov/3-8/archived.htm#scoring.

Test Administration

NYSTP Grades 3–8 ELA Tests were administered at the classroom level during January 2009. The testing window for Grades 3, 4, and 5 was January 12–16. The testing window for Grades 6, 7, and 8 was January 20–23. The makeup test administration window for Grades 3, 4, and 5 was January 20–23, and for Grades 6, 7, and 8, it was January 26–30. 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 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 2008–09 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.emsc.nysed.gov/3-8/update-rev-dec05.htm (see Attachment C).

Scoring of Constructed-Response Items

The scoring of constructed-response 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 field test. 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 2008, 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, a DVD was 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, the ELA Frequently Asked Questions (FAQs) document, and a DVD 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 DVDs 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, ELA FAQs, and DVD, 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. To affirm that all schools across the state adhered to scoring guidelines and policies, approximately five percent 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 February 2009. These data were used for all data analysis. Phase 2 involved submitting "straggler files" to CTB/McGraw-Hill in early-March 2009. 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	0	196918
Out of grade	195	196723
No grade	0	196723
Duplicate record	0	196723
Non-public and out-of-district schools	2179	194544
Missing values for ALL items on OP form	1	194543
Out-of-range CR scores	0	194543

Table 7b. NYSTP ELA Grade 4 Data Cleaning

Exclusion Rule	# Deleted	# Cases Remain
Initial # of cases	0	203666
Out of grade	147	203519
No grade	0	203519
Duplicate record	0	203519
Non-public and out-of-district schools	11240	192279
Missing values for ALL items on OP form	4	192275
Out-of-range CR scores	0	192275

Table 7c. NYSTP ELA Grade 5 Data Cleaning

Exclusion Rule	# Deleted	# Cases Remain
Initial # of cases	0	195325
Out of grade	120	195205
No grade	0	195205
Duplicate record	0	195205
Non-public and out-of-district schools	2031	193174
Missing values for ALL items on OP form	1	193173
Out-of-range CR scores	0	193173

Table 7d. NYSTP ELA Grade 6 Data Cleaning

Exclusion Rule	# Deleted	# Cases Remain
Initial # of cases	0	201715
Out of grade	126	201589
No grade	0	201589
Duplicate record	0	201589
Non-public and out-of-district schools	4573	197016
Missing values for ALL items on OP form	6	197010
Out-of-range CR scores	0	197010

Table 7e. NYSTP ELA Grade 7 Data Cleaning

Exclusion Rule	# Deleted	# Cases Remain
Initial # of cases	0	203828
Out of grade	174	203654
No grade	2	203652
Duplicate record	0	203652
Non-public and out-of-district schools	2171	201481
Missing values for ALL items on OP form	0	201481
Out-of-range CR scores	0	201481

Table 7f. NYSTP ELA Grade 8 Data Cleaning

Exclusion Rule	# Deleted	# Cases Remain
Initial # of cases	0	218453
Out of grade	220	218233
No grade	4	218229
Duplicate record	0	218229
Non-public and out-of-district schools	12300	205929
Missing values for ALL items on OP form	1	205928
Out-of-range CR scores	0	205928

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 = 194543)

Demographic Category		N-count	% of Total N-count
NRC	NYC	67964	34.97
	Big cities	8112	4.17
	Urban/Suburban	16072	8.27
	Rural	11464	5.90
	Average needs	58413	30.06
	Low needs	29895	15.38
	Charter	2420	1.25
Ethnicity	Asian	15648	8.04
	Black	35850	18.43
	Hispanic	40919	21.03
	American Indian	918	0.47
	Multi-Racial	624	0.32
	White	100486	51.65
	Unknown	98	0.05
ELL	No	177406	91.19
	Yes	17137	8.81
SWD	No	168938	86.84
	Yes	25605	13.16
SUA	No	150880	77.56
	Yes	43663	22.44

Table 8b. Grade 4 Sample Characteristics (N = 192275)

Demographic Category		N-count	% of Total N-count
NRC	NYC	66407	34.58
	Big cities	7889	4.11
	Urban/Suburban	15710	8.18
	Rural	11446	5.96
	Average needs	58826	30.63
	Low needs	29748	15.49
	Charter	2022	1.05
Ethnicity	Asian	14442	7.51
	Black	35861	18.65
	Hispanic	40296	20.96
	American Indian	905	0.47
	Multi-Racial	471	0.24
	White	100219	52.12
	Unknown	81	0.04
ELL	No	178195	92.68
	Yes	14080	7.32
SWD	No	164716	85.67
	Yes	27559	14.33
SUA	No	148147	77.05
	Yes	44128	22.95

Table 8c. Grade 5 Sample Characteristics (N = 193173)

Demographic Category		N-count	% of Total N-count
NRC	NYC	66242	34.34
	Big cities	7471	3.87
	Urban/Suburban	15460	8.01
	Rural	11455	5.94
	Average needs	59926	31.07
	Low needs	30561	15.84
	Charter	1783	0.92
Ethnicity	Asian	14363	7.44
	Black	35763	18.51
	Hispanic	39978	20.70
	American Indian	937	0.49
	Multi-Racial	467	0.24
	White	101566	52.58
	Unknown	99	0.05
ELL	No	181190	93.80
	Yes	11983	6.20
SWD	No	164025	84.91
	Yes	29148	15.09
SUA	No	148786	77.02
	Yes	44387	22.98

Table 8d. Grade 6 Sample Characteristics (N = 197010)

Demographic Category		N-count	% of Total N-count
NRC	NYC	67827	34.48
	Big cities	7422	3.77
	Urban/Suburban	15036	7.64
	Rural	11388	5.79
	Average needs	60880	30.95
	Low needs	30948	15.73
	Charter	3198	1.63
Ethnicity	Asian	14703	7.46
	Black	37889	19.23
	Hispanic	40321	20.47
	American Indian	901	0.46
	Multi-Racial	428	0.22
	White	102695	52.13
	Unknown	73	0.04
ELL	No	186663	94.75
	Yes	10347	5.25
SWD	No	166920	84.73
	Yes	30090	15.27
SUA	No	155509	78.93
	Yes	41501	21.07

Table 8e. Grade 7 Sample Characteristics (N = 201481)

Demographic Category		N-count	% of Total N-count
NRC	NYC	69201	34.42
	Big cities	7632	3.80
	Urban/Suburban	15260	7.59
	Rural	12037	5.99
	Average needs	62778	31.23
	Low needs	31715	15.78
	Charter	2421	1.20
Ethnicity	Asian	14788	7.34
	Black	38093	18.91
	Hispanic	40905	20.30
	American Indian	929	0.46
	Multi-Racial	388	0.19
	White	106298	52.76
	Unknown	80	0.04
ELL	No	191921	95.26
	Yes	9560	4.74
SWD	No	170703	84.72
	Yes	30778	15.28
SUA	No	161200	80.01
	Yes	40281	19.99

Table 8f. Grade 8 Sample Characteristics (N = 205928)

Demographic Category		N-count	% of Total N-count
NRC	NYC	71114	34.63
	Big cities	7643	3.72
	Urban/Suburban	15455	7.53
	Rural	12293	5.99
	Average needs	64777	31.54
	Low needs	31949	15.56
	Charter	2134	1.04
Ethnicity	Asian	14937	7.25
	Black	39164	19.02
	Hispanic	41597	20.20
	American Indian	1007	0.49
	Multi-Racial	290	0.14
	White	108866	52.87
	Unknown	67	0.03
ELL	No	196243	95.30
	Yes	9685	4.70
SWD	No	175802	85.37
	Yes	30126	14.63
SUA	No	165374	80.31
	Yes	40554	19.69

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 VII, "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 each 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 percentage 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 0.40 to 0.98. For Grade 3, the item p-values were between 0.40 and 0.98 with a mean of 0.79. For Grade 4, the item p-values were between 0.48 and 0.93 with a mean of 0.76. For Grade 5, the item p-values were between 0.56 and 0.95 with a mean of 0.81. For Grade 6, the item p-values were between 0.52 and 0.97 with a mean of 0.83. For Grade 7, the item p-values were between 0.45 and 0.97 with a mean of 0.81. For Grade 8, the item p-values were between 0.53 and 0.94 with a mean of 0.77. These mean 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	194543	0.86	0.04	0.04	10.38	*85.83	2.19	1.52	-0.38	*0.44	-0.14	-0.13	0.44
2	194543	0.71	0.09	0.07	14.83	6.04	*70.63	8.34	-0.22	-0.18	*0.45	-0.29	0.45
3	194543	0.89	0.15	0.08	4.33	3.91	*88.89	2.64	-0.36	-0.25	*0.51	-0.21	0.51
4	194543	0.40	0.23	0.10	11.50	36.12	12.12	*39.93	-0.10	0.01	-0.15	*0.17	0.17
5	194543	0.83	0.31	0.12	*82.71	3.01	7.12	6.73	*0.44	-0.24	-0.21	-0.27	0.44
6	194543	0.92	0.09	0.14	1.64	2.16	3.75	*92.23	-0.25	-0.23	-0.18	*0.38	0.38
7	194543	0.86	0.10	0.07	3.19	3.65	*86.40	6.60	-0.30	-0.23	*0.40	-0.16	0.40
8	194543	0.90	0.11	0.09	4.40	*89.81	2.39	3.19	-0.26	*0.47	-0.26	-0.27	0.47
9	194543	0.55	0.13	0.17	9.11	22.45	13.61	*54.53	-0.23	-0.11	-0.01	*0.24	0.24
10	194543	0.58	0.16	0.09	20.91	11.91	*57.84	9.09	-0.23	-0.13	*0.40	-0.20	0.40
11	194543	0.79	0.18	0.13	4.81	10.78	4.76	*79.35	-0.24	-0.21	-0.25	*0.43	0.43
12	194543	0.84	0.17	0.04	4.34	*84.29	5.42	5.74	-0.26	*0.48	-0.27	-0.25	0.48
13	194543	0.93	0.21	0.05	*93.24	2.82	1.67	2.02	*0.45	-0.25	-0.24	-0.26	0.45
14	194543	0.76	0.44	0.08	7.73	*76.28	6.40	9.08	-0.32	*0.47	-0.23	-0.17	0.47
15	194543	0.59	0.61	0.10	19.92	9.91	*59.06	10.39	-0.14	-0.14	*0.33	-0.18	0.33
16	194543	0.75	0.25	0.10	7.00	*75.26	3.45	13.93	-0.23	*0.45	-0.27	-0.23	0.45
17	194543	0.88	0.32	0.23	4.75	4.61	2.53	*87.56	-0.26	-0.26	-0.27	*0.48	0.48
18	194543	0.81	0.38	0.08	*80.89	6.28	7.05	5.32	*0.53	-0.25	-0.31	-0.27	0.53
19	194543	0.78	0.43	0.09	10.18	*77.78	2.73	8.78	-0.31	*0.44	-0.23	-0.16	0.44
20	194543	0.67	0.85	0.03	9.01	4.60	18.61	*66.90	-0.27	-0.21	-0.15	*0.41	0.41
21	194543	0.65	1.51	15.29	36.40	46.80							
22	194543	0.89	0.05	0.03	5.28	*88.89	5.25	0.49	-0.20	*0.31	-0.18	-0.18	0.31
23	194543	0.74	0.06	0.06	*73.91	6.62	13.22	6.14	*0.29	-0.13	-0.16	-0.15	0.29
24	194543	0.91	0.08	0.05	*90.85	1.14	3.24	4.63	*0.48	-0.16	-0.27	-0.34	0.48
25	194543	0.90	0.11	0.01	2.33	1.06	6.13	*90.35	-0.25	-0.15	-0.18	*0.33	0.33
26	194543	0.98	0.21	0.67	3.19	95.94							
27	194543	0.80	0.26	4.56	30.40	64.78							
28	194543	0.86	0.18	5.59	5.96	14.21	74.05						

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	192275	0.77	0.02	0.03	19.01	2.56	1.35	*77.04	-0.24	-0.26	-0.16	*0.37	0.37
2	192275	0.93	0.02	0.02	*93.27	4.31	0.87	1.51	*0.36	-0.25	-0.18	-0.17	0.36
3	192275	0.89	0.03	0.04	1.57	*89.06	2.88	6.40	-0.14	*0.38	-0.14	-0.31	0.38
4	192275	0.53	0.09	0.06	9.07	10.15	*52.53	28.10	-0.16	-0.14	*0.26	-0.08	0.26
5	192275	0.77	0.05	0.08	6.09	9.14	7.48	*77.17	-0.08	-0.25	-0.25	*0.38	0.38
6	192275	0.72	0.07	0.07	3.68	17.35	7.29	*71.53	-0.18	-0.16	-0.16	*0.31	0.31
7	192275	0.87	0.06	0.04	*87.48	2.02	7.43	2.98	*0.36	-0.19	-0.22	-0.19	0.36
8	192275	0.93	0.06	0.11	*92.82	1.88	1.46	3.67	*0.31	-0.21	-0.17	-0.16	0.31
9	192275	0.87	0.10	0.17	2.05	4.66	6.06	*86.95	-0.21	-0.18	-0.28	*0.41	0.41
10	192275	0.86	0.09	0.04	3.28	*86.04	7.22	3.33	-0.29	*0.50	-0.27	-0.28	0.50
11	192275	0.79	0.08	0.06	*79.37	4.24	10.85	5.40	*0.49	-0.26	-0.29	-0.23	0.49
12	192275	0.83	0.07	0.04	4.97	*83.48	6.20	5.23	-0.24	*0.47	-0.28	-0.23	0.47
13	192275	0.83	0.14	0.05	7.60	6.19	*82.52	3.50	-0.31	-0.20	*0.46	-0.23	0.46
14	192275	0.88	0.12	0.04	7.19	*87.73	2.21	2.71	-0.25	*0.41	-0.21	-0.23	0.41
15	192275	0.92	0.13	0.06	*92.15	1.70	4.41	1.54	*0.39	-0.22	-0.23	-0.22	0.39
16	192275	0.64	0.17	0.08	16.77	4.50	14.11	*64.37	-0.10	-0.27	-0.21	*0.35	0.35
17	192275	0.81	0.21	0.03	6.03	*80.67	9.99	3.07	-0.23	*0.48	-0.33	-0.19	0.48
18	192275	0.81	0.22	0.04	8.23	5.20	*81.32	5.00	-0.28	-0.26	*0.49	-0.24	0.49
19	192275	0.77	0.30	0.06	5.00	3.63	*76.79	14.22	-0.20	-0.19	*0.33	-0.16	0.33
20	192275	0.78	0.32	0.04	*78.01	2.59	3.98	15.07	*0.33	-0.20	-0.18	-0.18	0.33
21	192275	0.90	0.33	0.05	*89.84	3.76	2.74	3.28	*0.42	-0.19	-0.26	-0.23	0.42
22	192275	0.68	0.43	0.08	6.57	*68.35	18.47	6.10	-0.14	*0.32	-0.15	-0.18	0.32
23	192275	0.65	0.57	0.21	18.27	5.43	10.74	*64.78	-0.14	-0.13	-0.08	*0.25	0.25
24	192275	0.50	1.00	0.10	13.84	*50.49	27.94	6.64	-0.21	*0.38	-0.10	-0.22	0.38
25	192275	0.68	1.16	0.12	*68.20	4.77	6.37	19.38	*0.41	-0.25	-0.20	-0.18	0.41
26	192275	0.48	1.30	0.08	27.30	*47.90	13.51	9.91	-0.19	*0.43	-0.19	-0.17	0.43
27	192275	0.64	1.46	0.05	*63.56	13.08	11.24	10.61	*0.42	-0.25	-0.16	-0.18	0.42
28	192275	0.81	1.55	0.02	3.45	6.17	7.52	*81.29	-0.27	-0.23	-0.21	*0.44	0.44
29	192275	0.65	0.08	1.20	10.60	31.29	41.81	15.02					
30	192275	0.63	0.10	1.62	12.87	32.44	37.26	15.71					
31	192275	0.68	0.14	1.93	20.87	46.62	30.44						

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	193173	0.91	0.01	0.01	*90.68	3.83	1.15	4.31	*0.33	-0.26	-0.16	-0.14	0.33
2	193173	0.88	0.02	0.03	2.06	6.37	3.07	*88.44	-0.19	-0.23	-0.23	*0.39	0.39
3	193173	0.87	0.05	0.01	2.09	*87.42	7.99	2.43	-0.18	*0.47	-0.33	-0.27	0.47
4	193173	0.92	0.05	0.02	2.45	3.93	*92.27	1.29	-0.27	-0.26	*0.43	-0.19	0.43
5	193173	0.74	0.07	0.03	*74.32	11.08	10.46	4.04	*0.45	-0.20	-0.31	-0.17	0.45
6	193173	0.90	0.04	0.05	*89.82	3.60	3.23	3.25	*0.39	-0.20	-0.17	-0.27	0.39
7	193173	0.75	0.10	0.06	9.90	9.53	4.92	*75.48	-0.20	-0.23	-0.19	*0.40	0.40
8	193173	0.56	0.09	0.03	30.72	*55.60	11.19	2.37	0.01	*0.13	-0.14	-0.14	0.13
9	193173	0.65	0.09	0.03	24.42	4.79	*64.63	6.04	-0.07	-0.17	*0.26	-0.23	0.26
10	193173	0.67	0.07	0.04	21.50	*66.95	9.13	2.33	-0.17	*0.34	-0.18	-0.22	0.34
11	193173	0.93	0.09	0.03	2.50	1.46	2.65	*93.27	-0.21	-0.21	-0.26	*0.41	0.41
12	193173	0.75	0.07	0.02	11.28	9.04	*75.22	4.37	-0.22	-0.18	*0.40	-0.26	0.40
13	193173	0.88	0.10	0.03	2.91	*88.07	4.92	3.98	-0.23	*0.36	-0.21	-0.16	0.36
14	193173	0.78	0.11	0.05	5.02	*78.04	7.55	9.23	-0.29	*0.49	-0.28	-0.22	0.49
15	193173	0.85	0.11	0.11	4.85	3.71	5.87	*85.35	-0.28	-0.26	-0.25	*0.49	0.49
16	193173	0.85	0.31	0.03	*84.90	6.73	3.06	4.97	*0.31	-0.11	-0.20	-0.21	0.31
17	193173	0.83	0.27	0.02	5.11	9.98	*83.47	1.15	-0.24	-0.21	*0.38	-0.18	0.38
18	193173	0.90	0.35	0.06	3.16	4.21	2.22	*90.00	-0.24	-0.32	-0.25	*0.49	0.49
19	193173	0.82	0.45	0.03	2.49	11.52	*82.05	3.45	-0.17	-0.27	*0.41	-0.20	0.41
20	193173	0.64	0.64	0.01	12.80	17.89	*63.87	4.79	-0.08	-0.14	*0.28	-0.23	0.28
21	193173	0.77	0.96	6.03	31.12	61.89							
22	193173	0.95	0.04	0.00	*95.46	1.13	2.32	1.04	*0.30	-0.10	-0.23	-0.16	0.30
23	193173	0.88	0.05	0.01	*87.56	8.70	1.19	2.49	*0.25	-0.15	-0.16	-0.14	0.25
24	193173	0.90	0.06	0.01	0.83	1.63	*90.08	7.39	-0.10	-0.19	*0.28	-0.19	0.28
25	193173	0.88	0.10	0.02	3.57	3.99	*87.74	4.58	-0.23	-0.10	*0.27	-0.12	0.27
26	193173	0.88	0.11	2.71	18.69	78.49							
27	193173	0.61	0.22	13.84	20.41	34.03	31.50						

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	% at 5	Pbis Option 1	Pbis Option 2	Pbis Option 3	Pbis Option 4	Pbis Key
1	197010	0.83	0.03	0.01	*82.91	0.89	9.37	6.79		*0.19	-0.18	-0.11	-0.09	0.19
2	197010	0.97	0.02	0.03	0.91	0.92	0.75	*97.36		-0.09	-0.11	-0.14	*0.20	0.20
3	197010	0.93	0.04	0.01	2.01	*92.62	1.20	4.12		-0.19	*0.35	-0.19	-0.21	0.35
4	197010	0.90	0.05	0.01	3.19	3.88	*89.65	3.22		-0.18	-0.31	*0.43	-0.22	0.43
5	197010	0.87	0.05	0.02	*87.40	5.69	2.22	4.63		*0.36	-0.19	-0.24	-0.19	0.36
6	197010	0.96	0.04	0.02	1.52	*95.66	1.67	1.10		-0.21	*0.34	-0.19	-0.18	0.34
7	197010	0.87	0.04	0.04	*87.37	2.52	3.90	6.14		*0.42	-0.21	-0.21	-0.27	0.42
8	197010	0.81	0.06	0.03	8.40	4.96	5.26	*81.31		-0.23	-0.25	-0.24	*0.44	0.44
9	197010	0.85	0.06	0.02	10.07	2.13	*85.41	2.32		-0.18	-0.21	*0.34	-0.22	0.34
10	197010	0.91	0.05	0.03	*91.20	3.19	2.09	3.44		*0.28	-0.13	-0.19	-0.15	0.28
11	197010	0.84	0.09	0.02	2.71	4.30	*84.05	8.83		-0.23	-0.12	*0.23	-0.07	0.23
12	197010	0.94	0.07	0.05	1.53	1.08	3.75	*93.52		-0.25	-0.20	-0.28	*0.43	0.43
13	197010	0.77	0.09	0.02	6.06	10.39	*77.06	6.38		-0.21	-0.17	*0.43	-0.30	0.43
14	197010	0.93	0.06	0.01	2.25	*92.63	3.40	1.64		-0.23	*0.43	-0.26	-0.23	0.43
15	197010	0.92	0.11	0.03	1.92	*92.23	2.67	3.04		-0.21	*0.42	-0.28	-0.21	0.42
16	197010	0.86	0.13	0.04	2.80	8.72	1.93	*86.38		-0.24	-0.38	-0.23	*0.53	0.53
17	197010	0.81	0.16	0.04	7.26	7.51	*81.17	3.87		-0.23	-0.18	*0.4	-0.25	0.40
18	197010	0.64	0.12	0.04	4.21	1.25	*64.40	29.98		-0.32	-0.23	*0.31	-0.12	0.31
19	197010	0.89	0.14	0.03	4.01	4.16	*89.15	2.51		-0.27	-0.29	*0.48	-0.23	0.48
20	197010	0.77	0.16	0.05	16.89	2.89	3.20	*76.80		-0.20	-0.29	-0.25	*0.41	0.41
21	197010	0.52	0.20	0.07	5.12	19.23	*51.94	23.43		-0.22	-0.12	*0.33	-0.16	0.33
22	197010	0.88	0.39	0.05	*88.07	6.68	1.72	3.09		*0.45	-0.29	-0.22	-0.22	0.45
23	197010	0.87	0.40	0.03	2.45	7.22	*87.25	2.65		-0.19	-0.34	*0.43	-0.12	0.43
24	197010	0.57	0.43	0.04	23.22	11.91	*57.44	6.96		-0.11	-0.18	*0.33	-0.20	0.33
25	197010	0.86	0.50	0.03	*85.93	6.75	3.45	3.34		*0.44	-0.26	-0.26	-0.20	0.44
26	197010	0.81	0.55	0.02	*80.56	3.56	7.58	7.74		*0.51	-0.21	-0.30	-0.28	0.51
27	197010	0.73	0.07	0.36	2.55	9.98	27.47	35.96	23.62					
28	197010	0.67	0.10	0.60	5.47	16.14	30.92	30.13	16.65					
29	197010	0.75	0.09	0.89	13.51	45.00	40.51							

