New York State Testing Program 2009: Mathematics, 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, Mathematics 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 State. The Mathematics Tests target student progress toward five content standards in Grades 3–7 and four content standards in Grade 8 as described in Section II, "Test Design and Development," subsection "Content Rationale." The Grades 3–8 Mathematics 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 schools, Grades 3, 4, 5, 6, 7, and 8 (and ungraded students of equivalent age) are the target population for the Grades 3–8 Mathematics Tests. 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/elintmath.html.

Test Use and Decisions Based on Assessment

The Grades 3–8 Mathematics Tests are used to measure the extent to which individual students achieve the New York State Learning Standards in mathematics 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 Mathematics 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 Mathematics Tests at each grade level. The scale scores are comparable within each grade level but not across grades because the Grades 3–8 Mathematics Tests are not on a vertical scale. The test scores are reported at the individual level and can also be aggregated. Detailed information on derivation and properties of scale scores is provided in Section VI, "IRT Scaling and Equating." The Grades 3–8 Mathematics Test 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 Score 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 mathematics proficiency cut scores reflect the abilities intended by the New York State Education Department. Performance of students on the Grades 3–8 Mathematics 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. More 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 Mathematics* and the *NYS Measurement Review Technical Report 2006 for Mathematics*.

Standard Performance Index Scores

Standard performance index (SPI) scores are obtained from the Grades 3–8 Mathematics Tests. The SPI score is an indicator of student ability, knowledge, and skills in specific learning standards and 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 files for 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 testing program.

Camera-copy versions of the regular tests are provided to the braille vendor, who then proceeds to create the braille editions. Proofs of the braille editions are submitted to NYSED for review and approval prior to reproduction of the braille editions.

Test Translations

Since these are tests of mathematical ability, the NYSTP 3–8 Mathematics tests are translated into five other languages: Chinese, Haitian-Creole, Korean, Russian, and Spanish. These tests are translated to provide students the opportunity to demonstrate mathematical ability independent of their command of the English language. Sample tests are available in each translated language at the following locations:

http://www.emsc.nysed.gov/3-8/math-sample/chinese/home.htm (Chinese),
http://www.emsc.nysed.gov/3-8/math-sample/haitian/home.htm (Haitian-Creole),
http://www.emsc.nysed.gov/3-8/math-sample/korean/home.htm (Korean),
http://www.emsc.nysed.gov/3-8/math-sample/russian/home.htm (Russian),
http://www.emsc.nysed.gov/3-8/math-sample/spanish/home.htm (Spanish).

In addition, each year's OP test translations are released and posted to NYSED's web site after the testing administration window is over.

English language learners may be provided with an oral translation of the Mathematics Tests when a written translation is not available in the student's native language. The following testing accommodations were made available to English language learners: time extension, separate testing location, bilingual glossaries, simultaneous use of English and alternative language editions, oral translation for lower-incidence languages, and writing responses in the native language.

Section II: Test Design and Development

Test Description

The Grades 3–8 Mathematics Tests are New York State Learning Standards-based criterion-referenced tests composed of multiple-choice (MC) and constructed-response (CR) items differentiated by maximum score point. MC items have a maximum score of 1, short-response (SR) items have a maximum score of 2, and extended response (ER) items have a maximum score of 3. The tests were administered in New York State classrooms during March 2009 over a two-day period for Grades 3, 5, 6, and 7 and over a three-day period for Grades 4 and 8. 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 tests 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. Book 1 contained only MC items. Book 2 and Book 3 contained only CR items. The 2009 *Teacher's Directions* (available online at http://www.nysedregents.org/testing/elaei/09exams/home.htm) and the 2009 *School Administrator's Manual* provide more detail on security, scheduling, classroom organization and preparation, test materials, and administration.

Table 1. NYSTP Mathematics 2009 Test Configuration

Grade	Day	Book	Number of Items				Allotted Time (minutes)	
			MC	SR	ER	Total	Testing	Prep
3	1	1	25	0	0	25	45	10
	2	2	0	4	2	6	40	10
	Totals		25	4	2	31	85	20
4	1	1	30	0	0	30	50	10
	2	2	0	7	2	9	50	10
	3	3	0	7	2	9	50	10
	Totals		30	14	4	48	150	30
5	1	1	26	0	0	26	45	10
	2	2	0	4	4	8	50	10
	Totals		26	4	4	34	95	20
6	1	1	25	0	0	25	45	10
	2	2	0	6	4	10	60	10
	Totals		25	6	4	35	105	20

(Continued on next page)

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

Grade	Day	Book	Number of Items				Allotted Time (minutes)	
			MC	SR	ER	Total	Testing	Prep
7	1	1	30	0	0	30	55	10
	2	2	0	4	4	8	55	10
	Totals		30	4	4	38	110	20
8	1	1	27	0	0	27	50	10
	1	2	0	4	2	6	40	10
	2	3	0	8	4	12	70	10
	Totals		27	12	6	45	160	30

Test Blueprint

The NYSTP Mathematics Tests assess students on the content and process strands of New York State Mathematics Learning Standard 3. The test items are indicators used to assess a variety of mathematics skills and abilities. Each item is aligned with one content-performance indicator for reporting purposes but is also aligned to one or more process-performance indicators, as appropriate for the concepts embodied in the task. As a result of the alignment to both process and content strands, the tests assess students' conceptual understanding, procedural fluency, and problem-solving abilities, rather than solely assessing their knowledge of isolated skills and facts. The five content strands, to which the items are aligned for reporting purposes, are Number Sense and Operations, Algebra, Geometry, Measurement, and Statistics and Probability. The distribution of score points across the strands 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 strand at that grade and the emphasis placed on those performance indicators by the blueprint-specifications panel members. Table 2 shows the Grades 3–8 Mathematics Test blueprint and actual number of score points in 2009 OP tests.

Table 2. NYSTP Mathematics 2009 Test Blueprint

Grade	Total Points	Content Strand	Target Points	Selected Points	Target % of Test	Selected % of Test
3	39	Number Sense and Operations	19	16	48.0	41.0
		Algebra	5	6	13.0	15.0
		Geometry	5	5	13.0	13.0
		Measurement	5	5	13.0	13.0
		Statistics and Probability	5	7	13.0	18.0

(Continued on next page)

Table 2. NYSTP Mathematics 2009 Test Blueprint (cont.)

Grade	Total Points	Content Strand	Target Points	Selected Points	Target % of Test	Selected % of Test
4	70	Number Sense and Operations	32	33	45.0	47.0
		Algebra	10	11	14.0	16.0
		Geometry	8	9	12.0	13.0
		Measurement	12	10	17.0	14.0
		Statistics and Probability	8	7	12.0	10.0
5	46	Number