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	201481	0.93	0.04	0.01	2.75	1.31	*93.29	2.61	-0.18	-0.19	*0.33	-0.20	0.33
2	201481	0.92	0.02	0.01	2.72	4.48	*91.79	0.97	-0.35	-0.22	*0.43	-0.17	0.43
3	201481	0.85	0.06	0.02	5.30	5.58	3.94	*85.09	-0.27	-0.17	-0.25	*0.42	0.42
4	201481	0.84	0.05	0.02	5.30	7.81	*83.97	2.86	-0.28	-0.34	*0.51	-0.20	0.51
5	201481	0.94	0.07	0.02	*93.94	2.49	2.13	1.36	*0.41	-0.25	-0.23	-0.21	0.41
6	201481	0.94	0.04	0.01	2.43	*94.01	2.47	1.05	-0.22	*0.43	-0.33	-0.16	0.43
7	201481	0.97	0.05	0.02	1.27	*96.63	0.77	1.27	-0.14	*0.27	-0.12	-0.18	0.27
8	201481	0.89	0.06	0.02	2.68	6.71	1.87	*88.66	-0.19	-0.29	-0.27	*0.44	0.44
9	201481	0.87	0.05	0.03	5.10	4.70	2.84	*87.27	-0.32	-0.20	-0.25	*0.46	0.46
10	201481	0.86	0.08	0.01	5.88	4.31	*85.74	3.98	-0.34	-0.28	*0.54	-0.26	0.54
11	201481	0.49	0.12	0.03	9.31	*49.17	28.28	13.09	-0.26	*0.33	-0.06	-0.18	0.33
12	201481	0.82	0.10	0.02	7.75	3.71	6.13	*82.29	-0.30	-0.34	-0.23	*0.53	0.53
13	201481	0.87	0.07	0.03	6.58	2.86	*86.65	3.81	-0.28	-0.29	*0.50	-0.26	0.50
14	201481	0.90	0.10	0.05	1.83	3.44	4.17	*90.40	-0.24	-0.28	-0.32	*0.51	0.51
15	201481	0.45	0.12	0.01	27.66	*45.46	13.01	13.74	-0.10	*0.27	-0.16	-0.09	0.27
16	201481	0.87	0.11	0.02	5.57	*87.19	5.14	1.97	-0.12	*0.18	-0.07	-0.11	0.18
17	201481	0.92	0.10	0.05	*91.76	3.02	2.09	2.98	*0.38	-0.22	-0.17	-0.22	0.38
18	201481	0.82	0.14	0.06	4.04	10.51	3.48	*81.77	-0.34	-0.28	-0.27	*0.53	0.53
19	201481	0.84	0.15	0.04	*83.67	10.51	3.11	2.53	*0.43	-0.24	-0.27	-0.23	0.43
20	201481	0.92	0.15	0.03	2.47	3.72	*92.09	1.54	-0.31	-0.26	*0.48	-0.22	0.48
21	201481	0.90	0.21	0.03	1.31	*89.78	1.38	7.30	-0.18	*0.29	-0.18	-0.17	0.29
22	201481	0.87	0.28	0.02	3.41	5.50	*86.69	4.10	-0.30	-0.24	*0.46	-0.20	0.46
23	201481	0.82	0.27	0.02	5.54	9.25	*82.20	2.72	-0.28	-0.25	*0.45	-0.19	0.45
24	201481	0.90	0.30	0.04	*89.87	4.86	2.62	2.31	*0.48	-0.28	-0.27	-0.25	0.48
25	201481	0.51	0.40	0.06	9.97	*50.85	1.83	36.89	-0.12	*0.22	-0.23	-0.08	0.22
26	201481	0.65	0.60	0.01	16.18	6.66	*64.55	12.00	-0.18	-0.23	*0.39	-0.17	0.39
27	201481	0.66	1.53	11.62	41.42	45.44							
28	201481	0.68	3.02	9.97	38.20	48.81							
29	201481	0.66	0.13	0.02	12.51	13.80	7.57	*65.97	-0.21	-0.30	-0.08	*0.41	0.41
30	201481	0.95	0.12	0.01	0.93	2.50	*94.57	1.88	-0.18	-0.15	*0.31	-0.19	0.31
31	201481	0.72	0.13	0.02	3.65	22.48	*72.46	1.26	-0.23	-0.25	*0.37	-0.14	0.37
32	201481	0.97	0.14	0.01	1.40	0.63	0.75	*97.07	-0.13	-0.12	-0.15	*0.24	0.24
33	201481	0.80	0.12	2.22	34.70	62.97							
34	201481	0.81	0.25	4.93	28.53	66.29							
35	201481	0.52	0.35	20.87	23.19	33.84	21.75						

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	% at 5	Pbis Option 1	Pbis Option 2	Pbis Option 3	Pbis Option 4	Pbis Key
1	205928	0.91	0.02	0.01	3.24	4.62	1.27	*90.83		-0.32	-0.17	-0.16	*0.38	0.38
2	205928	0.86	0.03	0.02	*86.11	1.87	9.30	2.67		*0.38	-0.20	-0.22	-0.23	0.38
3	205928	0.93	0.03	0.04	*92.99	0.72	4.29	1.94		*0.32	-0.13	-0.21	-0.19	0.32
4	205928	0.84	0.05	0.01	5.13	4.82	*84.01	5.98		-0.17	-0.19	*0.36	-0.23	0.36
5	205928	0.56	0.08	0.02	27.73	5.72	9.99	*56.46		-0.03	-0.24	-0.23	*0.28	0.28
6	205928	0.91	0.04	0.02	3.11	4.24	1.78	*90.81		-0.18	-0.12	-0.17	*0.27	0.27
7	205928	0.88	0.06	0.02	2.93	6.18	2.52	*88.28		-0.26	-0.27	-0.26	*0.47	0.47
8	205928	0.75	0.08	0.02	9.31	*74.84	8.13	7.63		-0.08	*0.32	-0.21	-0.22	0.32
9	205928	0.74	0.06	0.02	21.79	*73.70	2.74	1.69		-0.24	*0.37	-0.24	-0.16	0.37
10	205928	0.94	0.04	0.02	*94.15	1.75	2.85	1.19		*0.4	-0.23	-0.25	-0.18	0.40
11	205928	0.81	0.11	0.02	1.51	9.84	*80.83	7.69		-0.20	-0.18	*0.32	-0.18	0.32
12	205928	0.89	0.08	0.03	*89.32	7.86	1.43	1.29		*0.33	-0.17	-0.25	-0.22	0.33
13	205928	0.86	0.06	0.02	3.26	5.88	*85.88	4.90		-0.23	-0.24	*0.46	-0.29	0.46
14	205928	0.62	0.10	0.03	7.50	5.47	24.58	*62.32		-0.19	-0.20	-0.10	*0.29	0.29
15	205928	0.62	0.07	0.03	5.13	*62.09	10.48	22.20		-0.17	*0.30	-0.12	-0.17	0.30
16	205928	0.84	0.09	0.03	1.59	6.61	*84.17	7.51		-0.24	-0.23	*0.33	-0.12	0.33
17	205928	0.71	0.10	0.04	*71.47	5.04	15.37	7.97		*0.23	-0.20	-0.07	-0.13	0.23
18	205928	0.71	0.09	0.03	5.21	*71.04	5.76	17.88		-0.18	*0.30	-0.23	-0.11	0.30
19	205928	0.53	0.09	0.03	*52.69	19.67	13.69	13.83		*0.39	-0.21	-0.19	-0.13	0.39
20	205928	0.70	0.09	0.04	4.45	17.20	*70.14	8.08		-0.22	-0.11	*0.33	-0.24	0.33
21	205928	0.85	0.19	0.03	5.90	4.92	4.44	*84.52		-0.31	-0.27	-0.29	*0.54	0.54
22	205928	0.84	0.19	0.03	7.16	5.43	3.55	*83.65		-0.27	-0.29	-0.28	*0.52	0.52
23	205928	0.76	0.21	0.05	8.79	7.79	*76.13	7.02		-0.15	-0.23	*0.42	-0.28	0.42
24	205928	0.77	0.27	0.04	4.07	9.93	*76.70	9.00		-0.27	-0.31	*0.46	-0.14	0.46
25	205928	0.53	0.30	0.05	*52.52	6.39	19.00	21.74		*0.24	-0.27	-0.05	-0.07	0.24
26	205928	0.69	0.30	0.02	9.12	*68.84	13.64	8.08		-0.25	*0.37	-0.10	-0.22	0.37
27	205928	0.70	0.18	0.94	4.94	12.84	27.57	31.75	21.78					
28	205928	0.71	0.14	0.63	4.00	11.54	27.52	34.70	21.47					
29	205928	0.76	0.22	1.35	12.56	43.63	42.23							

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 8, which had a point biserial of 0.13. Point biserials for correct answer options (pbis*) on the tests ranged 0.13–0.56. For Grade 3, the pbis* were between 0.17 and 0.53. For Grade 4, the pbis* were between 0.25 and 0.50. For Grade 5, the pbis* were between 0.13 and 0.49. For Grade 6, pbis* were between 0.19 and 0.53. For Grade 7, the pbis* were between 0.18 and 0.54. For Grade 8, the pbis* were between 0.23 and 0.54.

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 4 in Grade 3 and item number 8 in Grade 5, which both had a point biserial of 0.01. 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 2009 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 0.83–0.88. Feldt-Raju reliability coefficients ranged 0.85–0.89. 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 4 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 VII, "Reliability and Standard Error of Measurement").

Table 10. NYSTP ELA 2009 Test Form Statistics and Reliability

Grade	Max RS	RS Mean	RS SD	P-value Mean	Cronbach's alpha	Feldt-Raju
3	33	26.15	5.59	0.79	0.86	0.87
4	39 (43 WGT)	28.71 (31.43 WGT)	6.76 (7.50 WGT)	0.74	0.88	0.89
5	31	24.84	4.99	0.80	0.83	0.85
6	39	31.06	5.99	0.80	0.86	0.88
7	41	32.31	6.67	0.79	0.88	0.89
8	39 (44 WGT)	29.36 (32.90 WGT)	6.53 (7.43 WGT)	0.75	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 2009 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	35850	40919	15648	102126	95227	99316	103612	88308	17137	177406
4	35861	40296	14442	101676	93884	98391	101452	88574	14080	178195
5	35763	39978	14363	103069	95029	98144	100628	90487	11983	181190
6	37889	40321	14703	104097	96263	100747	101673	91828	10347	186663
7	38093	40905	14788	107695	98144	103337	104130	94493	9560	191921
8	39164	41597	14937	110230	101082	104846	106505	96726	9685	196243

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	0
4	4
5	4
6	5
7	5
8	7

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).

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 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 2009 samples were comparable to 2008 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	2008 Grade 3 Population	2009 Grade 3 Sample	2008 Grade 4 Population	2009 Grade 4 Sample
	%	%	%	%
NRC SUBGROUPS				
NYC	35.34	34.94	35.39	34.54
Big cities	4.14	4.17	3.91	4.10
Urban/Suburban	8.09	8.26	7.95	8.17
Rural	5.86	5.89	5.84	5.95
Average needs	29.90	30.03	30.19	30.59
Low needs	15.02	15.37	15.38	15.47
Charter	1.54	1.24	1.22	1.05
Missing	0.11	0.10	0.12	0.12
ETHNICITY				
Asian	7.31	8.04	7.20	7.51
Black	19.21	18.43	19.28	18.65
Hispanics	21.17	21.03	20.91	20.96
American Indian	0.50	0.47	0.48	0.47
Multi-Racial	0.12	0.32	0.10	0.24
White	51.67	51.65	51.98	52.12
Unknown	0.03	0.05	0.04	0.04
ELL STATUS				
No	91.53	91.19	92.89	92.68
Yes	8.47	8.81	7.11	7.32
DISABILITY				
No	86.71	86.84	85.33	85.67
Yes	13.29	13.16	14.67	14.33
ACCOMMODATIONS				
No	80.06	77.56	79.14	77.05
Yes	19.94	22.44	20.86	22.95

Table 14. Grades 5 and 6 Demographic Statistics

Demographics	2008 Grade 5 Population	2009 Grade 5 Sample	2008 Grade 6 Population	2009 Grade 6 Sample
	%	%	%	%
NRC SUBGROUPS				
NYC	34.95	34.29	34.49	34.43
Big cities	3.80	3.87	3.77	3.77
Urban/Suburban	7.68	8.00	7.64	7.63
Rural	5.74	5.93	5.88	5.78
Average needs	30.51	31.02	30.93	30.90
Low needs	15.52	15.82	15.75	15.71
Charter	1.67	0.92	1.38	1.62
Missing	0.13	0.14	0.16	0.16
ETHNICITY				
Asian	7.36	7.44	7.25	7.46
Black	19.35	18.51	18.95	19.23
Hispanics	20.50	20.7	20.25	20.47
American Indian	0.46	0.49	0.46	0.46
Multi-Racial	0.08	0.24	0.08	0.22
White	52.20	52.58	52.97	52.13
Unknown	0.04	0.05	0.04	0.04
ELL STATUS				
No	94.25	93.8	95.10	94.75
Yes	5.75	6.20	4.90	5.25
DISABILITY				
No	85.03	84.91	84.83	84.73
Yes	14.97	15.09	15.17	15.27
ACCOMMODATIONS				
No	79.73	77.02	80.77	78.93
Yes	20.27	22.98	19.23	21.07

Table 15. Grades 7 and 8 Demographic Statistics

Demographics	2008 Grade 7 Population	2009 Grade 7 Sample	2008 Grade 8 Population	2009 Grade 8 Sample
	%	%	%	%
NRC SUBGROUPS				
NYC	34.47	34.35	34.21	34.53
Big cities	3.80	3.79	3.99	3.71
Urban/Suburban	7.60	7.57	7.64	7.51
Rural	6.04	5.97	6.29	5.97
Average needs	31.42	31.16	31.67	31.46
Low needs	15.37	15.74	15.20	15.51
Charter	1.11	1.20	0.68	1.04
Missing	0.20	0.22	0.31	0.27
ETHNICITY				
Asian	7.01	7.34	6.83	7.25
Black	19.27	18.91	19.33	19.02
Hispanics	20.00	20.30	19.58	20.20
American Indian	0.49	0.46	0.51	0.49
Multi-Racial	0.06	0.19	0.06	0.14
White	53.14	52.76	53.67	52.87
Unknown	0.03	0.04	0.03	0.03
ELL STATUS				
No	95.53	95.26	95.88	95.30
Yes	4.47	4.74	4.12	4.70
DISABILITY				
No	85.35	84.72	85.54	85.37
Yes	14.65	15.28	14.46	14.63
ACCOMMODATIONS				
No	81.60	80.01	81.80	80.31
Yes	18.40	19.99	18.20	19.69

.

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. These default values of the *c*-parameter were obtained during the FT calibration and were held unchanged between field test and OP administrations. For the Grades 3–8 ELA tests, all calibration estimation results are reasonable. The summary of calibration results is presented in Table 16.

Table 16. NYSTP ELA 2009 Calibration Results

Grade	Largest *a*-parameter	*b*-parameter/ Gamma Range		# Items with Default *c*-parameter	Theta Mean	Theta Standard Deviation	# Students
3	2.228	-3.864	1.137	13	0.12	1.304	194543
4	2.498	-3.290	1.037	18	0.04	1.165	192275
5	2.474	-3.671	0.237	15	0.17	1.385	193173
6	2.387	-4.043	0.349	15	0.10	1.268	197010
7	2.636	-4.063	1.078	14	0.08	1.218	201481
8	2.484	-3.771	0.293	11	0.06	1.182	205928

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 2009 ELA Tests for Grades 3, 4, 5, and 7 demonstrated good model fit. Items 27 and 28 in Grade 6 and item 27 in Grade 8 exhibited poor item-model fit statistics. The fact that so few items were flagged for poor fit across all ELA Tests further supports the use of the chosen models. Item fit statistics are presented in Tables 17–22.

Table 17. ELA Grade 3 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	Z observed	Z-critical	Fit OK?
1	3PL	48.71	7	187943	11.15	501.18	Y
2	3PL	189.98	7	187943	48.90	501.18	Y
3	3PL	140.75	7	187943	35.75	501.18	Y
4	3PL	698.13	7	187943	184.71	501.18	Y
5	3PL	97.96	7	187943	24.31	501.18	Y
6	3PL	92.41	7	187943	22.83	501.18	Y
7	3PL	225.27	7	187943	58.34	501.18	Y
8	3PL	187.25	7	187943	48.17	501.18	Y
9	3PL	316.35	7	187943	82.68	501.18	Y
10	3PL	789.65	7	187943	209.17	501.18	Y
11	3PL	202.36	7	187943	52.21	501.18	Y
12	3PL	168.24	7	187943	43.09	501.18	Y
13	3PL	233.20	7	187943	60.46	501.18	Y
14	3PL	338.25	7	187943	88.53	501.18	Y
15	3PL	716.34	7	187943	189.58	501.18	Y
16	3PL	253.63	7	187943	65.91	501.18	Y
17	3PL	150.06	7	187943	38.24	501.18	Y
18	3PL	536.59	7	187943	141.54	501.18	Y
19	3PL	142.45	7	187943	36.20	501.18	Y
20	3PL	566.29	7	187943	149.48	501.18	Y
21	2PPC	2419.67	17	187943	412.05	501.18	Y
22	3PL	84.45	7	187943	20.70	501.18	Y
23	3PL	423.54	7	187943	111.33	501.18	Y
24	3PL	113.02	7	187943	28.33	501.18	Y
25	3PL	63.88	7	187943	15.20	501.18	Y
26	2PPC	459.67	17	187943	75.92	501.18	Y
27	2PPC	564.28	17	187943	93.86	501.18	Y
28	2PPC	327.18	26	187943	41.77	501.18	Y

Table 18. ELA Grade 4 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	Z observed	Z-critical	Fit OK?
1	3PL	1229.34	7	189703	326.68	505.88	Y
2	3PL	79.29	7	189703	19.32	505.88	Y
3	3PL	52.08	7	189703	12.05	505.88	Y
4	3PL	84.39	7	189703	20.68	505.88	Y
5	3PL	333.97	7	189703	87.39	505.88	Y
6	3PL	390.27	7	189703	102.43	505.88	Y
7	3PL	50.18	7	189703	11.54	505.88	Y
8	3PL	209.65	7	189703	54.16	505.88	Y
9	3PL	301.92	7	189703	78.82	505.88	Y
10	3PL	329.94	7	189703	86.31	505.88	Y
11	3PL	114.51	7	189703	28.73	505.88	Y
12	3PL	123.53	7	189703	31.14	505.88	Y
13	3PL	92.16	7	189703	22.76	505.88	Y
14	3PL	408.00	7	189703	107.17	505.88	Y
15	3PL	217.26	7	189703	56.19	505.88	Y
16	3PL	42.60	7	189703	9.52	505.88	Y
17	3PL	358.78	7	189703	94.02	505.88	Y
18	3PL	170.37	7	189703	43.66	505.88	Y
19	3PL	406.88	7	189703	106.87	505.88	Y
20	3PL	67.50	7	189703	16.17	505.88	Y
21	3PL	360.03	7	189703	94.35	505.88	Y
22	3PL	68.62	7	189703	16.47	505.88	Y
23	3PL	307.71	7	189703	80.37	505.88	Y
24	3PL	203.89	7	189703	52.62	505.88	Y
25	3PL	140.43	7	189703	35.66	505.88	Y
26	3PL	1020.24	7	189703	270.80	505.88	Y
27	3PL	225.07	7	189703	58.28	505.88	Y
28	3PL	452.20	7	189703	118.98	505.88	Y
29	2PPC	2109.92	35	189703	248.00	505.88	Y
30	2PPC	2969.11	35	189703	350.69	505.88	Y
31	2PPC	936.04	26	189703	126.20	505.88	Y

Table 19. ELA Grade 5 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	Z observed	Z-critical	Fit OK?
1	3PL	338.37	7	184365	88.56	491.64	Y
2	3PL	85.15	7	184365	20.89	491.64	Y
3	3PL	240.95	7	184365	62.53	491.64	Y
4	3PL	173.50	7	184365	44.50	491.64	Y
5	3PL	536.44	7	184365	141.50	491.64	Y
6	3PL	109.66	7	184365	27.44	491.64	Y
7	3PL	388.26	7	184365	101.90	491.64	Y
8	3PL	1353.02	7	184365	359.74	491.64	Y
9	3PL	774.71	7	184365	205.18	491.64	Y
10	3PL	173.11	7	184365	44.39	491.64	Y
11	3PL	101.64	7	184365	25.29	491.64	Y
12	3PL	526.41	7	184365	138.82	491.64	Y
13	3PL	127.86	7	184365	32.30	491.64	Y
14	3PL	371.31	7	184365	97.37	491.64	Y
15	3PL	366.59	7	184365	96.10	491.64	Y
16	3PL	18.51	7	184365	3.08	491.64	Y
17	3PL	195.14	7	184365	50.28	491.64	Y
18	3PL	286.60	7	184365	74.73	491.64	Y
19	3PL	574.44	7	184365	151.66	491.64	Y
20	3PL	1198.06	7	184365	318.32	491.64	Y
21	2PPC	1115.66	17	184365	188.42	491.64	Y
22	3PL	156.20	7	184365	39.87	491.64	Y
23	3PL	130.67	7	184365	33.05	491.64	Y
24	3PL	38.38	7	184365	8.39	491.64	Y
25	3PL	54.59	7	184365	12.72	491.64	Y
26	2PPC	474.58	17	184365	78.47	491.64	Y
27	2PPC	2442.19	26	184365	335.06	491.64	Y

Table 20. ELA Grade 6 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	Z observed	Z-critical	Fit OK?
1	3PL	538.06	7	190378	141.93	507.68	Y
2	3PL	108.58	7	190378	27.15	507.68	Y
3	3PL	94.32	7	190378	23.34	507.68	Y
4	3PL	67.00	7	190378	16.04	507.68	Y
5	3PL	318.81	7	190378	83.34	507.68	Y
6	3PL	159.10	7	190378	40.65	507.68	Y
7	3PL	259.30	7	190378	67.43	507.68	Y
8	3PL	195.68	7	190378	50.43	507.68	Y
9	3PL	97.76	7	190378	24.26	507.68	Y
10	3PL	206.67	7	190378	53.36	507.68	Y
11	3PL	1856.47	7	190378	494.29	507.68	Y
12	3PL	111.27	7	190378	27.87	507.68	Y
13	3PL	162.02	7	190378	41.43	507.68	Y
14	3PL	89.81	7	190378	22.13	507.68	Y
15	3PL	248.93	7	190378	64.66	507.68	Y
16	3PL	232.33	7	190378	60.22	507.68	Y
17	3PL	97.52	7	190378	24.19	507.68	Y
18	3PL	392.07	7	190378	102.91	507.68	Y
19	3PL	283.79	7	190378	73.97	507.68	Y
20	3PL	244.85	7	190378	63.57	507.68	Y
21	3PL	1261.69	7	190378	335.33	507.68	Y
22	3PL	550.92	7	190378	145.37	507.68	Y
23	3PL	385.46	7	190378	101.15	507.68	Y
24	3PL	630.84	7	190378	166.73	507.68	Y
25	3PL	146.32	7	190378	37.24	507.68	Y
26	3PL	545.82	7	190378	144.01	507.68	Y
27	2PPC	5291.90	44	190378	559.43	507.68	N
28	2PPC	6128.82	44	190378	648.64	507.68	N
29	2PPC	2042.63	26	190378	279.66	507.68	Y

Table 21. ELA Grade 7 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	Z observed	Z-critical	Fit OK?
1	3PL	144.59	7	197390	36.77	526.37	Y
2	3PL	315.93	7	197390	82.57	526.37	Y
3	3PL	51.47	7	197390	11.89	526.37	Y
4	3PL	165.80	7	197390	42.44	526.37	Y
5	3PL	82.99	7	197390	20.31	526.37	Y
6	3PL	73.35	7	197390	17.73	526.37	Y
7	3PL	160.80	7	197390	41.11	526.37	Y
8	3PL	50.83	7	197390	11.72	526.37	Y
9	3PL	69.50	7	197390	16.70	526.37	Y
10	3PL	155.70	7	197390	39.74	526.37	Y
11	3PL	152.30	7	197390	38.83	526.37	Y
12	3PL	313.83	7	197390	82.00	526.37	Y
13	3PL	129.11	7	197390	32.64	526.37	Y
14	3PL	140.60	7	197390	35.71	526.37	Y
15	3PL	245.31	7	197390	63.69	526.37	Y
16	3PL	1292.12	7	197390	343.46	526.37	Y
17	3PL	156.07	7	197390	39.84	526.37	Y
18	3PL	206.43	7	197390	53.30	526.37	Y
19	3PL	432.22	7	197390	113.65	526.37	Y
20	3PL	113.79	7	197390	28.54	526.37	Y
21	3PL	117.50	7	197390	29.53	526.37	Y
22	3PL	119.19	7	197390	29.98	526.37	Y
23	3PL	153.65	7	197390	39.19	526.37	Y
24	3PL	151.77	7	197390	38.69	526.37	Y
25	3PL	470.18	7	197390	123.79	526.37	Y
26	3PL	319.58	7	197390	83.54	526.37	Y
27	2PPC	839.00	17	197390	140.97	526.37	Y
28	2PPC	500.38	17	197390	82.90	526.37	Y
29	3PL	406.04	7	197390	106.65	526.37	Y
30	3PL	319.41	7	197390	83.50	526.37	Y
31	3PL	265.29	7	197390	69.03	526.37	Y
32	3PL	156.05	7	197390	39.83	526.37	Y
33	2PPC	212.17	17	197390	33.47	526.37	Y
34	2PPC	420.37	17	197390	69.18	526.37	Y
35	2PPC	2073.91	26	197390	283.99	526.37	Y

Table 22. ELA Grade 8 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	Z observed	Z-critical	Fit OK?
1	3PL	559.76	7	202831	147.73	540.88	Y
2	3PL	140.52	7	202831	35.69	540.88	Y
3	3PL	86.71	7	202831	21.30	540.88	Y
4	3PL	42.22	7	202831	9.41	540.88	Y
5	3PL	116.15	7	202831	29.17	540.88	Y
6	3PL	109.14	7	202831	27.30	540.88	Y
7	3PL	179.78	7	202831	46.18	540.88	Y
8	3PL	23.61	7	202831	4.44	540.88	Y
9	3PL	152.19	7	202831	38.80	540.88	Y
10	3PL	75.83	7	202831	18.39	540.88	Y
11	3PL	265.04	7	202831	68.96	540.88	Y
12	3PL	364.83	7	202831	95.63	540.88	Y
13	3PL	93.90	7	202831	23.22	540.88	Y
14	3PL	196.42	7	202831	50.62	540.88	Y
15	3PL	213.85	7	202831	55.28	540.88	Y
16	3PL	329.22	7	202831	86.12	540.88	Y
17	3PL	360.51	7	202831	94.48	540.88	Y
18	3PL	265.31	7	202831	69.04	540.88	Y
19	3PL	252.52	7	202831	65.62	540.88	Y
20	3PL	897.81	7	202831	238.08	540.88	Y
21	3PL	494.91	7	202831	130.40	540.88	Y
22	3PL	262.35	7	202831	68.24	540.88	Y
23	3PL	89.68	7	202831	22.10	540.88	Y
24	3PL	1064.70	7	202831	282.68	540.88	Y
25	3PL	392.71	7	202831	103.08	540.88	Y
26	3PL	149.82	7	202831	38.17	540.88	Y
27	2PPC	5139.55	44	202831	543.19	540.88	N
28	2PPC	4433.82	44	202831	467.96	540.88	Y
29	2PPC	1400.51	26	202831	190.61	540.88	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 2009 Grades 3–8 ELA assessments were calibrated and equated to the associated 2008 assessments, using two separate equating procedures.

In the first equating procedure, the new 2009 OP forms were pre-equated to the corresponding 2008 assessments. Prior to pre-equating, the FT items administered in 2008 were placed onto the OP scales in each grade. The equating of 2008 FT items to the 2008 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 2007, 2006, 2005, and (for Grades 4 and 8 only) 2003. All items field tested between 2003 and 2007 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 2003, 2005, 2006, 2007, and 2008 was used to select the 2009 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 2009 forms with the target 2008 OP forms. (Note that the 2008 OP TCC and SE curves were based on OP parameters and the 2009 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 2009 ELA OP data were re-calibrated after the 2009 OP administration. FT parameters for all MC items in OP tests were used as anchors to transform the 2009 OP item parameters to the OP scale. The CR items were not used as anchors in order to avoid potential error associated with rater effect. The MC items contained in the 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 (*M*1 and *M*2) 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 and that relationship defined by OP anchor item parameter estimates. This places the transformed parameters for the OP test items onto the New York State OP scale.

In this procedure, new 2009 OP parameter estimates were obtained for all items. The *a*-parameters and *b*-parameters were estimated freely while *c*-parameters of anchor items were fixed to their FT parameter 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:

$$M1 = A * M1_{Ft}$$
$$M2 = A * M2_{Ft} + B$$

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 $M1_{Ft}$ and $M2_{Ft}$ are the transformation constants previously used to place the anchor item FT parameter estimates onto the NYS scale.

The *A* and *B* values are derived from the input (FT) and estimate (OP) values of anchor items. Anchor input or FT values are known item parameter estimates entered into equating. Anchor estimate or OP 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 *A* and *B* constants are computed as follows:

$$A = \frac{SD_{Op}}{SD_{Ft}}$$

$$B = (Mean_{OP} - \frac{SD_{Op}}{SD_{Ft}} Mean_{Ft})$$

where

SD_{Op} is the standard deviation of anchor estimates in scale score metric.
SD_{Ft} is the standard deviation of anchor input values in scale score metric.
$Mean_{Op}$ is the mean of anchor estimates in scale score metric.
$Mean_{Ft}$ is the mean of anchor input in scale score metric.

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 23 presents the 2009 OP transformation constants for New York State Grades 3–8 ELA Tests.

Table 23. NYSTP ELA 2009 Final Transformation Constants

Grade	*M*1	*M*2
3	26.229	667.363
4	29.304	669.738
5	20.730	670.945
6	16.096	663.940
7	18.228	664.953
8	23.037	659.281

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. In the New York State Testing Program, different anchor sets are used each year to minimize item exposure that could adversely affect the accuracy of the equatings.

Anchor Item Evaluation

Anchor items were evaluated using several procedures. Procedures 1 and 2 evaluate the overall anchor set, while procedures 3, 4, and 5 evaluate 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 FT and the 2009 OP administration.

2. Correlations of anchor input and estimates of *a*- and *b*-parameters and p-values. Correlations of anchor input and estimate of *a*- and *b*-parameters and p-values 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 and p-values 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 FT estimates and the other on transformed estimates from the 2009 OP calibration. 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.

4. Delta plots (differences in the standardized proportion correct value). The delta-plot method relies on the differences in the standardized proportion correct value (p-value). P-values of the anchor items based on the FT (years 2003, 2005, 2006, 2007, and/or 2008) and the 2009 OP administration were calculated. The p-values were then converted to *z*-scores that correspond to the (1-p)th percentiles. A rule to identify outlier items that are functioning differentially between the two groups with respect to the level of difficulty is to draw the perpendicular distance to the line-of-best-fit. The fitted line is chosen to minimize the sum of squared perpendicular distances of the points to the line. Items lying more than two standard deviations from the fitted line are flagged as outliers.

5. Lord's chi-square criterion. Lord's χ^2 criterion involves significance testing of both item difficulty and discrimination parameters simultaneously for each item and evaluating the results based on the chi-square distribution table. (For details see Divgi, 1985; Lord, 1980.) If the null hypothesis that the item difficulty and discrimination parameters are equal is true, the item is not flagged for differential performance. If the null hypothesis is rejected and the observed value for χ^2 is greater than the critical χ^2 value, the items are flagged for performance differences between the two item administrations.

Table 24 provides a summary of anchor item evaluation and item flags.

Table 24. ELA Anchor Evaluation Summary

Grade	Number of Anchors	Anchor Input/ Estimate Correlation			Flagged Anchors (item numbers)			
		a-par	*b*-par	p-value	RMSD *a*-par	RMSD *b*-par	Delta	Lord's Chi-Square
3	24	0.895	0.969	0.988				
4	28	0.870	0.969	0.973	12,27	8,27	6,27	27
5	24	0.891	0.898	0.964		13	19	
6	26	0.864	0.823	0.858	3, 19	20	20	20
7	30	0.831	0.946	0.987	30	30	16	30
8	26	0.731	0.806	0.946	3,10	17		19, 22

In all cases the overall TCC alignment for anchor set input and estimate was good. The correlations for input and estimated p-values were over 0.90 for all grades except Grade 6, for which the correlation for input and estimate p-values was 0.86. Correlations for

b-parameter input and estimates ranged from 0.81 for Grade 8 to 0.97 for Grades 3 and 4. Correlations for *a*-parameter input and estimate ranged from 0.73 for Grade 8 to 0.90 for Grade 3. Correlations between *a*-parameter input and estimates for Grade 8 and correlations between *b*-parameter input and estimates for Grades 5, 6, and 8 were slightly below the NYS criterion.

Overall TCC alignment for anchor set input and estimate was very good (see Figures 1–6). In addition, correlations between parameter input and estimates were satisfactory for Grades 3–8. Therefore, despite the fact that some individual items were flagged by multiple methods in Grades 4, 6, and 7, no anchors were removed from any of the anchor sets. It was determined that removal of flagged anchors from Grades 4, 6, and 7 anchor sets had only minimal effect on item parameter estimates and minimal or no effect on the scoring tables.

An investigation of lower than expected correlations for Grade 8 revealed that under OP administration conditions students performed better on anchor items 1–10, field tested in 2008, while performing slightly worse on the remaining anchor items, which were field tested in 2003 or 2005–2007. It is possible that familiarity of items field tested in 2008 could have contributed to better performance on these items when they were administered operationally. Also, several factors (including population changes, curriculum changes, instruction method changes, etc.) might have contributed to item parameter changes between 2003/2005–2007 FT and 2009 OP administration. However, because the overall anchor set TCC alignment for Grade 8 was good (see Figure 6), all anchor items were retained for this grade. Similarly, all anchor items were retained for the remaining grades. Retaining all anchor items in all grades allowed for adequate anchor item content coverage and maintenance of anchor set reliability.

Figure 1. ELA Grade 3 Anchor Set TCC Alignment

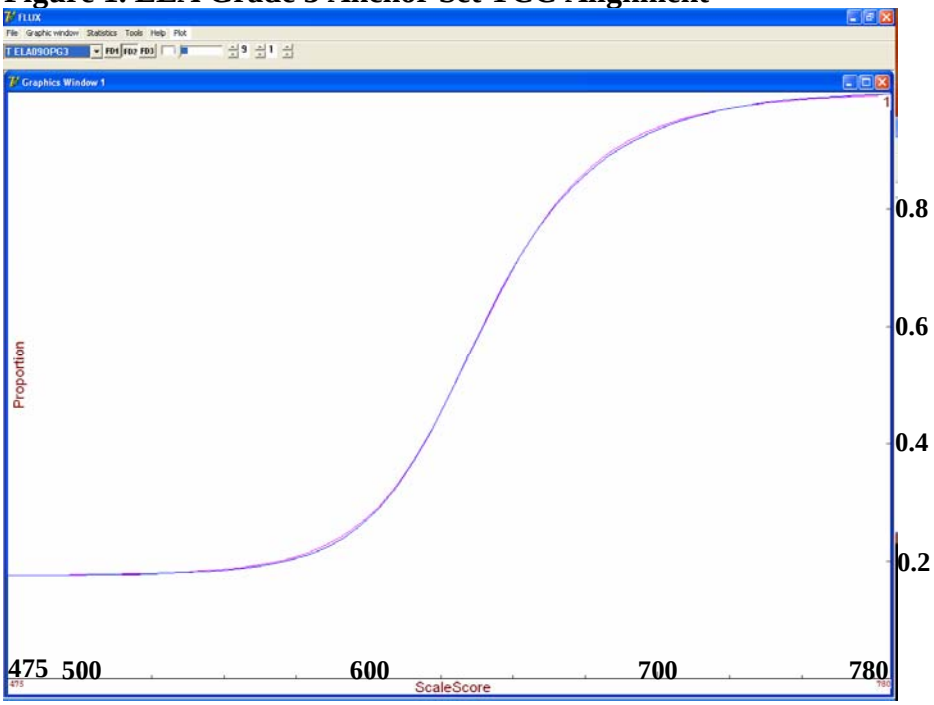


Figure 2. ELA Grade 4 Anchor Set TCC Alignment.

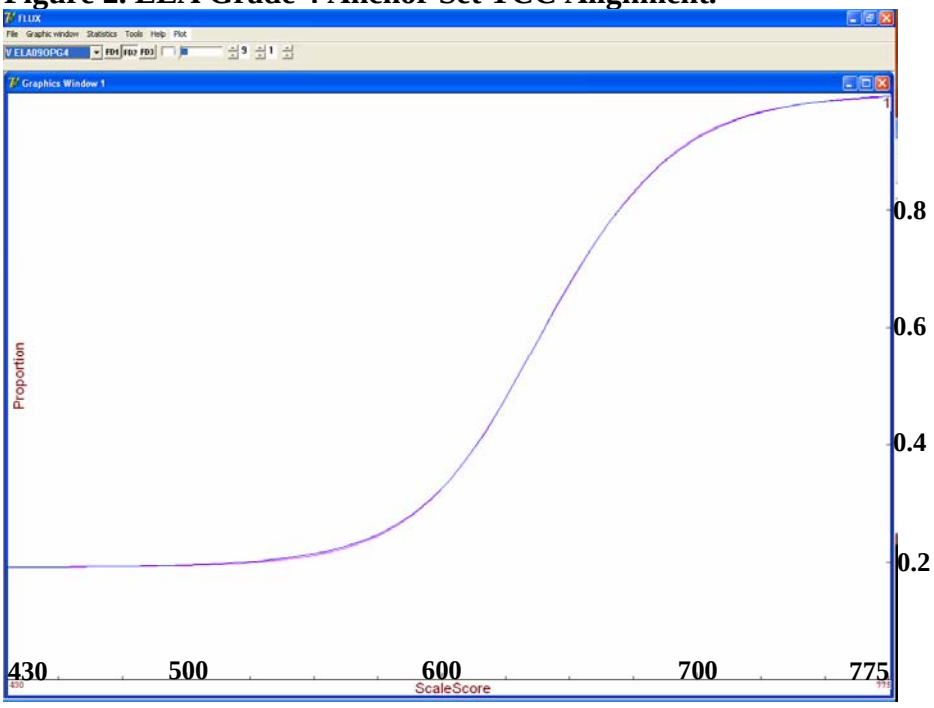


Figure 3. ELA Grade 5 Anchor Set TCC Alignment

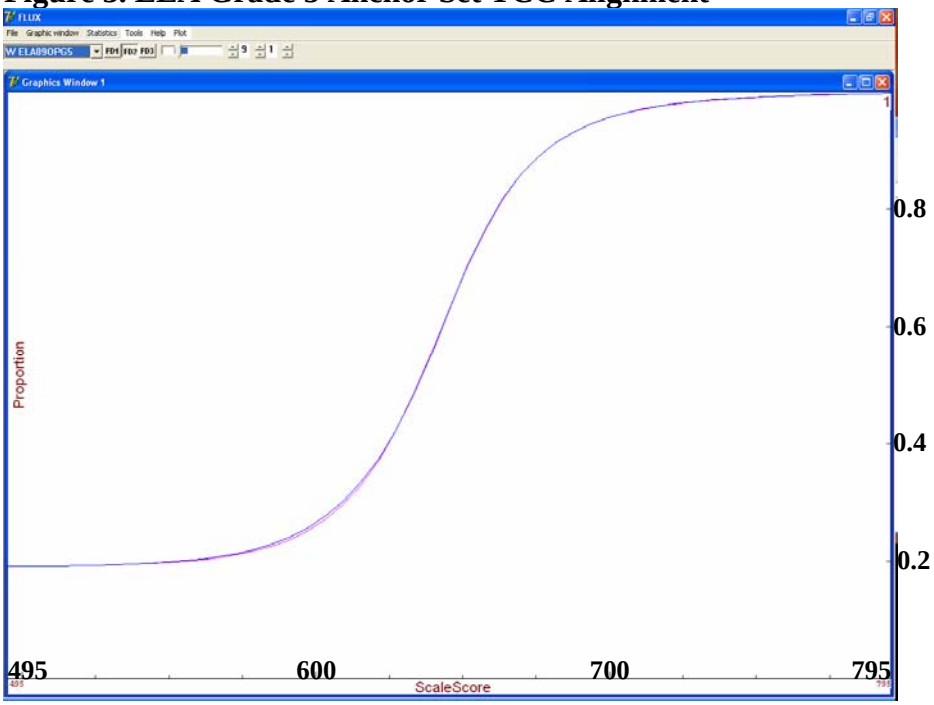


Figure 4. ELA Grade 6 Anchor Set TCC Alignment

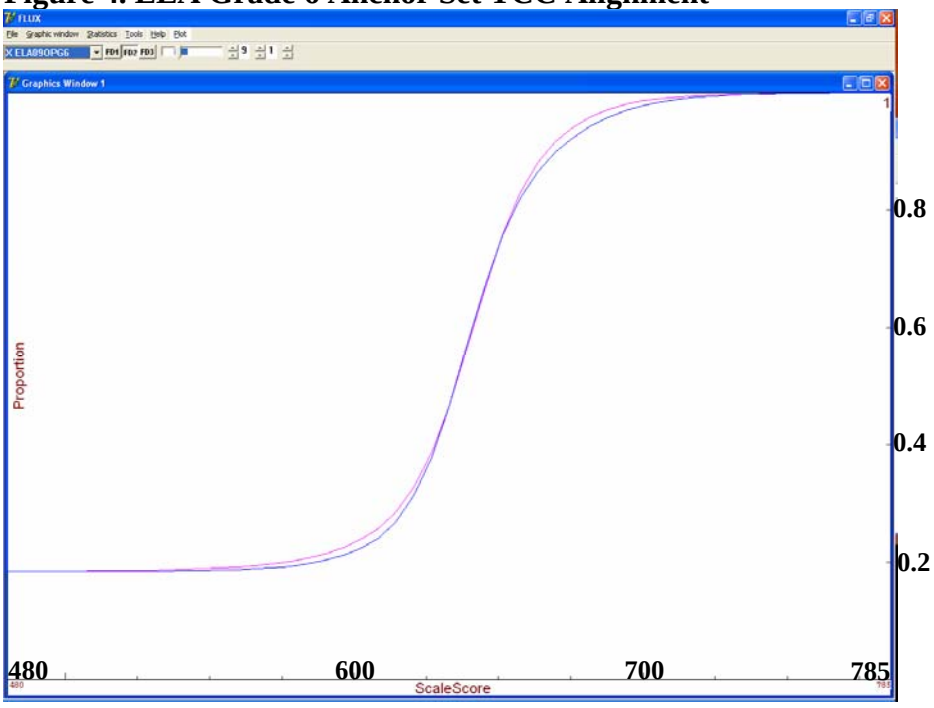


Figure 5. ELA Grade 7 Anchor Set TCC Alignment

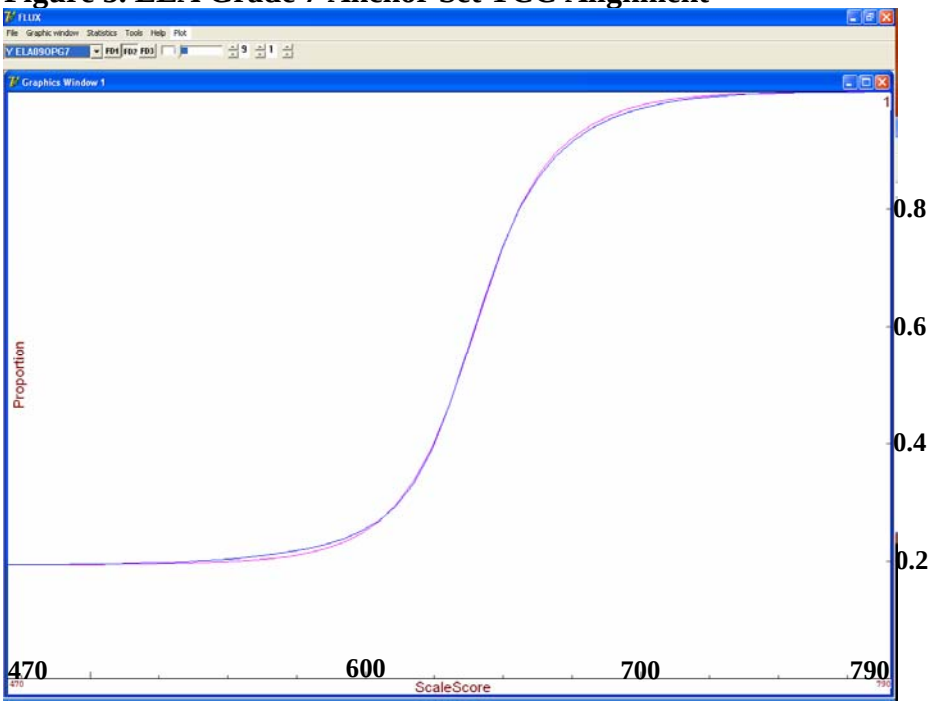
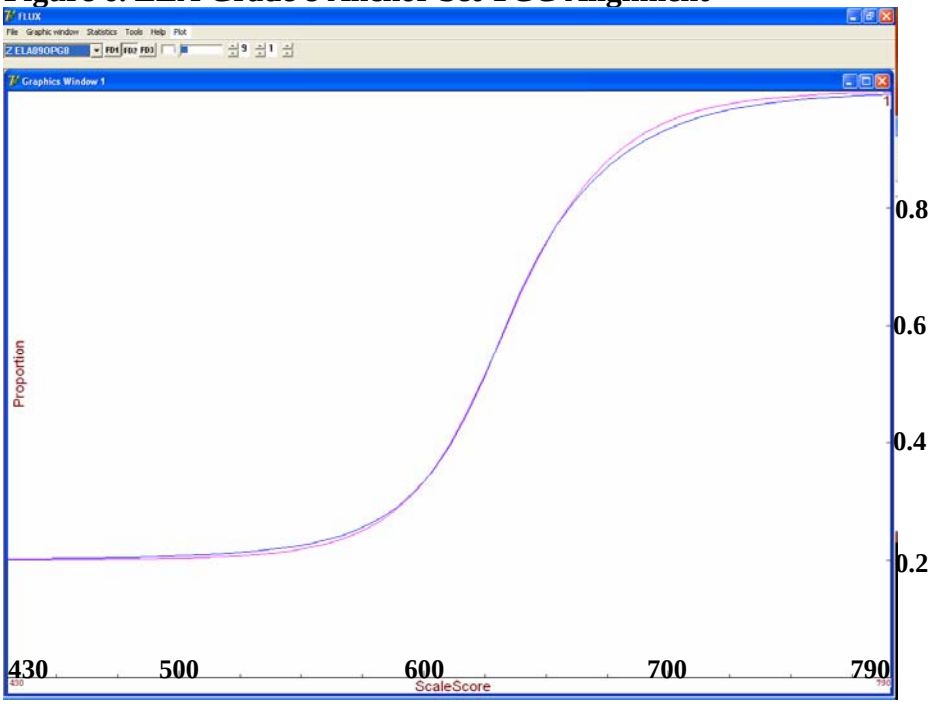


Figure 6. ELA Grade 8 Anchor Set TCC Alignment



Note that in Figures 1–6 anchor input parameters are represented by a blue TCC, and anchor estimate parameters are represented by a pink TCC. The *x*-axis is a theta scale expressed in scale score metric. The y-axis is the proportion of the anchor items that the students can

answer correctly. As seen in all the figures, the alignment of anchor input and estimate parameters is very good, indicating overall good stability of anchor parameters between FT and OP test administrations.

The anchor sets used to equate new OP assessments to the NYS scale are MC items only, and these items are representative of the test blueprint. The CR items were not included in anchor sets in order to avoid potential error associated with possible rater effects.

Item Parameters

The item parameters were estimated by the software PARDUX (Burket, 2002) and are presented in Tables 25–30. 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 25. 2009 Operational Item Parameter Estimates, Grade 3

Item	Max Pts	*a*-par/alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3
1	1	0.034	627.989	0.116	
2	1	0.032	649.109	0.085	
3	1	0.050	631.022	0.200	
4	1	0.017	705.799	0.178	
5	1	0.037	637.993	0.200	
6	1	0.034	612.862	0.143	
7	1	0.031	627.320	0.200	
8	1	0.043	626.552	0.200	
9	1	0.019	677.639	0.200	
10	1	0.038	668.792	0.173	
11	1	0.035	642.431	0.200	
12	1	0.040	635.139	0.166	
13	1	0.047	619.371	0.200	
14	1	0.040	647.650	0.177	
15	1	0.023	664.709	0.141	
16	1	0.038	649.582	0.200	
17	1	0.045	632.492	0.200	
18	1	0.046	640.840	0.116	
19	1	0.032	641.403	0.144	
20	1	0.030	656.577	0.166	
21	2	0.051	32.359	33.980	
22	1	0.024	612.725	0.200	
23	1	0.020	642.502	0.200	
24	1	0.047	625.786	0.200	
25	1	0.027	613.058	0.200	
26	2	0.044	26.043	25.296	
27	2	0.033	19.538	21.291	
28	3	0.026	16.271	15.822	15.404

Table 26. 2009 Operational Item Parameter Estimates, Grade 4

Item	Max Pts	*a*-par/ alpha	*b*-par/ gamma1	*c*-par/ gamma2	Gamma3	gamma4
1	1	0.021	638.904	0.200		
2	1	0.032	608.403	0.200		
3	1	0.028	618.990	0.200		
4	1	0.017	685.122	0.200		
5	1	0.026	643.019	0.200		
6	1	0.017	646.180	0.200		
7	1	0.025	619.434	0.200		
8	1	0.025	601.872	0.200		
9	1	0.030	625.418	0.200		
10	1	0.043	634.197	0.183		
11	1	0.038	643.989	0.181		
12	1	0.036	636.972	0.200		
13	1	0.036	638.558	0.200		
14	1	0.031	624.271	0.200		
15	1	0.033	614.064	0.200		
16	1	0.022	661.493	0.180		
17	1	0.035	639.312	0.151		
18	1	0.041	643.748	0.236		
19	1	0.021	639.417	0.200		
20	1	0.020	635.513	0.200		
21	1	0.032	620.259	0.200		
22	1	0.018	654.074	0.200		
23	1	0.014	658.382	0.200		
24	1	0.029	679.584	0.149		
25	1	0.025	653.768	0.133		
26	1	0.050	681.904	0.169		
27	1	0.036	667.792	0.231		
28	1	0.031	635.770	0.134		
29	4	0.056	32.254	34.815	36.896	39.455
30	4	0.064	37.327	40.187	42.423	44.836
31	3	0.048	27.946	30.791	33.162	

Table 27. 2009 Operational Item Parameter Estimates, Grade 5

Item	Max Pts	*a*-par/ alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3
1	1	0.035	628.482	0.200	
2	1	0.044	637.989	0.200	
3	1	0.061	643.985	0.200	
4	1	0.060	635.710	0.200	
5	1	0.052	657.610	0.187	
6	1	0.045	635.605	0.200	
7	1	0.039	654.429	0.200	
8	1	0.013	681.662	0.200	
9	1	0.021	662.568	0.185	
10	1	0.031	661.533	0.175	
11	1	0.056	632.149	0.200	
12	1	0.038	652.787	0.156	
13	1	0.038	635.845	0.200	
14	1	0.054	651.837	0.102	
15	1	0.064	648.283	0.240	
16	1	0.030	636.676	0.200	
17	1	0.037	642.330	0.170	
18	1	0.070	640.966	0.166	
19	1	0.046	648.323	0.200	
20	1	0.023	663.323	0.170	
21	2	0.079	49.848	52.220	
22	1	0.041	618.482	0.200	
23	1	0.024	623.100	0.200	
24	1	0.028	623.127	0.200	
25	1	0.026	625.819	0.200	
26	2	0.045	27.625	28.842	
27	3	0.042	27.122	27.481	28.562

Table 28. 2009 Operational Item Parameter Estimates, Grade 6

Item	Max Pts	*a*-par/ alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3	gamma4	gamma5
1	1	0.019	621.886	0.200			
2	1	0.041	606.575	0.200			
3	1	0.055	632.201	0.252			
4	1	0.068	639.168	0.207			
5	1	0.048	637.412	0.200			
6	1	0.064	626.575	0.200			
7	1	0.057	638.305	0.109			
8	1	0.058	644.884	0.108			
9	1	0.043	638.316	0.200			
10	1	0.037	624.962	0.200			
11	1	0.023	624.975	0.200			
12	1	0.081	635.108	0.200			
13	1	0.056	650.389	0.200			
14	1	0.074	635.568	0.200			
15	1	0.069	635.393	0.200			
16	1	0.087	643.653	0.113			
17	1	0.050	644.654	0.159			
18	1	0.033	656.224	0.128			
19	1	0.079	640.325	0.146			
20	1	0.048	647.579	0.127			
21	1	0.046	668.442	0.148			
22	1	0.083	644.414	0.273			
23	1	0.065	641.772	0.200			
24	1	0.044	666.219	0.200			
25	1	0.065	643.139	0.200			
26	1	0.087	650.336	0.200			
27	5	0.087	53.122	54.617	55.905	57.554	59.224
28	5	0.087	52.889	54.917	56.236	57.768	59.262
29	3	0.090	55.037	57.791	60.340		

Table 29. 2009 Operational Item Parameter Estimates, Grade 7

Item	Max Pts	*a*-par/alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3
1	1	0.044	623.440	0.200	
2	1	0.058	632.397	0.200	
3	1	0.049	640.354	0.200	
4	1	0.069	645.412	0.200	
5	1	0.061	628.503	0.200	
6	1	0.069	630.243	0.200	
7	1	0.044	611.709	0.200	
8	1	0.054	635.326	0.136	
9	1	0.056	637.670	0.122	
10	1	0.076	642.750	0.121	
11	1	0.036	671.284	0.114	
12	1	0.072	646.710	0.157	
13	1	0.070	642.525	0.203	
14	1	0.085	639.904	0.247	
15	1	0.041	680.260	0.214	
16	1	0.017	605.331	0.200	
17	1	0.047	628.656	0.200	
18	1	0.073	647.280	0.161	
19	1	0.049	642.129	0.200	
20	1	0.072	634.659	0.200	
21	1	0.033	624.198	0.200	
22	1	0.058	640.656	0.200	
23	1	0.052	643.632	0.149	
24	1	0.068	637.736	0.198	
25	1	0.028	679.290	0.232	
26	1	0.044	659.812	0.167	
27	2	0.063	40.064	42.043	
28	2	0.068	42.975	44.822	
29	1	0.049	660.147	0.199	
30	1	0.043	619.438	0.200	
31	1	0.056	660.921	0.362	
32	1	0.041	606.949	0.200	
33	2	0.051	30.500	33.378	
34	2	0.052	31.860	33.469	
35	3	0.064	41.750	42.202	43.730

Table 30. 2009 Operational Item Parameter Estimates, Grade 8

Item	Max Pts	*a*-par/ alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3	gamma4	gamma5
1	1	0.036	613.245	0.124			
2	1	0.033	623.347	0.200			
3	1	0.036	614.675	0.366			
4	1	0.029	621.777	0.136			
5	1	0.019	658.507	0.117			
6	1	0.025	602.988	0.200			
7	1	0.052	626.788	0.173			
8	1	0.025	637.725	0.200			
9	1	0.030	641.929	0.200			
10	1	0.053	617.305	0.320			
11	1	0.025	626.774	0.200			
12	1	0.029	612.766	0.200			
13	1	0.048	631.560	0.248			
14	1	0.022	655.817	0.200			
15	1	0.022	652.924	0.153			
16	1	0.026	621.397	0.200			
17	1	0.015	634.933	0.200			
18	1	0.021	641.211	0.200			
19	1	0.034	664.028	0.139			
20	1	0.035	661.496	0.440			
21	1	0.063	634.981	0.197			
22	1	0.054	633.150	0.133			
23	1	0.036	639.519	0.174			
24	1	0.039	637.992	0.127			
25	1	0.017	669.173	0.159			
26	1	0.031	648.561	0.200			
27	5	0.078	46.276	48.147	49.538	51.301	53.043
28	5	0.076	44.520	46.451	47.857	49.650	51.589
29	3	0.072	42.434	44.869	47.620		

Test Characteristic Curves

Test characteristic curves (TCCs) provide an overview of the tests in the IRT scale score metric. The 2008 and 2009 TCCs were generated using final OP item parameters for all test items administered in 2008 and 2009. 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 2008 and 2009 TCCs and SE curves are presented in Figures 7–12. 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 2008 OP curves are considered to be target curves for the 2009 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 2008 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 7. Grade 3 ELA 2008 and 2009 OP TCCs and SE curves

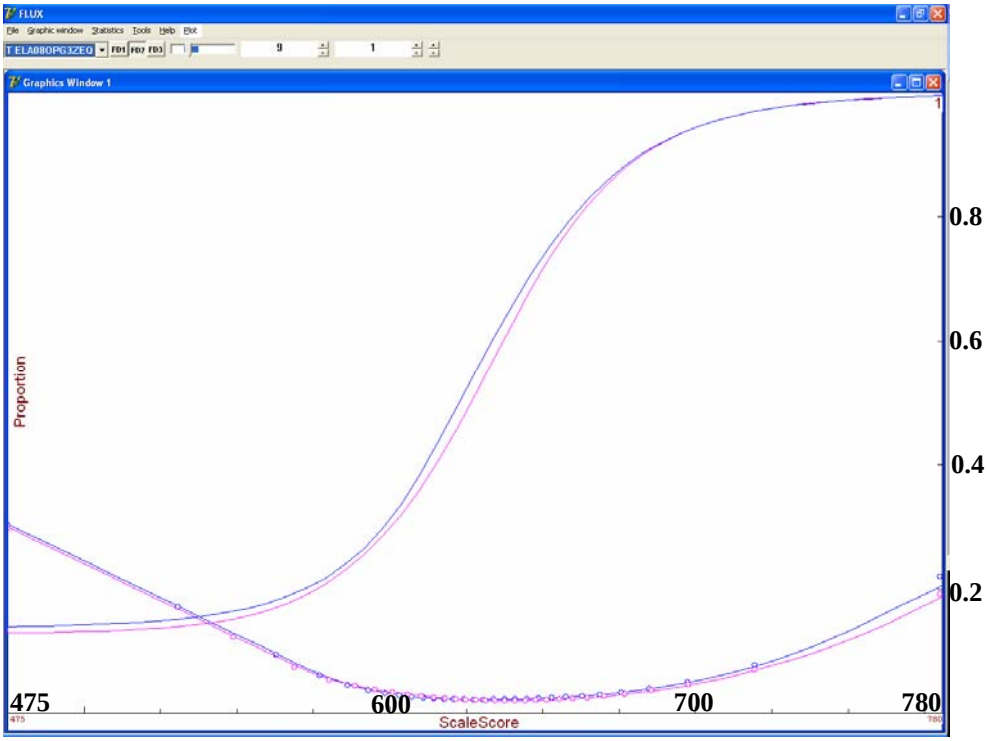


Figure 8. Grade 4 ELA 2008 and 2009 OP TCCs and SE curves

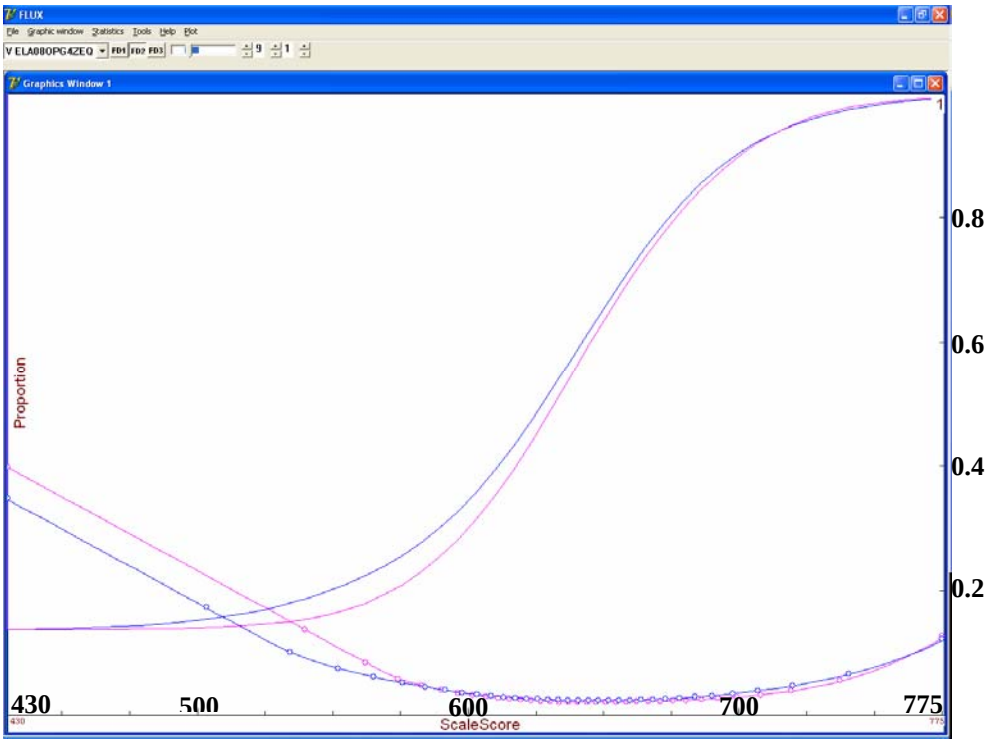


Figure 9. Grade 5 ELA 2008 and 2009 OP TCCs and SE curves

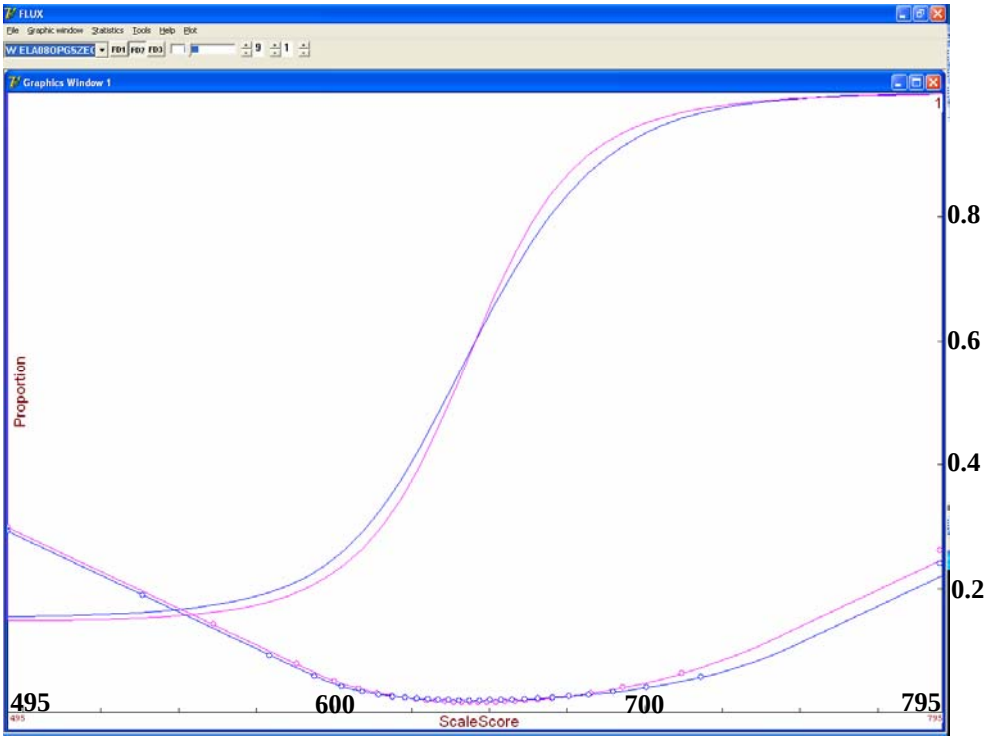


Figure 10. Grade 6 ELA 2008 and 2009 TCCs and SE curves

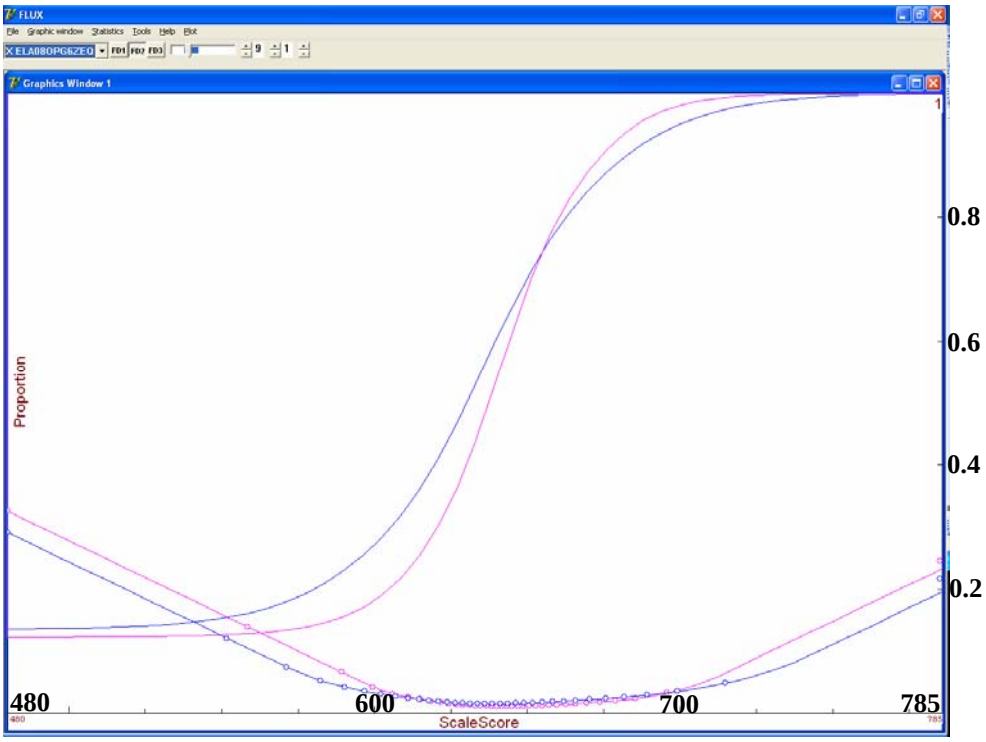


Figure 11. Grade 7 ELA 2008 and 2008 TCCs and SE curves

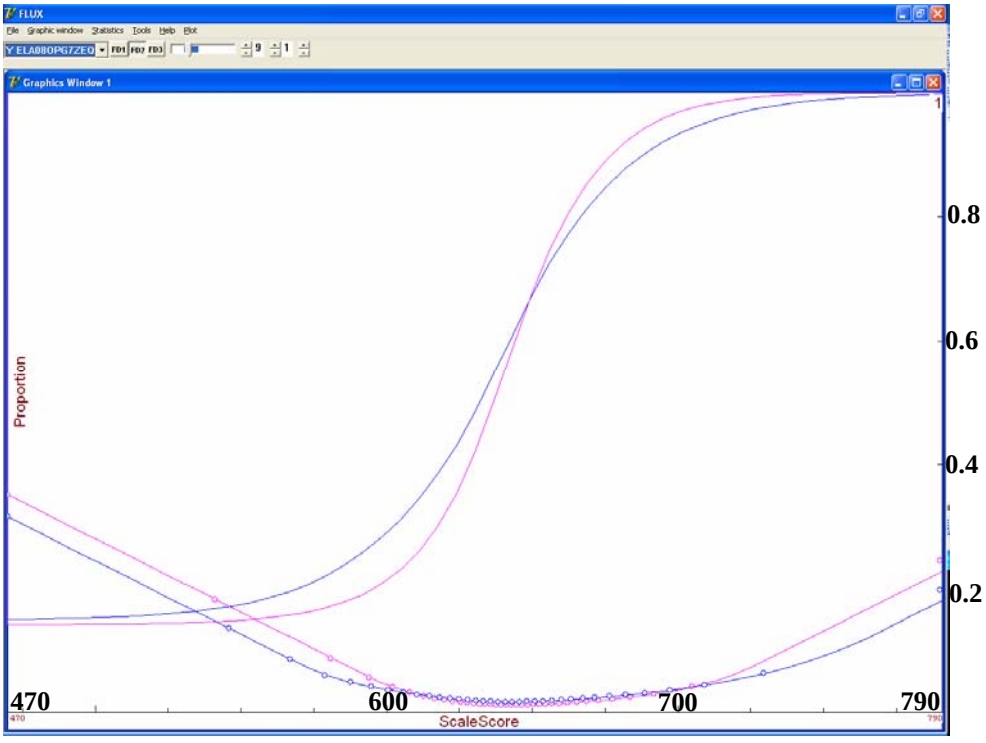
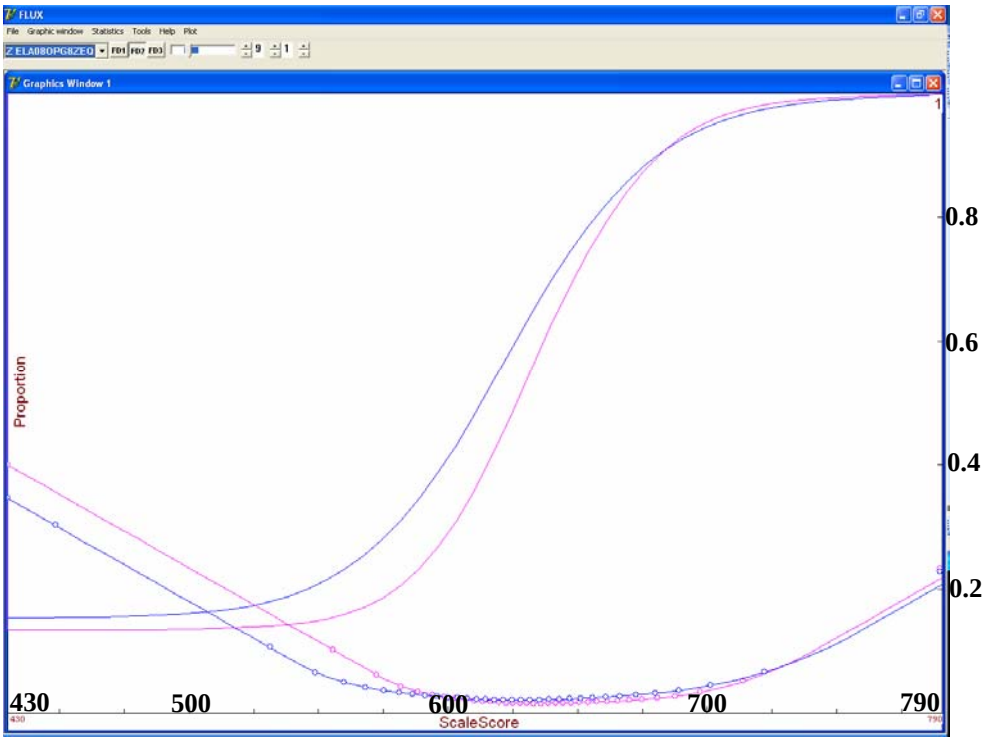


Figure 12. Grade 8 ELA 2008 and 2009 TCCs and SE curves



As seen in Figures 7–12, good alignments of 2008 and 2009 TCCs and SE curves were found for Grades 3, 4, and 5. The TCCs for Grade 6 and 7 were somewhat less well aligned at the lower and upper ends of the scale (indicating that the 2009 form tended to be slightly more difficult for lower-ability students and be slightly easier for the high-ability students), and the TCCs for Grade 8 were less well aligned at the lower and middle parts of the ability scales (indicating that the 2009 test form tended to be slightly more difficult for lower and middle-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:

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

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

Consistently 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 2008 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:

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

where

$SE(\hat{\theta})$ 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 31. Grade 3 Raw Score-to-Scale Score (with Standard Error)

Raw Score	Scale Score	Standard Error
0	475	125
1	475	125
2	475	125
3	475	125
4	475	125
5	549	51
6	569	31
7	580	22
8	589	18
9	595	16
10	601	14
11	606	13
12	610	11
13	614	11
14	618	10
15	622	10
16	625	9
17	628	9
18	631	9
19	635	9
20	638	9
21	641	9
22	644	9
23	648	9
24	651	9
25	655	10
26	660	10

(Continued on next page)

Table 31. Grade 3 Raw Score-to-Scale Score (with Standard Error) (cont.)

Raw Score	Scale Score	Standard Error
27	665	11
28	670	12
29	677	13
30	686	15
31	698	19
32	720	30
33	780	81

Table 32. Grade 4 Raw Score-to-Scale Score (with Standard Error)

Weighted Raw Score	Scale Score	Standard Error
0	430	163
1	430	163
2	430	163
3	430	163
4	430	163
5	430	163
6	536	56
7	558	35
8	570	25
9	578	20
10	585	17
11	591	15
12	597	14
13	602	13
14	606	12
15	610	12
16	614	11
17	618	10
18	621	10
19	625	10
20	628	9
21	631	9
22	634	9
23	637	9
24	641	9
25	644	9
26	647	9
27	650	9

(Continued on next page)

Table 32. Grade 4 Raw Score-to-Scale Score (with Standard Error) (cont.)

Weighted Raw Score	Scale Score	Standard Error
28	653	9
29	657	9
30	660	9
31	664	10
32	668	10
33	672	10
34	676	10
35	680	11
36	685	11
37	691	12
38	696	12
39	703	13
40	711	15
41	721	17
42	737	24
43	775	55

Table 33. Grade 5 Raw Score-to-Scale Score (with Standard Error)

Raw Score	Scale Score	Standard Error
0	495	126
1	495	126
2	495	126
3	495	126
4	495	126
5	561	60
6	588	33
7	600	21
8	608	16
9	614	13
10	619	11
11	623	10
12	627	9
13	630	8
14	633	8
15	636	7
16	639	7
17	641	7
18	644	7
19	647	7
20	649	7

(Continued on next page)

Table 33. Grade 5 Raw Score-to-Scale Score (with Standard Error) (cont.)

Raw Score	Scale Score	Standard Error
21	652	7
22	655	7
23	658	8
24	662	8
25	666	9
26	670	10
27	676	11
28	683	13
29	693	17
30	712	27
31	795	110

Table 34. Grade 6 Raw Score-to-Scale Score (with Standard Error)

Raw Score	Scale Score	Standard Error
0	480	137
1	480	137
2	480	137
3	480	137
4	480	137
5	558	59
6	589	28
7	599	18
8	606	13
9	611	11
10	615	9
11	619	8
12	622	7
13	624	7
14	627	6
15	629	6
16	631	5
17	633	5
18	635	5
19	637	5
20	638	5
21	640	5
22	642	5
23	643	5
24	645	5
25	647	5

(Continued on next page)

Table 34. Grade 6 Raw Score-to-Scale Score (with Standard Error) (cont.)

Raw Score	Scale Score	Standard Error
26	649	5
27	651	5
28	653	5
29	655	5
30	657	5
31	660	6
32	663	6
33	666	6
34	669	7
35	674	8
36	679	8
37	685	10
38	696	14
39	785	103

Table 35. 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	541	76
7	581	37
8	594	24
9	602	17
10	608	13
11	613	11
12	617	9
13	620	8
14	623	8
15	625	7
16	628	6
17	630	6
18	632	6
19	634	5
20	636	5
21	638	5
22	639	5

(Continued on next page)

Table 35. Grade 7 Raw Score-to-Scale Score (with Standard Error) (cont.)

Raw Score	Scale Score	Standard Error
23	641	5
24	643	5
25	645	5
26	647	5
27	648	5
28	650	5
29	652	5
30	655	6
31	657	6
32	659	6
33	662	7
34	665	7
35	669	8
36	673	8
37	678	9
38	684	10
39	692	13
40	705	18
41	790	103

Table 36. Grade 8 Raw Score to Scale Score (with Standard Error)

Weighted Raw Score	Scale Score	Standard Error
0	430	166
1	430	166
2	430	166
3	430	166
4	430	166
5	430	166
6	554	42
7	570	26
8	579	18
9	585	15
10	590	13
11	595	12
12	599	11
13	602	10
14	606	9
15	609	9
16	612	8
17	615	8

(Continued on next page)

Table 36. Grade 8 Raw Score-to-Scale Score (with Standard Error) (cont.)

Weighted Raw Score	Scale Score	Standard Error
18	617	8
19	620	8
20	622	7
21	625	7
22	627	7
23	630	7
24	632	7
25	634	7
26	637	7
27	639	7
28	642	7
29	644	7
30	647	7
31	650	8
32	653	8
33	656	8
34	659	8
35	662	8
36	666	9
37	670	9
38	674	10
39	679	10
40	684	11
41	691	13
42	700	16
43	717	25
44	790	98

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 F.

For the 2009 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 (scale score of 650 for all grades). Table 37 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 37. SPI Target Ranges

Grade	Objective	# Items	Total Points	Level III Cut SPI Target Range
3	8	9	9	61–79
	9	13	16	69–82
	10	5	5	45–59
4	8	11	11	65–78
	9	14	17	54–66
	10	5	8	54–68
5	8	13	14	62–76
	9	9	9	60–78
	10	4	5	52–71
6	8	10	10	60–75
	9	12	16	65–78
	10	6	10	58–71
7	8	15	17	70–81
	9	12	13	65–76
	10	7	8	53–67
8	8	10	14	66–80
	9	13	13	69–79
	10	5	9	56–70

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 of greatest value to improving 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 Grade 3, 4, and 8 tests; and two items were flagged on the Grade 5, 6, and 7 test, as is shown in Table 38. As indicated in the classical DIF analysis section, items flagged for DIF do not necessarily display bias.

Table 38. Number of Items Flagged for DIF by the Linn-Harnisch Method

Grade	Number of Flagged Items
3	0
4	0
5	2
6	2
7	2
8	0

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

Section VII: 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 39 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 39. 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	198123	28	33	0.86	2.11	0.87	2.04
4	195634	31	39	0.88	2.38	0.89	2.23
5	197522	27	31	0.83	2.06	0.85	1.97
6	197674	29	39	0.86	2.26	0.88	2.04
7	202400	35	41	0.88	2.31	0.89	2.19
8	207083	29	39	0.86	2.48	0.88	2.25

All the coefficients for total test reliability are in the range of 0.83–0.89, 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 40 presents reliabilities for the MC subsets.

Table 40. 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	198123	24	0.84	1.74	0.84	1.73
4	195634	28	0.85	1.95	0.85	1.90
5	197522	24	0.81	1.68	0.81	1.67
6	197674	26	0.84	1.64	0.84	1.63
7	202400	30	0.86	1.75	0.87	1.73
8	207083	26	0.82	1.90	0.82	1.89

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 41 should be interpreted with caution.

Table 41. 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	198123	4	9	0.55	1.14	0.60	1.08
4	195634	3	11	0.80	1.02	0.81	0.99
5	197522	3	7	0.54	1.11	0.60	1.03
6	197674	3	13	0.80	1.12	0.82	1.06
7	202400	5	11	0.70	1.36	0.72	1.31
8	207083	3	13	0.82	1.11	0.85	1.03

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 42a–42f, 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 exception of Grade 5 NRC = 6 (Low Needs districts). Feldt-Raju reliability coefficients, which tend to be larger than the Cronbach alpha estimates for the same group, were all larger than or equal to 0.80 with the exceptions of Grade 3 NRC = 7 (Charter schools), Grade 5 NRC = 6 (Low Needs districts) and NRC = 7 (Charter schools) and Grade 6 unknown ethnicity group. All other test reliability alpha statistics were in the 0.81–0.89 range, indicating very good test internal consistency (reliability) for analyzed subgroups of examinees.

Table 42a. 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	198123	0.86	2.11	0.87	2.04
Gender	Female	96707	0.85	2.02	0.86	1.97
	Male	101416	0.87	2.18	0.88	2.11
Ethnicity	Asian	15808	0.83	1.88	0.84	1.84
	Black	37517	0.86	2.30	0.87	2.23
	Hispanic	42112	0.86	2.29	0.87	2.23
	American Indian	934	0.84	2.28	0.85	2.22
	Multi-Racial	642	0.83	2.02	0.84	1.97
	White	101012	0.84	1.92	0.85	1.85
	Unknown	98	0.84	2.00	0.84	1.95
NRC	New York City	70202	0.87	2.21	0.88	2.14
	Big 4 Cites	8193	0.87	2.41	0.87	2.33
	High Needs Urban/Suburban	16107	0.86	2.22	0.87	2.15
	High Needs Rural	11493	0.85	2.16	0.86	2.10
	Average Needs	58500	0.84	2.01	0.85	1.96
	Low Needs	29916	0.81	1.80	0.81	1.76
	Charter	3493	0.78	2.06	0.79	2.03
SWD	All Codes	26929	0.87	2.57	0.88	2.48
SUA	All Codes	45137	0.87	2.48	0.88	2.40
SWD/SUA	SUA=504 plan codes	22779	0.87	2.60	0.88	2.51
ELL/SUA	SUA=ELL codes	15579	0.85	2.47	0.86	2.39
ELL	ELL = Y	17491	0.86	2.48	0.87	2.40

Table 42b. 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	195634	0.88	2.38	0.89	2.23
Gender	Female	95246	0.87	2.33	0.89	2.19
	Male	100388	0.88	2.41	0.90	2.27
Ethnicity	Asian	14588	0.86	2.24	0.88	2.08
	Black	37452	0.87	2.51	0.88	2.39
	Hispanic	41415	0.87	2.51	0.89	2.39
	American Indian	916	0.87	2.46	0.88	2.35
	Multi-Racial	486	0.88	2.35	0.89	2.22
	White	100696	0.87	2.25	0.88	2.12
	Unknown	81	0.84	2.27	0.86	2.14
NRC	New York City	68638	0.88	2.48	0.89	2.33
	Big 4 Cites	7943	0.88	2.57	0.89	2.45
	High Needs Urban/Suburban	15738	0.87	2.44	0.88	2.32
	High Needs Rural	11475	0.87	2.40	0.89	2.28
	Average Needs	58934	0.86	2.28	0.88	2.16
	Low Needs	29766	0.84	2.10	0.86	1.99
	Charter	2905	0.83	2.35	0.84	2.27
SWD	All Codes	28896	0.88	2.66	0.89	2.55
SUA	All Codes	45648	0.88	2.63	0.89	2.52
SWD/SUA	SUA=504 plan codes	25827	0.88	2.67	0.88	2.57
ELL/SUA	SUA=ELL codes	12627	0.85	2.63	0.85	2.55
ELL	ELL = Y	14382	0.85	2.64	0.86	2.55

Table 42c. 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	197522	0.83	2.06	0.85	1.97
Gender	Female	96843	0.82	2.02	0.84	1.94
	Male	100679	0.84	2.09	0.86	1.99
Ethnicity	Asian	14523	0.83	1.90	0.84	1.81
	Black	37980	0.82	2.26	0.84	2.18
	Hispanic	41408	0.83	2.24	0.84	2.16
	American Indian	953	0.84	2.16	0.85	2.07
	Multi-Racial	486	0.81	2.02	0.82	1.93
	White	102073	0.80	1.89	0.82	1.81
	Unknown	99	0.83	1.97	0.84	1.88

(Continued on next page)

Table 42c. 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	68531	0.84	2.20	0.85	2.10
	Big 4 Cites	7506	0.84	2.32	0.86	2.24
	High Needs Urban/Suburban	15484	0.83	2.14	0.84	2.06
	High Needs Rural	11477	0.82	2.02	0.83	1.95
	Average Needs	60033	0.81	1.93	0.82	1.85
	Low Needs	30582	0.77	1.77	0.78	1.71
	Charter	3613	0.79	2.21	0.80	2.15
SWD	All Codes	30705	0.84	2.42	0.85	2.36
SUA	All Codes	46138	0.84	2.39	0.85	2.32
SWD/SUA	SUA=504 plan codes	27907	0.84	2.43	0.85	2.37
ELL/SUA	SUA=ELL codes	10483	0.82	2.45	0.82	2.40
ELL	ELL = Y	12309	0.82	2.46	0.83	2.41

Table 42d. 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	197674	0.86	2.26	0.88	2.04
Gender	Female	96510	0.85	2.18	0.88	1.98
	Male	101164	0.86	2.30	0.89	2.08
Ethnicity	Asian	14726	0.86	2.07	0.89	1.82
	Black	38107	0.85	2.41	0.87	2.23
	Hispanic	40522	0.86	2.42	0.88	2.22
	American Indian	902	0.85	2.35	0.88	2.16
	Multi-Racial	431	0.84	2.17	0.87	1.95
	White	102913	0.83	2.10	0.86	1.90
	Unknown	73	0.75	2.07	0.80	1.88
NRC	New York City	68151	0.87	2.40	0.89	2.17
	Big 4 Cites	7490	0.86	2.46	0.88	2.28
	High Needs Urban/Suburban	15067	0.86	2.32	0.88	2.13
	High Needs Rural	11422	0.84	2.23	0.87	2.04
	Average Needs	61034	0.83	2.12	0.86	1.94
	Low Needs	30969	0.81	1.93	0.84	1.76
	Charter	3215	0.82	2.30	0.84	2.16

(Continued on next page)

Table 42d. 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	30362	0.86	2.58	0.88	2.44
SUA	All Codes	41788	0.86	2.57	0.88	2.41
SWD/SUA	SUA=504 plan codes	27009	0.86	2.58	0.87	2.45
ELL/SUA	SUA=ELL codes	8248	0.85	2.66	0.87	2.51
ELL	ELL = Y	10414	0.85	2.67	0.87	2.52

Table 42e. 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	202400	0.88	2.31	0.89	2.19
Gender	Female	98510	0.87	2.24	0.88	2.12
	Male	103890	0.89	2.36	0.90	2.25
Ethnicity	Asian	14822	0.89	2.17	0.90	2.04
	Black	38408	0.87	2.49	0.88	2.40
	Hispanic	41120	0.88	2.49	0.89	2.40
	American Indian	935	0.88	2.42	0.89	2.34
	Multi-Racial	400	0.85	2.26	0.87	2.15
	White	106633	0.86	2.12	0.87	2.02
	Unknown	82	0.86	2.19	0.87	2.07
NRC	New York City	69611	0.89	2.45	0.90	2.33
	Big 4 Cites	7685	0.89	2.57	0.89	2.48
	High Needs Urban/Suburban	15326	0.87	2.42	0.88	2.32
	High Needs Rural	12101	0.87	2.28	0.88	2.19
	Average Needs	63001	0.85	2.16	0.86	2.06
	Low Needs	31745	0.82	1.97	0.84	1.88
	Charter	2476	0.83	2.37	0.84	2.28
SWD	All Codes	31155	0.88	2.67	0.89	2.60
SUA	All Codes	40645	0.88	2.66	0.89	2.59
SWD/SUA	SUA=504 plan codes	27456	0.88	2.67	0.88	2.61
ELL/SUA	SUA=ELL codes	7629	0.86	2.76	0.87	2.70
ELL	ELL = Y	9603	0.86	2.76	0.87	2.70

Table 42f. 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	207083	0.86	2.48	0.88	2.25
Gender	Female	101526	0.85	2.40	0.87	2.20
	Male	105557	0.86	2.53	0.89	2.29
Ethnicity	Asian	14982	0.86	2.37	0.89	2.10
	Black	39551	0.85	2.61	0.87	2.42
	Hispanic	41871	0.85	2.65	0.88	2.42
	American Indian	1012	0.85	2.57	0.87	2.38
	Multi-Racial	301	0.84	2.37	0.87	2.18
	White	109299	0.84	2.30	0.86	2.12
	Unknown	67	0.83	2.49	0.86	2.22
NRC	New York City	71642	0.86	2.62	0.88	2.37
	Big 4 Cites	7686	0.87	2.71	0.89	2.50
	High Needs Urban/Suburban	15544	0.86	2.55	0.88	2.35
NRC	High Needs Rural	12362	0.85	2.44	0.87	2.26
	Average Needs	65108	0.83	2.32	0.86	2.15
	Low Needs	31981	0.80	2.12	0.83	1.98
	Charter	2164	0.81	2.43	0.83	2.30
SWD	All Codes	30617	0.85	2.74	0.86	2.58
SUA	All Codes	41006	0.85	2.76	0.87	2.58
SWD/SUA	SUA=504 plan codes	27446	0.84	2.74	0.86	2.59
ELL/SUA	SUA=ELL codes	8039	0.82	2.84	0.85	2.66
ELL	ELL = Y	9730	0.83	2.83	0.85	2.66

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 39. SEMs ranged 1.97–2.48, which is reasonable and small. In other words, the error of measurement from the observed test score ranged from approximately ± 2 to ± 4 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 42a–42f. The SEMs associated with all reliability estimates for all subpopulations are in the range of 1.71–2.83, 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 I.

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 43 and 44 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 43 depicts the consistency study results based on the range of performance levels for all grades. Overall, between 74% and 82% 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 0.55–0.63.

Table 43. Decision Consistency (All Cuts)

Grade	N-count	Agreement	Inconsistency	Kappa
3	198123	0.7384	0.2616	0.5455
4	195634	0.8031	0.1969	0.6116
5	197522	0.7552	0.2448	0.5453
6	197674	0.8217	0.1783	0.6323
7	202400	0.8213	0.1787	0.6187
8	207083	0.7971	0.2029	0.6234

Table 44 depicts the consistency study results based on two performance levels (passing and not passing) as defined by the Level III cut. Overall, about 88%–92% 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 0.69–0.74.

Table 44. Decision Consistency (Level III Cut)

Grade	N-count	Agreement	Inconsistency	Kappa
3	198123	0.8917	0.1083	0.7085
4	195634	0.9036	0.0964	0.7315
5	197522	0.9084	0.0916	0.6929
6	197674	0.9125	0.0875	0.7251
7	202400	0.9169	0.0831	0.7417
8	207083	0.8763	0.1237	0.7164

Accuracy

The results of classification accuracy are presented in Table 45. Included in the table are case counts (N-count), 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 81%–86% of the time across all performance levels and approximately 91%–94% of the time in regards to the Level III cut score.

Table 45. 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	198123	**0.8067**	0.1346	0.0588	**0.9237**	0.0349	0.0415
4	195634	**0.8545**	0.1021	0.0434	**0.9296**	0.0417	0.0287
5	197522	**0.8088**	0.1449	0.0464	**0.9354**	0.0293	0.0353
6	197674	**0.8614**	0.1020	0.0366	**0.9379**	0.0310	0.0311
7	202400	**0.8600**	0.1112	0.0288	**0.9374**	0.0401	0.0225
8	207083	**0.8519**	0.0935	0.0546	**0.9124**	0.0407	0.0468

Section VIII: Summary of Operational Test Results

This section summarizes the distribution of OP scale score results on the New York State 2009 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, 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 as disability 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 I.

Scale Score Distribution Summary

Scale score distribution summary tables are presented and discussed in Tables 46–52. In Table 46, scale score statistics for total populations of students from public and charter schools are presented. In Tables 47–52, 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 46. ELA Grades 3–8 Scale Score Distribution Summary

Grade	N-count	SS Mean	SS Std Dev	10^{th} %tile	25^{th} %tile	50^{th} %tile	75^{th} %tile	90^{th} %tile
3	198123	669.97	35.81	631	651	670	686	720
4	195634	669.93	34.72	631	650	672	691	711
5	197522	675.47	34.58	644	655	670	683	712
6	197674	667.31	27.64	643	653	663	674	685
7	202400	667.19	27.06	641	652	665	678	692
8	207083	661.09	30.82	627	644	659	679	691

Grade 3

Scale score statistics and N-counts of demographic groups for Grade 3 are presented in Table 47. The population scale score mean was 669.97 with a standard deviation of 35.81. By gender subgroup, Females outperformed Males, and the difference was more than eight 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 (681.70). 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. SWD, SUA, and ELL subgroups scored, on average, approximately three-fourths of one standard deviation below the mean scale score for the population. The SWD subgroup, which had a scale score mean about 32 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 670: Asian (677), White (676) and Low Needs districts (686).

Table 47. 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	198123	669.97	35.81	631	651	670	686	720
Gender	Female	96707	674.23	35.16	638	655	670	686	720
	Male	101416	665.91	35.96	625	648	665	686	698
Ethnicity	Asian	15808	681.70	35.74	644	660	677	698	720
	Black	37517	657.99	32.85	622	641	660	677	698
	Hispanic	42112	657.89	33.05	622	641	660	677	698
	American Indian	934	659.18	31.07	625	641	660	677	698
	Multi-Racial	642	674.19	34.14	641	655	670	686	720
	White	101012	677.69	35.29	641	660	677	698	720
	Unknown	98	676.14	37.17	641	660	677	686	720
NRC	New York City	70202	663.69	35.72	625	644	665	686	698
	Big 4 Cities	8193	651.13	33.71	614	631	651	670	686
	High Needs Urban/Suburban	16107	663.33	33.97	625	644	665	677	698
	High Needs Rural	11493	665.56	32.15	631	648	665	686	698
	Average Needs	58500	674.73	34.02	638	655	670	686	720
	Low Needs	29916	686.26	35.29	651	665	686	698	720
	Charter	3493	669.85	28.15	641	651	670	686	698
ELL	ELL = Y	17491	645.32	31.63	610	628	648	665	677
SWD	All Codes	26929	637.57	36.17	601	618	641	660	677
SUA	All Codes	45137	645.06	34.69	606	625	648	665	686
SWD/SUA	SUA=504 Plan codes	22779	634.26	35.37	595	614	638	655	670
ELL/SUA	SUA=ELL codes	15579	646.63	30.87	610	631	648	665	677

Grade 4

Scale score statistics and N-counts of demographic groups for Grade 4 are presented in Table 48. The Grade 4 population (All Students) mean was 669.93, with a standard deviation of 34.72. By gender subgroup, Females outperformed Males, but the difference was less than nine 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

(682.32). 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 subgroup had a scale score mean nearly 35 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 672: Female (676), Asian (680), White (676), Average Needs districts (676), and Low Needs districts (685).

Table 48. 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	195634	669.93	34.72	631	650	672	691	711
Gender	Female	95246	674.16	34.00	634	653	676	696	711
	Male	100388	665.92	34.92	625	647	668	685	703
Ethnicity	Asian	14588	682.32	34.47	644	660	680	703	721
	Black	37452	658.26	32.98	621	641	660	680	696
	Hispanic	41415	657.53	33.36	621	641	660	676	696
	American Indian	916	660.16	31.49	625	644	660	680	696
	Multi-Racial	486	672.31	34.31	631	653	672	691	711
	White	100696	677.65	33.10	641	660	676	696	711
	Unknown	81	678.72	31.55	641	660	676	691	721
NRC	New York City	68638	662.69	35.43	625	644	664	685	703
	Big 4 Cities	7943	652.38	35.30	614	634	653	672	691
	High Needs Urban/Suburban	15738	663.74	32.97	628	647	664	685	696
	High Needs Rural	11475	665.52	32.56	628	647	668	685	703
	Average Needs	58934	675.14	32.01	637	657	676	691	711
	Low Needs	29766	686.80	31.47	653	668	685	703	721
	Charter	2905	666.79	26.23	634	650	664	685	696
ELL	ELL = Y	14382	638.88	33.06	602	625	644	660	672
SWD	All Codes	28896	635.15	38.01	591	618	641	660	676
SUA	All Codes	45648	641.92	36.13	602	625	647	664	680
SWD/SUA	SUA=504 Plan codes	25827	632.84	37.83	591	614	637	657	672
ELL/SUA	SUA= ELL codes	12627	640.21	31.39	606	628	644	660	672

Grade 5

Scale score summary statistics for Grade 5 students are in Table 49. Overall, the scale score mean was 675.47, with a standard deviation of 34.58. The difference between mean scale scores by gender groups was very small (about four 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 a half of standard deviation below the population mean. SWD, SUA, and ELL subgroups scored approximately three-fourths standard deviation below the mean scale score for the population. The SWD subgroup, which had a scale score mean nearly 35 scale score units (one standard deviation) below the population mean, was the lowest performing group analyzed. At the 50th percentile, the following groups exceeded the population score of 670: Asian (676), White (676), Average Needs districts (676), and Low Needs districts (683).

Table 49. 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	197522	675.47	34.58	644	655	670	683	712
Gender	Female	96843	677.20	34.60	644	658	670	693	712
	Male	100679	673.80	34.47	641	655	670	683	712
Ethnicity	Asian	14523	685.73	39.31	649	662	676	693	712
	Black	37980	663.02	28.12	636	649	662	676	693
	Hispanic	41408	664.31	29.36	636	649	662	676	693
	American Indian	953	669.23	30.72	639	652	666	683	693
	Multi-Racial	486	677.96	34.02	647	658	670	693	712
	White	102073	683.20	35.36	652	662	676	693	712
	Unknown	99	680.93	37.50	649	662	676	693	712
NRC	New York City	68531	668.58	33.13	639	649	666	683	693
	Big 4 Cities	7506	658.99	29.22	630	644	658	670	693
	High Needs Urban/Suburban	15484	669.89	31.13	639	652	666	683	693
	High Needs Rural	11477	674.27	30.76	644	658	670	683	712
	Average Needs	60033	680.80	34.13	649	662	676	693	712
	Low Needs	30582	689.27	36.93	655	670	683	693	712
	Charter	3613	665.96	26.66	641	652	662	676	693
ELL	ELL = Y	12309	646.32	24.42	623	636	647	658	670
SWD	All Codes	30705	650.02	26.89	623	636	649	662	676
SUA	All Codes	46138	653.16	26.84	627	639	652	666	683
SWD/SUA	SUA=504 Plan codes	27907	648.93	26.24	623	636	649	662	676
ELL/SUA	SUA= ELL codes	10483	647.30	24.21	623	636	649	658	670

Grade 6

Scale score summary statistics for Grade 6 students are in Table 50. The scale score mean was 667.31, with a standard deviation of 27.64. 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 third of standard deviation

below the population mean. SWD, SUA, and ELL subgroups scored over two-thirds 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 663: Asian (674), Multi-Racial (666), White (669), Average Needs districts (666), and Low Needs districts (674).

Table 50. 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	197674	667.31	27.64	643	653	663	674	685
Gender	Female	96510	670.97	29.69	645	655	666	679	696
	Male	101164	663.82	25.03	642	651	663	674	685
Ethnicity	Asian	14726	678.96	35.49	649	660	674	685	696
	Black	38107	658.43	20.71	638	647	657	666	679
	Hispanic	40522	658.74	22.02	637	647	657	669	679
	American Indian	902	659.79	21.97	640	649	657	669	679
	Multi-Racial	431	670.74	33.29	647	655	666	679	696
	White	102913	672.35	28.70	649	657	669	679	696
	Unknown	73	677.32	35.01	651	660	666	679	696
NRC	New York City	68151	662.19	26.29	638	649	660	669	685
	Big 4 Cities	7490	656.60	20.20	637	645	655	666	679
	High Needs Urban/Suburban	15067	661.87	22.70	640	651	660	669	679
	High Needs Rural	11422	664.99	23.65	643	653	663	674	685
	Average Needs	61034	670.84	27.46	647	657	666	679	696
	Low Needs	30969	678.72	31.60	653	663	674	685	696
	Charter	3215	660.97	17.96	642	651	660	669	679
ELL	ELL = Y	10414	643.74	17.69	624	635	645	653	663
SWD	All Codes	30362	647.02	18.22	629	638	649	657	666
SUA	All Codes	41788	648.91	18.58	629	640	649	660	669
SWD/SUA	SUA=504 Plan codes	27009	646.29	17.76	627	638	647	657	666
ELL/SUA	SUA= ELL codes	8248	644.44	17.62	627	637	645	655	663

Grade 7

Scale score statistics and N-counts of demographic groups for Grade 7 are presented in Table 51. The population scale score mean was 667.19 and the population standard deviation was 27.06. By gender subgroup, Females outperformed Males, but the difference was about one-fifth 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 (675.00). The students from the Big 4 Cities scored below their peers from schools with other NRC designations and about a half of standard deviation below the population mean. SWD and SUA subgroups scored approximately four-fifths standard deviation below the mean scale score for the population. The ELL subgroup, which had a scale score mean slightly more than 31 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: Female (669), Asian (673), White (673), Average Needs districts (669), and Low Needs districts (678).

Table 51. 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	202400	667.19	27.06	641	652	665	678	692
Gender	Female	98510	669.90	27.32	643	655	669	684	692
	Male	103890	664.63	26.56	638	650	662	678	692
Ethnicity	Asian	14822	675.00	31.83	647	659	673	684	705
	Black	38408	656.83	20.66	636	647	657	669	678
	Hispanic	41120	656.46	22.61	634	645	657	669	678
	American Indian	935	659.19	24.58	636	648	659	669	684
	Multi-Racial	400	670.09	27.05	647	657	665	678	705
	White	106633	674.04	27.40	648	659	673	684	692
	Unknown	82	669.89	31.38	650	657	669	684	692
NRC	New York City	69611	660.34	25.49	636	647	659	673	684
	Big 4 Cities	7685	653.16	23.48	630	641	652	665	678
	High Needs Urban/Suburban	15326	660.72	22.97	638	648	659	673	684
	High Needs Rural	12101	665.84	23.38	641	652	665	678	692
	Average Needs	63001	672.12	25.85	648	659	669	684	692
	Low Needs	31745	680.38	29.34	655	665	678	692	705
	Charter	2476	662.21	19.86	643	650	662	673	684
ELL	ELL = Y	9603	636.35	23.58	613	628	639	650	657
SWD	All Codes	31155	645.02	21.71	623	636	647	657	665
SUA	All Codes	40645	645.93	22.35	623	636	648	659	669
SWD/SUA	SUA=504 Plan codes	27456	644.39	21.47	623	634	647	657	665
ELL/SUA	SUA=ELL codes	7629	636.92	23.27	617	628	641	650	657

Grade 8

Scale score statistics and N-counts of demographic groups for Grade 8 are presented in Table 52. The population scale score mean was 661.09 with a standard deviation of 30.82. 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 standard deviation below the population mean. SWD and SUA subgroups scored approximately one standard deviation below the mean scale score for the population. The ELL subgroup, which had a scale score mean just below 38 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 659: Female (662), Asian (670), White (666), Average Needs districts (666), and Low Needs districts (674).

Table 52. 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	207083	661.09	30.82	627	644	659	679	691
Gender	Female	101526	665.42	30.89	632	647	662	679	700
	Male	105557	656.93	30.17	625	642	656	674	691
Ethnicity	Asian	14982	671.23	34.68	634	653	670	691	700
	Black	39551	649.59	26.14	622	634	650	662	679
	Hispanic	41871	649.28	27.96	620	634	650	666	679
	American Indian	1012	652.51	26.37	622	637	653	666	684
	Multi-Racial	301	664.72	33.10	637	647	659	679	700
	White	109299	668.46	30.13	637	653	666	684	700
	Unknown	67	664.10	25.64	634	644	662	684	700
NRC	New York City	71642	653.17	29.58	622	637	653	670	684
	Big 4 Cities	7686	644.48	29.48	612	630	644	659	674
	High Needs Urban/Suburban	15544	654.06	28.02	622	639	653	670	684
	High Needs Rural	12362	659.23	27.68	630	644	659	674	691
	Average Needs	65108	666.63	29.06	637	650	666	679	700
	Low Needs	31981	676.70	30.74	647	659	674	691	717
	Charter	2164	658.01	24.29	632	644	656	670	684
ELL	ELL = Y	9730	623.25	29.98	595	612	627	639	650
SWD	All Codes	30617	632.96	26.61	606	620	634	647	659
SUA	All Codes	41006	634.34	27.74	606	622	637	650	662
SWD/SUA	SUA=504 Plan codes	27446	632.25	26.22	606	620	634	647	659
ELL/SUA	SUA=ELL codes	8039	623.62	29.74	595	612	627	639	650

Performance Level Distribution Summary

Percentage of students in each performance level was computed based on performance levels scale score ranges established during the 2006 Standard Setting for all grades except Grade 3,

6, and 7, Level IV. The adjustment of the Level IV cuts scores in these grades is based on a NYSED policy decision in 2008. Otherwise, a perfect raw score will be required for a student to be classified into Level IV. In 2008, the NYS Technical Advisory Group endorsed a policy decision by NYSED to adjust the Level IV cut in future test administration so that students were not required to earn a perfect raw score in order to achieve a Level IV. Information on the cut score adjustment was posted on the NYSED web site at http://www.emsc.nysed.gov/irts/ela-math/ela-math-08/2009ELAScaleScoretoPerformanceLevels.html

Table 53 shows the ELA cut scores used for classification of students to the four performance level categories in 2009.

Table 53. ELA Grades 3–8 Performance Level Cut Scores

Grade	Level II Cut	Level III Cut	Level IV Cut
3	616	650	720
4	612	650	716
5	608	650	711
6	598	650	696
7	600	650	705
8	602	650	715

Tables 54–60 show the performance level distribution for all examinees from public and charter school with valid scores. Table 54 presents performance level data for total populations of students in Grades 3–8. Tables 55–60 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 54. 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	198123	4.75	19.37	65.17	10.72	75.89
4	195634	4.28	18.76	69.69	7.27	76.96
5	197522	0.62	17.09	68.72	13.57	82.29
6	197674	0.13	18.87	71.98	9.02	81.00
7	202400	0.42	19.15	73.51	6.91	80.42
8	207083	1.72	29.66	63.75	4.87	68.62

Grade 3

Performance level distributions and N-counts of demographic groups for Grade 3 are presented in Table 55. Statewide, 75.89% of third-graders were Level III or Level IV. 6.20% of Male students were Level I, as compared to only 3.22% of Female students. The percentage of students in Levels III and IV varied widely by ethnicity and NRC subgroups. About 90% of Low Needs district students and about 87% 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 26%–46% of students who were in Level I or Level II. About one-fifth of students with ELL, SWD, or SUA status were in Level I and only about 2% 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 55. 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	198123	4.75	19.37	65.17	10.72	75.89
Gender	Female	96707	3.22	16.48	67.69	12.61	80.30
	Male	101416	6.20	22.12	62.77	8.91	71.68
Ethnicity	Asian	15808	2.07	10.64	70.31	16.99	87.30
	Black	37517	7.72	28.88	58.57	4.83	63.40
	Hispanic	42112	7.90	28.31	59.00	4.78	63.78
	American Indian	934	5.14	29.12	60.60	5.14	65.74
	Multi-Racial	642	2.34	15.42	70.72	11.53	82.25
	White	101012	2.76	13.41	69.39	14.44	83.83
	Unknown	98	3.06	12.24	71.43	13.27	84.70
NRC	New York City	70202	6.56	24.10	61.42	7.92	69.34
	Big 4 Cities	8193	11.20	34.93	50.52	3.34	53.86
	High Needs Urban/Suburban	16107	6.36	24.35	62.23	7.06	69.29
	High Needs Rural	11493	4.68	21.96	66.14	7.21	73.35
	Average Needs	58500	3.05	15.22	69.35	12.38	81.73
	Low Needs	29916	1.27	8.31	70.65	19.77	90.42
	Charter	3493	1.40	19.15	71.66	7.79	79.45
ELL	ELL = Y	17491	14.04	38.34	46.17	1.45	47.62
SWD	All Codes	26929	22.75	40.54	34.91	1.80	36.71
SUA	All Codes	45137	16.31	36.90	44.46	2.33	46.79
SWD/SUA	SUA=504 Plan codes	22779	25.15	42.35	31.30	1.20	32.50
ELL/SUA	SUA=ELL codes	15579	12.52	38.37	47.63	1.48	49.11

Grade 4

Performance level distributions and N-counts of demographic groups for Grade 4 are presented in Table 56. Across New York, approximately 77% 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 (91.90%), and the SWD subgroup had the lowest (37.66%). Students in the Black, Hispanic, and American Indian subgroups had percentages classified in Levels III and IV below 70%, which was more than 10% below the other ethnic subgroups. Nearly twice as many Big 4 City students were in Level I than the population. About a fifth of the students with ELL, SWD, or SUA status were in Level I (over three times the amount of the Statewide rate of 4.28%) 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 56. 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	195634	4.28	18.76	69.69	7.27	76.96
Gender	Female	95246	2.99	16.79	71.01	9.20	80.21
	Male	100388	5.50	20.62	68.43	5.45	73.88
Ethnicity	Asian	14588	2.02	11.26	72.77	13.95	86.72
	Black	37452	6.45	28.57	62.06	2.93	64.99
	Hispanic	41415	7.04	28.11	62.14	2.71	64.85
	American Indian	916	5.79	25.22	66.05	2.95	69.00
	Multi-Racial	486	4.53	15.84	71.60	8.02	79.62
	White	100696	2.66	12.31	75.20	9.84	85.04
	Unknown	81	0.00	12.35	76.54	11.11	87.65
NRC	New York City	68638	5.97	25.16	63.63	5.24	68.87
	Big 4 Cities	7943	9.11	33.58	54.83	2.48	57.31
	High Needs Urban/Suburban	15738	5.24	22.19	68.33	4.24	72.57
	High Needs Rural	11475	4.94	20.61	69.72	4.73	74.45
	Average Needs	58934	2.72	13.89	75.27	8.13	83.40
	Low Needs	29766	1.30	6.80	77.33	14.57	91.90
	Charter	2905	1.76	21.41	73.46	3.37	76.83
ELL	ELL = Y	14382	14.74	44.33	40.72	0.20	40.92
SWD	All Codes	28896	21.16	41.18	37.12	0.54	37.66
SUA	All Codes	45648	15.61	38.47	45.22	0.71	45.93
SWD/SUA	SUA=504 Plan codes	25827	22.73	42.45	34.50	0.32	34.82
ELL/SUA	SUA=ELL codes	12627	13.13	44.91	41.78	0.17	41.96

Grade 5

Performance level distributions and N-counts of demographic groups for Grade 5 are presented in Table 57. About 82% 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 (93.90%). Fewer Male students were in the Level I category than was observed with Grades 3 and 4, by a few percentage points. Students in the American Indian, Black, and Hispanic subgroups had rates around 72% of students classified in Levels III and IV, approximately 10% less than other ethnic subgroups. Over three times as many Big 4

City students were in Level I than the population's rate. About 3% of the students with ELL, SWD, or SUA status were in Level I (approximately four to five times as many as the Statewide rate of 0.62%), yet only about 50% 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, High Needs Rural, Average Needs districts, and Low Needs districts.

Table 57. 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	197522	0.62	17.09	68.72	13.57	82.29
Gender	Female	96843	0.44	15.36	69.78	14.42	84.20
	Male	100679	0.79	18.76	67.70	12.76	80.46
Ethnicity	Asian	14523	0.47	10.25	67.44	21.83	89.27
	Black	37980	0.95	28.72	64.63	5.70	70.33
	Hispanic	41408	1.16	27.17	65.13	6.54	71.67
	American Indian	953	0.52	22.88	67.58	9.02	76.60
	Multi-Racial	486	0.41	14.40	70.99	14.20	85.19
	White	102073	0.30	9.62	71.87	18.21	90.08
	Unknown	99	1.01	12.12	67.68	19.19	86.87
NRC	New York City	68531	0.98	24.43	64.83	9.77	74.60
	Big 4 Cities	7506	1.83	34.73	58.35	5.09	63.44
	High Needs Urban/Suburban	15484	0.65	20.71	69.24	9.40	78.64
	High Needs Rural	11477	0.48	15.15	73.29	11.08	84.37
	Average Needs	60033	0.29	10.95	72.53	16.23	88.76
	Low Needs	30582	0.15	5.95	70.87	23.03	93.90
	Charter	3613	0.17	24.74	69.00	6.09	75.09
ELL	ELL = Y	12309	3.54	53.64	41.83	0.98	42.81
SWD	All Codes	30705	3.00	48.48	46.35	2.18	48.53
SUA	All Codes	46138	2.41	42.89	52.05	2.65	54.70
SWD/SUA	SUA=504 Plan codes	27907	3.12	49.95	45.10	1.82	46.92
ELL/SUA	SUA=ELL codes	10483	3.14	52.05	43.78	1.04	44.82

Grade 6

Performance level distributions and N-counts of demographic groups for Grade 6 are presented in Table 58. Statewide, 81% 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 (93.58%), and the ELL, SWD, and SUA subgroups had the fewest. Students in the American Indian, Black, and Hispanic subgroups had around 70% 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 percentage of students placed in Level I for all subgroups are less than 1%. 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 58. 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	197674	0.13	18.87	71.98	9.02	81.00
Gender	Female	96510	0.07	15.56	72.57	11.81	84.38
	Male	101164	0.19	22.03	71.42	6.36	77.78
Ethnicity	Asian	14726	0.12	10.24	70.61	19.03	89.64
	Black	38107	0.15	30.48	66.10	3.27	69.37
	Hispanic	40522	0.22	30.06	65.90	3.82	69.72
	American Indian	902	0.33	25.61	69.51	4.55	74.06
	Multi-Racial	431	0.46	13.46	74.71	11.37	86.08
	White	102913	0.08	11.38	76.75	11.79	88.54
	Unknown	73	0.00	8.22	78.08	13.70	91.78
NRC	New York City	68151	0.23	27.19	66.06	6.52	72.58
	Big 4 Cities	7490	0.19	34.43	62.67	2.71	65.38
	High Needs Urban/Suburban	15067	0.15	23.42	71.57	4.86	76.43
	High Needs Rural	11422	0.05	18.46	75.07	6.41	81.48
	Average Needs	61034	0.05	12.42	77.10	10.42	87.52
	Low Needs	30969	0.03	6.40	76.63	16.95	93.58
	Charter	3215	0.00	23.61	73.06	3.33	76.39
ELL	ELL = Y	10414	0.85	63.46	35.46	0.22	35.68
SWD	All Codes	30362	0.64	55.02	43.77	0.58	44.35
SUA	All Codes	41788	0.55	50.21	48.43	0.82	49.25
SWD/SUA	SUA=504 Plan codes	27009	0.66	56.61	42.36	0.37	42.73
ELL/SUA	SUA=ELL codes	8248	0.86	61.42	37.50	0.22	37.72

Grade 7

Performance level distributions and N-counts of demographic groups for Grade 7 are presented in Table 59. In Grade 7, 80.42% of the students were in Levels III and IV. Over 6% more Female than Male students were classified in these two proficiency levels. Close to 40% of Big 4 Cities students were in Levels I and II. Above 94% of Low Needs students were in Levels III and IV. About 25% 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 57–75% 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, High Needs Rural, Average Needs districts, and Low Needs districts.

Table 59. 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	202400	0.42	19.15	73.51	6.91	80.42
Gender	Female	98510	0.27	16.03	75.55	8.15	83.70
	Male	103890	0.57	22.11	71.59	5.74	77.33
Ethnicity	Asian	14822	0.57	12.21	74.58	12.64	87.22
	Black	38408	0.51	32.37	65.31	1.81	67.12
	Hispanic	41120	0.84	32.26	64.94	1.96	66.90
	American Indian	935	0.86	25.24	71.12	2.78	73.90
	Multi-Racial	400	0.25	16.00	73.25	10.50	83.75
	White	106633	0.21	10.27	79.64	9.88	89.52
	Unknown	82	1.22	8.54	80.49	9.76	90.25
NRC	New York City	69611	0.69	28.53	66.58	4.20	70.78
	Big 4 Cities	7685	1.18	39.18	57.97	1.67	59.64
	High Needs Urban/Suburban	15326	0.48	25.53	70.70	3.30	74.00
	High Needs Rural	12101	0.29	17.96	76.52	5.23	81.75
	Average Needs	63001	0.16	11.42	80.09	8.33	88.42
	Low Needs	31745	0.08	5.75	80.05	14.11	94.16
	Charter	2476	0.16	21.08	75.97	2.79	78.76
ELL	ELL = Y	9603	3.76	70.78	25.42	0.04	25.46
SWD	All Codes	31155	1.82	54.58	43.19	0.40	43.59
SUA	All Codes	40645	1.88	52.50	45.05	0.56	45.61
SWD/SUA	SUA=504 Plan codes	27456	1.86	55.96	41.85	0.33	42.18
ELL/SUA	SUA=ELL codes	7629	3.71	69.75	26.50	0.04	26.54

Grade 8

Performance level distributions and N-counts of demographic groups for Grade 8 are presented in Table 60. In Grade 8, 68.62% of the students were in Levels III and IV. About 10% more Female than Male students were in Levels III or IV. Over 44% of American Indian, Black, and Hispanic students were in Levels I and II. Over 88% of Low Needs students were in Levels III and IV, while fewer than 13% 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 70% 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 60. 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	207083	1.72	29.66	63.75	4.87	68.62
Gender	Female	101526	1.05	25.33	67.38	6.23	73.62
	Male	105557	2.36	33.82	60.26	3.56	63.82
Ethnicity	Asian	14982	1.78	18.59	70.26	9.37	79.63
	Black	39551	2.41	45.65	50.54	1.39	51.94
	Hispanic	41871	3.36	44.00	51.15	1.49	52.65
	American Indian	1012	1.98	42.59	53.56	1.88	55.43
	Multi-Racial	301	1.66	23.59	70.10	4.65	74.75
	White	109299	0.83	19.79	72.54	6.84	79.38
	Unknown	67	0.00	31.34	62.69	5.97	68.66
NRC	New York City	71642	2.72	40.48	53.99	2.82	56.80
	Big 4 Cities	7686	4.83	50.72	43.20	1.26	44.46
	High Needs Urban/Suburban	15544	2.07	38.90	56.54	2.49	59.03
	High Needs Rural	12362	1.31	30.97	64.14	3.58	67.72
	Average Needs	65108	0.82	21.28	72.05	5.85	77.90
	Low Needs	31981	0.27	11.66	77.81	10.26	88.07
	Charter	2164	0.69	31.19	65.57	2.54	68.11
ELL	ELL = Y	9730	13.63	73.74	12.61	0.02	12.63
SWD	All Codes	30617	7.56	67.47	24.81	0.16	24.97
SUA	All Codes	41006	7.57	64.56	27.58	0.29	27.87
SWD/SUA	SUA=504 Plan codes	27446	7.71	68.62	23.56	0.10	23.66
ELL/SUA	SUA=ELL codes	8039	13.48	73.55	12.94	0.02	12.96

Section IX: Longitudinal Comparison of Results

This section provides longitudinal comparison of OP scale score results on the New York State 2006–2009 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 61.

Table 61. 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	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	2009	195634	669.93	34.72	4.28	18.76	69.69	7.27	76.96
	2008	196367	666.40	39.90	7.34	21.37	62.85	8.44	71.29
	2007	197306	664.70	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	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	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	2009	202400	667.19	27.06	0.42	19.15	73.51	6.91	80.42
	2008	205946	662.30	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	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

As seen in Table 61, an increase in scale score means was observed for all ELA grades between the 2006 and 2009 test administrations. The least gain was observed for Grades 3 and 4 for which total gain was 2 and 5 scale score points, respectively, between 2006 and 2009 test administrations. The largest gain in scale score points between 2006 and 2009 test administrations was noted for Grades 5 and 7 (12 and 15 scale score points, respectively). Grades 6 and 8 gained around 11 scale score points. Relatively steady yearly gain was

noticed for Grades 5, 7, and 8 with the overall population mean scale score increase of 11 or more scale points between years 2006 and 2009. An increase of approximately 5 scale score points was observed for Grade 6 between years 2006 and 2007. No score change was noticed for Grade 6 between administration years 2007 and 2008, but another 6 scale score points was observed between years 2008 and 2009. 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. Overall, the mean scale score increase for Grades 3 and 4 was less than 5 scale score point between administration years 2006 and 2009.

The variability of scale score distribution decreased steadily across years for ELA Grades 6, 7, and 8. 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. For Grades 3 and 4, the variability of scale score distribution decreased in 2009. The standard deviations for these grades decreased from about 40 scale score points in 2006, 2007, and 2008 to approximately 35 points in 2009. 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.

Following evaluation of the pattern of means scale score change between the 2006 and 2009 ELA Test administrations, a longitudinal trend of proficiency score distribution was evaluated. The percentage of students classified in Levels III and IV increased only slightly for Grades 3 and 4 between years 2006 and 2008 and increased about 6% between years 2008 and 2009. Grades 5 and 7 proficiency score trends indicated slight increases (approximately 1%) in the percentage of students classified in Levels III and IV between years 2006 and 2007 and a larger increase in the percentage of students classified in Levels III and IV between years 2007 and 2008. The percentage of Grade 5 students classified in Levels III and IV increased from approximately 68% to 78% between years 2007 and 2008, and the percentage of Grade 7 students classified in Levels III and IV increased from approximately 58% to 70% between years 2007 and 2008. Between years 2008 and 2009, another 5% of Grade 5 students and 10% of Grade 7 students were classified in Levels III and IV. The percentage of Grade 6 students classified in Levels III and IV was observed to be steadily increasing (approximately 3% each year) between the 2006 and 2008 test administrations and jumped 14% during years 2008 and 2009. It was also observed that while approximately 8% more Grade 8 students were classified in Levels III and IV in 2007 than in 2006, no increase in the percentage of students classified in the two highest proficiency levels was observed between years 2007 and 2008, but a 13% increase was observed between years 2008 and 2009.

Overall, the percentage of students classified in Levels III and IV increased least for ELA Grades 3 and 4 between years 2006 and 2009 (7% and 9% respectively). The other ELA grades' proficiency score trends indicated much larger increases in the percentage of students classified in Levels III and IV between years 2006 and 2009: 15% for Grade 5, 20% for Grades 6 and 8, and 24% for Grade 7.

In summary, the mean scale score change and the change in percentage of students classified in Levels III and IV was not uniform across grades during the four years of test administrations. As expected, the mean scale score change was found to be in alignment with the performance levels score trend between years 2006 and 2009.

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 2009 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 2009 Operational Test Passages

Title	Passage Type	Word Count	Avg. Prose Score	Fry Graph	Spache/ Dale-Chall*	Flesch-Kincaid Formula
GRADE 3						
Book 1 (Reading)						
Chicks & Chickens	Info	235	4.77	3.77	2.89	3.12
If It Was Sunlight Shining	Lit-Poem	80	n/a	n/a	n/a	n/a
Float in the Ocean	Info-How to	175	5.31	4.42	2.94	3.71
Harold's Hundred Days of School	Lit-Fiction	435	3.69	2.42	2.74	1.85
Readability Averages			**4.59**	**3.54**	**2.86**	**2.89**
Book 2 (Listening)						
More Than Leaves	Lit-Fiction	395	3.71	2.12	2.39	1.54
GRADE 4						
Book 1 (Reading)						
Comets	Info	290	5.09	4.43	3.52	3.69
The Tortoise, the Hare, and the Penguin	Lit-Fiction	490	6.40	6.04	4.00	4.85
Your Nose Knows	Info-How to	425	5.51	4.73	2.90	3.97
The Missing Homework	Lit-Poem	110	n/a	n/a	n/a	n/a
Geneva	Lit-Fiction	315	5.07	4.10	2.99	3.45

(Continued on next page)

Table A1. Readability Summary Information for 2009 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)						
The Bell That Knew the Truth	Lit-Fiction	510	5.73	4.69	3.60	3.98
Book 3 (Reading pair)						
Swan Song	Lit-Fiction	315	3.00	1.80	2.97	1.05
After the Error	Info-Essay	355	4.36	3.23	3.03	2.66
Readability Averages			**4.90**	**4.06**	**3.24**	**3.28**
GRADE 5						
Book 1 (Reading)						
Lion at School	Lit-Fiction	535	5.80	5.15	5–6	4.34
Waiting for the Little Penguins	Info-Article	465	6.67	6.17	7–8	4.98
Frozen Bubbles	Info-Article	315	6.47	6.51	5–6	5.36
The Red Fox	Lit-Fiction	590	6.16	5.75	5–6	4.62
Readability Averages			**6.28**	**5.90**	**5–6**	**4.83**
Book 2 (Listening)						
Snorkeling for Bass	Info-Essay	515	5.59	5.12	7–8	4.22
GRADE 6						
Book 1 (Reading)						
The Wise Fools of Gotham	Lit-Fiction	690	6.35	6.23	7–8	5.04
A Boy Who Makes a Difference	Info-Article	445	8.51	8.33	7–8	8.01
The Wolf at My Window	Lit-Fiction	710	5.50	4.52	7–8	3.79
The Clever Crow	Info-Article	445	7.25	7.06	7–8	6.19
Book 1 (Reading)						
Under the Rice Moon	Lit-Fiction	645	7.02	6.84	5–6	6.08
Book 2 (Listening)						
About Me	Lit-Fiction	655	5.89	5.70	5–6	4.65
Book 3 (Reading pair)						
Climbin' Ryan	Info-Article	515	7.69	7.56	9–10	6.78
Natalya's Happy Hugged Hens	Info-Article	515	7.89	7.76	7–8	7.32
Readability Averages			**7.17**	**6.90**	**7–8**	**6.17**

(Continued on next page)

Table A1. Readability Summary Information for 2009 Operational Test Passages (cont.)

Title	Passage Type	Word Count	Avg. Prose Score	Fry Graph	Spache/ Dale-Chall*	Flesch-Kincaid Formula
GRADE 7						
Book 1 (Reading)						
Hearing Voices	Info-Article	335	10.07	9.63	9–10	8.44
Diondra	Lit-Fiction	635	3.89	2.83	5–6	2.21
The Beagle Brigade	Info-Article	280	9.31	9.56	11–12	7.78
Young Author Profile: Amelia Atwater-Rhodes	Info-Article	605	6.08	5.66	7–8	4.64
The Herring Choker	Lit-Fiction	605	4.05	3.18	5–6	2.56
Readability Averages			**6.68**	**6.17**	**7–8**	**5.13**
Book 2 (Listening)						
Growing Up … and Up	Info-Bio	395	9.37	8.78	7–8	8.46
GRADE 8						
Book 1 (Reading)						
Old Champ	Lit-Fiction	670	5.47	4.56	7–8	3.85
Why We Play	Info-Article	420	8.93	8.44	9–10	7.81
Mother Has a Job	Lit-Fiction	355	9.30	8.96	9–10	9.14
Winter Dark	Lit-Poem	40	n/a	n/a	n/a	n/a
Rescue at Pea Island	Info-Article	800	8.23	7.72	9–10	6.82
Book 2 (Listening)						
Music to His Ears	Info-Article	485	9.23	8.36	9–10	8.08
Book 3 (Reading pair)						
Drawing Calvin and Hobbes	Info-Essay	915	9.47	9.01	9–10	9.15
Lucky Break	Info-Essay	535	6.36	6.30	7–8	5.08
Readability Averages			**7.96**	**7.50**	**9–10**	**6.98**

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
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 2009 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 2009 item selection will include 2006, 2007, and 2008 FT items for Grades 3, 5, 6, and 7 and 2003, 2006, 2007, and 2008 FT items for Grades 4 and 8. 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 it is 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 2009 OP forms and the 2008 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, 2009 full selection, 2009 MC items only)
    - Item number in 2009 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**	**6.32**	**22.56**	**22.56**
		2	1.19	4.25	26.81
		3	1.06	3.79	30.60
	SWD	**1**	**6.88**	**24.56**	**24.56**
		2	1.32	4.73	29.28
		3	1.10	3.92	33.20
	SUA	**1**	**6.85**	**24.46**	**24.46**
		2	1.25	4.46	28.91
		3	1.09	3.90	32.81
	SWD /SUA	**1**	**6.74**	**24.08**	**24.08**
		2	1.38	4.92	29.00
		3	1.08	3.85	32.85
	ELL /SUA	**1**	**6.23**	**22.25**	**22.25**
		2	1.20	4.28	26.53
		3	1.07	3.81	30.34
4	ELL	**1**	**5.99**	**19.31**	**19.31**
		2	1.19	3.84	23.15
		3	1.10	3.55	26.70
		4	1.01	3.25	29.95
	SWD	**1**	**6.92**	**22.34**	**22.34**
		2	1.20	3.87	26.20
		3	1.07	3.46	29.66
	SUA	**1**	**6.88**	**22.20**	**22.20**
		2	1.21	3.91	26.11
		3	1.07	3.44	29.54
	SWD /SUA	**1**	**6.89**	**22.22**	**22.22**
		2	1.19	3.83	26.06
		3	1.09	3.51	29.57

(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	ELL /SUA	**1**	**5.85**	**18.87**	**18.87**
		2	1.18	3.81	22.68
		3	1.11	3.57	26.25
		4	1.01	3.27	29.52
5	ELL	**1**	**5.08**	**18.83**	**18.83**
		2	1.20	4.43	23.26
		3	1.09	4.02	27.28
		4	1.05	3.88	31.16
		5	1.01	3.73	34.89
	SWD	**1**	**5.61**	**20.78**	**20.78**
		2	1.28	4.74	25.51
		3	1.03	3.83	29.34
		4	1.03	3.80	33.14
	SUA	**1**	**5.61**	**20.78**	**20.78**
		2	1.25	4.62	25.40
		3	1.05	3.87	29.27
		4	1.02	3.79	33.06
	SWD /SUA	**1**	**5.61**	**20.78**	**20.78**
		2	1.30	4.81	25.58
		3	1.03	3.83	29.41
		4	1.03	3.80	33.22
	ELL /SUA	**1**	**5.05**	**18.70**	**18.70**
		2	1.18	4.37	23.07
		3	1.09	4.05	27.12
		4	1.05	3.90	31.02
		5	1.01	3.75	34.77
6	ELL	**1**	**5.89**	**20.31**	**20.31**
		2	1.22	4.21	24.52
		3	1.11	3.83	28.35
		4	1.07	3.70	32.05
	SWD	**1**	**6.35**	**21.90**	**21.90**
		2	1.22	4.22	26.12
		3	1.17	4.04	30.15
		4	1.03	3.55	33.70
	SUA	**1**	**6.47**	**22.31**	**22.31**
		2	1.21	4.17	26.48
		3	1.14	3.92	30.39
		4	1.05	3.62	34.01

(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**	**6.38**	**21.99**	**21.99**
		2	1.24	4.28	26.26
		3	1.16	4.00	30.27
		4	1.03	3.55	33.82
	ELL /SUA	**1**	**5.94**	**20.48**	**20.48**
		2	1.22	4.20	24.68
		3	1.12	3.85	28.53
		4	1.06	3.67	32.19
7	ELL	**1**	**6.77**	**19.33**	**19.33**
		2	1.33	3.79	23.13
		3	1.14	3.26	26.39
		4	1.10	3.13	29.52
		5	1.05	3.00	32.52
	SWD	**1**	**7.38**	**21.09**	**21.09**
		2	1.32	3.78	24.88
		3	1.22	3.48	28.35
		4	1.06	3.03	31.38
		5	1.04	2.96	34.35
	SUA	**1**	**7.60**	**21.72**	**21.72**
		2	1.34	3.81	25.53
		3	1.19	3.39	28.92
		4	1.07	3.04	31.96
		5	1.02	2.90	34.86
	SWD /SUA	**1**	**7.44**	**21.25**	**21.25**
		2	1.32	3.77	25.02
		3	1.25	3.56	28.58
		4	1.05	3.01	31.59
		5	1.04	2.98	34.57
	ELL /SUA	**1**	**6.83**	**19.52**	**19.52**
		2	1.33	3.79	23.31
		3	1.14	3.25	26.56
		4	1.09	3.12	29.68
		5	1.05	2.99	32.67
		6	1.01	2.90	35.57
8	ELL	**1**	**5.27**	**18.16**	**18.16**
		2	1.22	4.21	22.37
		3	1.08	3.72	26.09
		4	1.05	3.63	29.72
		5	1.02	3.53	33.25

(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**	**5.78**	**19.93**	**19.93**
		2	1.17	4.02	23.95
		3	1.09	3.75	27.70
		4	1.04	3.59	31.29
	SUA	**1**	**5.97**	**20.58**	**20.58**
		2	1.16	4.00	24.58
		3	1.05	3.62	28.19
		4	1.03	3.55	31.74
	SWD /SUA	**1**	**5.76**	**19.84**	**19.84**
		2	1.18	4.06	23.90
		3	1.09	3.75	27.65
		4	1.04	3.60	31.25
	ELL /SUA	**1**	**5.31**	**18.31**	**18.31**
		2	1.22	4.19	22.50
		3	1.08	3.72	26.22
		4	1.06	3.67	29.89
		5	1.02	3.52	33.41

Appendix E—Items Flagged for DIF

These tables support the DIF information in Section V, "Operational Test Data Collection and Classical Analyses," 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 2009 Classical DIF Item Flags

Grade	Item #	Subgroup	DIF	SMD	Mantel-Haenszel	Delta
4	4	Female	Against	-0.10	No Flag	No Flag
4	12	Asian	Against	No Flag	181.83	-0.46
4	29	ELL	Against	-0.12	No Flag	No Flag
4	31	Asian	In Favor	0.10	No Flag	No Flag
4	31	Female	In Favor	0.10	No Flag	No Flag
5	9	Female	Against	-0.10	No Flag	No Flag
5	23	Asian	Against	No Flag	979.75	-2.01
5	24	Asian	Against	No Flag	723.95	-1.91
5	27	Asian	In Favor	0.10	n/a	n/a
5	27	Female	In Favor	0.11	n/a	n/a
5	27	ELL	Against	-0.16	n/a	n/a
6	4	Female	Against	No Flag	1696.15	-1.69
6	11	ELL	Against	-0.12	No Flag	No Flag
6	18	ELL	In Favor	0.10	No Flag	No Flag
6	28	Female	In Favor	0.16	n/a	n/a
6	29	Female	In Favor	0.10	n/a	n/a
6	29	ELL	Against	-0.10	n/a	n/a
7	11	Asian	In Favor	0.11	No Flag	No Flag
7	18	ELL	In Favor	0.11	n/a	n/a
7	21	Hispanic	Against	No Flag	1062.98	-1.53
7	21	Asian	Against	No Flag	629.14	-1.75
7	25	Black	Against	-0.13	No Flag	No Flag
7	25	Hispanic	Against	-0.10	No Flag	No Flag
7	25	Asian	Against	-0.15	1471.44	-1.66
7	35	Black	Against	-0.10	n/a	n/a
7	35	Hispanic	Against	-0.10	n/a	n/a
7	35	ELL	Against	-0.19	n/a	n/a
8	4	ELL	Against	-0.10	No Flag	No Flag
8	9	ELL	In Favor	0.10	No Flag	No Flag

(Continued on next page)

Table E1. NYSTP ELA 2009 Classical DIF Item Flags (cont.)

Grade	Item #	Subgroup	DIF	SMD	Mantel-Haenszel	Delta
8	14	Black	In Favor	0.10	No Flag	No Flag
8	14	Asian	In Favor	0.11	No Flag	No Flag
8	18	Black	In Favor	0.10	No Flag	No Flag
8	27	Black	Against	-0.16	n/a	n/a
8	27	Hispanic	Against	-0.11	n/a	n/a
8	27	High Needs	Against	-0.13	n/a	n/a
8	27	ELL	Against	-0.24	n/a	n/a
8	28	Females	In Favor	0.17	n/a	n/a
8	29	Females	In Favor	0.11	n/a	n/a
8	29	ELL	Against	-0.16	n/a	n/a

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})
5	27	Black	Against	-0.105
5	27	ELL	Against	-0.178
6	11	ELL	Against	-0.130
6	28	Male	Against	-0.109
7	25	Asian	Against	-0.112
7	35	ELL	Against	-0.197

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	48.71	7	187943	11.15	501.18	Y
2	3PL	189.98	7	187943	48.90	501.18	Y
3	3PL	140.75	7	187943	35.75	501.18	Y
4	3PL	698.13	7	187943	184.71	501.18	Y
5	3PL	97.96	7	187943	24.31	501.18	Y
6	3PL	92.41	7	187943	22.83	501.18	Y
7	3PL	225.27	7	187943	58.34	501.18	Y
8	3PL	187.25	7	187943	48.17	501.18	Y
9	3PL	316.35	7	187943	82.68	501.18	Y
10	3PL	789.65	7	187943	209.17	501.18	Y
11	3PL	202.36	7	187943	52.21	501.18	Y
12	3PL	168.24	7	187943	43.09	501.18	Y
13	3PL	233.20	7	187943	60.46	501.18	Y
14	3PL	338.25	7	187943	88.53	501.18	Y
15	3PL	716.34	7	187943	189.58	501.18	Y
16	3PL	253.63	7	187943	65.91	501.18	Y
17	3PL	150.06	7	187943	38.24	501.18	Y
18	3PL	536.59	7	187943	141.54	501.18	Y
19	3PL	142.45	7	187943	36.20	501.18	Y
20	3PL	566.29	7	187943	149.48	501.18	Y
21	2PPC	2419.67	17	187943	412.05	501.18	Y
22	3PL	84.45	7	187943	20.70	501.18	Y
23	3PL	423.54	7	187943	111.33	501.18	Y
24	3PL	113.02	7	187943	28.33	501.18	Y
25	3PL	63.88	7	187943	15.20	501.18	Y
26	2PPC	459.67	17	187943	75.92	501.18	Y
27	2PPC	564.28	17	187943	93.86	501.18	Y
28	2PPC	327.18	26	187943	41.77	501.18	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	1229.34	7	189703	326.68	505.87	Y
2	3PL	79.29	7	189703	19.32	505.87	Y
3	3PL	52.08	7	189703	12.05	505.87	Y
4	3PL	84.39	7	189703	20.68	505.87	Y
5	3PL	333.97	7	189703	87.39	505.87	Y
6	3PL	390.27	7	189703	102.43	505.87	Y
7	3PL	50.18	7	189703	11.54	505.87	Y
8	3PL	209.65	7	189703	54.16	505.87	Y
9	3PL	301.92	7	189703	78.82	505.87	Y
10	3PL	329.94	7	189703	86.31	505.87	Y
11	3PL	114.51	7	189703	28.73	505.87	Y
12	3PL	123.53	7	189703	31.14	505.87	Y
13	3PL	92.16	7	189703	22.76	505.87	Y
14	3PL	408.00	7	189703	107.17	505.87	Y
15	3PL	217.26	7	189703	56.19	505.87	Y
16	3PL	42.60	7	189703	9.52	505.87	Y
17	3PL	358.78	7	189703	94.02	505.87	Y
18	3PL	170.37	7	189703	43.66	505.87	Y
19	3PL	406.88	7	189703	106.87	505.87	Y
20	3PL	67.50	7	189703	16.17	505.87	Y
21	3PL	360.03	7	189703	94.35	505.87	Y
22	3PL	68.62	7	189703	16.47	505.87	Y
23	3PL	307.71	7	189703	80.37	505.87	Y
24	3PL	203.89	7	189703	52.62	505.87	Y
25	3PL	140.43	7	189703	35.66	505.87	Y
26	3PL	1020.24	7	189703	270.80	505.87	Y
27	3PL	225.07	7	189703	58.28	505.87	Y
28	3PL	452.20	7	189703	118.98	505.87	Y
29	2PPC	2109.92	35	189703	248.00	505.87	Y
30	2PPC	2969.11	35	189703	350.69	505.87	Y
31	2PPC	936.04	26	189703	126.20	505.87	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	338.37	7	184365	88.56	491.64	Y
2	3PL	85.15	7	184365	20.89	491.64	Y
3	3PL	240.95	7	184365	62.53	491.64	Y
4	3PL	173.50	7	184365	44.50	491.64	Y
5	3PL	536.44	7	184365	141.50	491.64	Y
6	3PL	109.66	7	184365	27.44	491.64	Y
7	3PL	388.26	7	184365	101.90	491.64	Y
8	3PL	1353.02	7	184365	359.74	491.64	Y
9	3PL	774.71	7	184365	205.18	491.64	Y
10	3PL	173.11	7	184365	44.39	491.64	Y
11	3PL	101.64	7	184365	25.29	491.64	Y
12	3PL	526.41	7	184365	138.82	491.64	Y
13	3PL	127.86	7	184365	32.30	491.64	Y
14	3PL	371.31	7	184365	97.37	491.64	Y
15	3PL	366.59	7	184365	96.10	491.64	Y
16	3PL	18.51	7	184365	3.08	491.64	Y
17	3PL	195.14	7	184365	50.28	491.64	Y
18	3PL	286.60	7	184365	74.73	491.64	Y
19	3PL	574.44	7	184365	151.66	491.64	Y
20	3PL	1198.06	7	184365	318.32	491.64	Y
21	2PPC	1115.66	17	184365	188.42	491.64	Y
22	3PL	156.20	7	184365	39.87	491.64	Y
23	3PL	130.67	7	184365	33.05	491.64	Y
24	3PL	38.38	7	184365	8.39	491.64	Y
25	3PL	54.59	7	184365	12.72	491.64	Y
26	2PPC	474.58	17	184365	78.47	491.64	Y
27	2PPC	2442.19	26	184365	335.06	491.64	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	538.06	7	190378	141.93	507.67	Y
2	3PL	108.58	7	190378	27.15	507.67	Y
3	3PL	94.32	7	190378	23.34	507.67	Y
4	3PL	67.00	7	190378	16.04	507.67	Y
5	3PL	318.81	7	190378	83.34	507.67	Y
6	3PL	159.10	7	190378	40.65	507.67	Y
7	3PL	259.30	7	190378	67.43	507.67	Y
8	3PL	195.68	7	190378	50.43	507.67	Y
9	3PL	97.76	7	190378	24.26	507.67	Y
10	3PL	206.67	7	190378	53.36	507.67	Y
11	3PL	1856.47	7	190378	494.29	507.67	Y
12	3PL	111.27	7	190378	27.87	507.67	Y
13	3PL	162.02	7	190378	41.43	507.67	Y
14	3PL	89.81	7	190378	22.13	507.67	Y
15	3PL	248.93	7	190378	64.66	507.67	Y
16	3PL	232.33	7	190378	60.22	507.67	Y
17	3PL	97.52	7	190378	24.19	507.67	Y
18	3PL	392.07	7	190378	102.91	507.67	Y
19	3PL	283.79	7	190378	73.97	507.67	Y
20	3PL	244.85	7	190378	63.57	507.67	Y
21	3PL	1261.69	7	190378	335.33	507.67	Y
22	3PL	550.92	7	190378	145.37	507.67	Y
23	3PL	385.46	7	190378	101.15	507.67	Y
24	3PL	630.84	7	190378	166.73	507.67	Y
25	3PL	146.32	7	190378	37.24	507.67	Y
26	3PL	545.82	7	190378	144.01	507.67	Y
27	2PPC	5291.90	44	190378	559.43	507.67	N
28	2PPC	6128.82	44	190378	648.64	507.67	N
29	2PPC	2042.63	26	190378	279.66	507.67	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	144.59	7	197390	36.77	526.37	Y
2	3PL	315.93	7	197390	82.57	526.37	Y
3	3PL	51.47	7	197390	11.89	526.37	Y
4	3PL	165.80	7	197390	42.44	526.37	Y
5	3PL	82.99	7	197390	20.31	526.37	Y
6	3PL	73.35	7	197390	17.73	526.37	Y
7	3PL	160.80	7	197390	41.11	526.37	Y
8	3PL	50.83	7	197390	11.72	526.37	Y
9	3PL	69.50	7	197390	16.70	526.37	Y
10	3PL	155.70	7	197390	39.74	526.37	Y
11	3PL	152.30	7	197390	38.83	526.37	Y
12	3PL	313.83	7	197390	82.00	526.37	Y
13	3PL	129.11	7	197390	32.64	526.37	Y
14	3PL	140.60	7	197390	35.71	526.37	Y
15	3PL	245.31	7	197390	63.69	526.37	Y
16	3PL	1292.12	7	197390	343.46	526.37	Y
17	3PL	156.07	7	197390	39.84	526.37	Y
18	3PL	206.43	7	197390	53.30	526.37	Y
19	3PL	432.22	7	197390	113.65	526.37	Y
20	3PL	113.79	7	197390	28.54	526.37	Y
21	3PL	117.50	7	197390	29.53	526.37	Y
22	3PL	119.19	7	197390	29.98	526.37	Y
23	3PL	153.65	7	197390	39.19	526.37	Y
24	3PL	151.77	7	197390	38.69	526.37	Y
25	3PL	470.18	7	197390	123.79	526.37	Y
26	3PL	319.58	7	197390	83.54	526.37	Y
27	2PPC	839.00	17	197390	140.97	526.37	Y
28	2PPC	500.38	17	197390	82.90	526.37	Y
29	3PL	406.04	7	197390	106.65	526.37	Y
30	3PL	319.41	7	197390	83.50	526.37	Y
31	3PL	265.29	7	197390	69.03	526.37	Y
32	3PL	156.05	7	197390	39.83	526.37	Y
33	2PPC	212.17	17	197390	33.47	526.37	Y
34	2PPC	420.37	17	197390	69.18	526.37	Y
35	2PPC	2073.91	26	197390	283.99	526.37	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	559.76	7	202831	147.73	540.88	Y
2	3PL	140.52	7	202831	35.69	540.88	Y
3	3PL	86.71	7	202831	21.30	540.88	Y
4	3PL	42.22	7	202831	9.41	540.88	Y
5	3PL	116.15	7	202831	29.17	540.88	Y
6	3PL	109.14	7	202831	27.30	540.88	Y
7	3PL	179.78	7	202831	46.18	540.88	Y
8	3PL	23.61	7	202831	4.44	540.88	Y
9	3PL	152.19	7	202831	38.80	540.88	Y
10	3PL	75.83	7	202831	18.39	540.88	Y
11	3PL	265.04	7	202831	68.96	540.88	Y
12	3PL	364.83	7	202831	95.63	540.88	Y
13	3PL	93.90	7	202831	23.22	540.88	Y
14	3PL	196.42	7	202831	50.62	540.88	Y
15	3PL	213.85	7	202831	55.28	540.88	Y
16	3PL	329.22	7	202831	86.12	540.88	Y
17	3PL	360.51	7	202831	94.48	540.88	Y
18	3PL	265.31	7	202831	69.04	540.88	Y
19	3PL	252.52	7	202831	65.62	540.88	Y
20	3PL	897.81	7	202831	238.08	540.88	Y
21	3PL	494.91	7	202831	130.40	540.88	Y
22	3PL	262.35	7	202831	68.24	540.88	Y
23	3PL	89.68	7	202831	22.10	540.88	Y
24	3PL	1064.70	7	202831	282.68	540.88	Y
25	3PL	392.71	7	202831	103.08	540.88	Y
26	3PL	149.82	7	202831	38.17	540.88	Y
27	2PPC	5139.55	44	202831	543.19	540.88	N
28	2PPC	4433.82	44	202831	467.96	540.88	Y
29	2PPC	1400.51	26	202831	190.61	540.88	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—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 I1. Grade 3 ELA 2009 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
475	387	0.20	387	0.20
549	348	0.18	735	0.37
569	451	0.23	1186	0.60
580	642	0.32	1828	0.92
589	831	0.42	2659	1.34
595	1007	0.51	3666	1.85
601	1176	0.59	4842	2.44
606	1380	0.70	6222	3.14
610	1489	0.75	7711	3.89
614	1691	0.85	9402	4.75
618	1969	0.99	11371	5.74
622	2141	1.08	13512	6.82
625	2435	1.23	15947	8.05
628	2701	1.36	18648	9.41
631	3101	1.57	21749	10.98
635	3672	1.85	25421	12.83
638	4277	2.16	29698	14.99
641	5013	2.53	34711	17.52
644	5839	2.95	40550	20.47
648	7222	3.65	47772	24.11
651	8551	4.32	56323	28.43
655	10767	5.43	67090	33.86
660	13012	6.57	80102	40.43
665	15540	7.84	95642	48.27
670	18596	9.39	114238	57.66

(Continued on next page)

Table I1. Grade 3 ELA 2009 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
677	20904	10.55	135142	68.21
686	21760	10.98	156902	79.19
698	19990	10.09	176892	89.28
720	14597	7.37	191489	96.65
780	6634	3.35	198123	100.00

Table I2. Grade 4 ELA 2009 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
430	352	0.18	352	0.18
536	281	0.14	633	0.32
558	420	0.21	1053	0.54
570	538	0.28	1591	0.81
578	605	0.31	2196	1.12
585	765	0.39	2961	1.51
591	810	0.41	3771	1.93
597	924	0.47	4695	2.40
602	1112	0.57	5807	2.97
606	1234	0.63	7041	3.60
610	1335	0.68	8376	4.28
614	1609	0.82	9985	5.10
618	1829	0.93	11814	6.04
621	2024	1.03	13838	7.07
625	2288	1.17	16126	8.24
628	2619	1.34	18745	9.58
631	3130	1.60	21875	11.18
634	3527	1.80	25402	12.98
637	4080	2.09	29482	15.07
641	4580	2.34	34062	17.41
644	5149	2.63	39211	20.04
647	5863	3.00	45074	23.04
650	6418	3.28	51492	26.32

(Continued on next page)

Table I2. Grade 4 ELA 2009 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
653	7183	3.67	58675	29.99
657	7698	3.93	66373	33.93
660	8827	4.51	75200	38.44
664	9456	4.83	84656	43.27
668	10216	5.22	94872	48.49
672	10942	5.59	105814	54.09
676	11381	5.82	117195	59.91
680	12125	6.20	129320	66.10
685	12454	6.37	141774	72.47
691	12068	6.17	153842	78.64
696	11137	5.69	164979	84.33
703	9193	4.70	174172	89.03
711	7231	3.70	181403	92.73
721	6108	3.12	187511	95.85
737	5542	2.83	193053	98.68
775	2581	1.32	195634	100.00

Table I3. Grade 5 ELA 2009 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
495	165	0.08	165	0.08
561	207	0.10	372	0.19
588	319	0.16	691	0.35
600	528	0.27	1219	0.62
608	680	0.34	1899	0.96
614	866	0.44	2765	1.40
619	1076	0.54	3841	1.94
623	1267	0.64	5108	2.59
627	1462	0.74	6570	3.33
630	1872	0.95	8442	4.27
633	2165	1.10	10607	5.37
636	2565	1.30	13172	6.67

(Continued on next page)

Table I3. Grade 5 ELA 2009 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
639	2937	1.49	16109	8.16
641	3552	1.80	19661	9.95
644	4232	2.14	23893	12.10
647	5102	2.58	28995	14.68
649	5987	3.03	34982	17.71
652	7207	3.65	42189	21.36
655	8488	4.30	50677	25.66
658	10243	5.19	60920	30.84
662	12231	6.19	73151	37.03
666	14964	7.58	88115	44.61
670	17510	8.86	105625	53.48
676	20379	10.32	126004	63.79
683	22311	11.30	148315	75.09
693	22397	11.34	170712	86.43
712	17948	9.09	188660	95.51
795	8862	4.49	197522	100.00

Table I4. Grade 6 ELA 2009 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
480	77	0.04	77	0.04
558	68	0.03	145	0.07
589	110	0.06	255	0.13
599	170	0.09	425	0.22
606	231	0.12	656	0.33
611	301	0.15	957	0.48
615	404	0.20	1361	0.69
619	507	0.26	1868	0.94
622	645	0.33	2513	1.27
624	728	0.37	3241	1.64
627	876	0.44	4117	2.08
629	1021	0.52	5138	2.60
631	1242	0.63	6380	3.23

(Continued on next page)

Table I4. Grade 6 ELA 2009 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
633	1387	0.70	7767	3.93
635	1571	0.79	9338	4.72
637	1884	0.95	11222	5.68
638	2198	1.11	13420	6.79
640	2606	1.32	16026	8.11
642	3072	1.55	19098	9.66
643	3479	1.76	22577	11.42
645	4253	2.15	26830	13.57
647	4925	2.49	31755	16.06
649	5806	2.94	37561	19.00
651	6697	3.39	44258	22.39
653	8192	4.14	52450	26.53
655	9231	4.67	61681	31.20
657	11035	5.58	72716	36.79
660	12691	6.42	85407	43.21
663	14583	7.38	99990	50.58
666	15865	8.03	115855	58.61
669	16832	8.52	132687	67.12
674	16794	8.50	149481	75.62
679	16092	8.14	165573	83.76
685	14273	7.22	179846	90.98
696	11187	5.66	191033	96.64
785	6641	3.36	197674	100.00

Table I5. Grade 7 ELA 2009 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
470	150	0.07	150	0.07
541	153	0.08	303	0.15
581	218	0.11	521	0.26
594	335	0.17	856	0.42
602	473	0.23	1329	0.66

(Continued on next page)

Table I5. Grade 7 ELA 2009 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
608	570	0.28	1899	0.94
613	671	0.33	2570	1.27
617	801	0.40	3371	1.67
620	881	0.44	4252	2.10
623	1029	0.51	5281	2.61
625	1107	0.55	6388	3.16
628	1343	0.66	7731	3.82
630	1445	0.71	9176	4.53
632	1569	0.78	10745	5.31
634	1784	0.88	12529	6.19
636	2043	1.01	14572	7.20
638	2244	1.11	16816	8.31
639	2644	1.31	19460	9.61
641	3006	1.49	22466	11.10
643	3302	1.63	25768	12.73
645	3934	1.94	29702	14.67
647	4632	2.29	34334	16.96
648	5287	2.61	39621	19.58
650	6022	2.98	45643	22.55
652	7196	3.56	52839	26.11
655	8169	4.04	61008	30.14
657	9571	4.73	70579	34.87
659	10865	5.37	81444	40.24
662	12269	6.06	93713	46.30
665	14108	6.97	107821	53.27
669	15421	7.62	123242	60.89
673	16848	8.32	140090	69.21
678	17186	8.49	157276	77.71
684	16775	8.29	174051	85.99
692	14360	7.09	188411	93.09
705	9895	4.89	198306	97.98
790	4094	2.02	202400	100.00

Table I6. Grade 8 ELA 2009 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
430	249	0.12	249	0.12
554	207	0.10	456	0.22
570	281	0.14	737	0.36
579	370	0.18	1107	0.53
585	458	0.22	1565	0.76
590	564	0.27	2129	1.03
595	654	0.32	2783	1.34
599	774	0.37	3557	1.72
602	898	0.43	4455	2.15
606	1044	0.50	5499	2.66
609	1173	0.57	6672	3.22
612	1307	0.63	7979	3.85
615	1502	0.73	9481	4.58
617	1710	0.83	11191	5.40
620	1948	0.94	13139	6.34
622	2214	1.07	15353	7.41
625	2499	1.21	17852	8.62
627	3010	1.45	20862	10.07
630	3316	1.60	24178	11.68
632	3865	1.87	28043	13.54
634	4290	2.07	32333	15.61
637	4919	2.38	37252	17.99
639	5769	2.79	43021	20.77
642	6402	3.09	49423	23.87
644	7383	3.57	56806	27.43
647	8169	3.94	64975	31.38
650	9050	4.37	74025	35.75
653	9884	4.77	83909	40.52
656	10716	5.17	94625	45.69
659	11635	5.62	106260	51.31
662	11960	5.78	118220	57.09
666	12375	5.98	130595	63.06
670	12403	5.99	142998	69.05
674	12160	5.87	155158	74.93

(Continued on next page)

Table I6. Grade 8 ELA 2009 SS Frequency Distribution, State (cont.)

SS	N-count	Percent	Cumulative Frequency	Cumulative Percent
679	11864	5.73	167022	80.65
684	11020	5.32	178042	85.98
691	9903	4.78	187945	90.76
700	9049	4.37	196994	95.13
717	6993	3.38	203987	98.50
790	3096	1.50	207083	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.

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

Dorans, N.J., A.P. Schmitt, and 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. and 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, and 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, and G.R. Burket. 1989. Experiences in the application of item response theory in test construction. *Applied Measurement in Education* 2:297–312.

Huynh, H. and 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. and S. Kotz. 1970. *Distributions in Statistics: Continuous Univariate Distributions,* Vol. 2. New York: John Wiley.

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

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

Lee, W., B.A. Hanson, and 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., and 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. and 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. and M.R. Novick. 1968. *Statistical Theories of Mental Test Scores.* Menlo Park, CA: Addison-Wesley.

Mehrens, W.A. and 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., and R.D. Bock. 1991. *PARSCALE*: *Parameter Scaling of Rating Data* [Computer program]. Chicago, IL: Scientific Software, Inc.

Novick, M.R. and 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. and 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. and 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, and 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. and 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, and 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, and A. Grima. 1993. Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement* 36:225–33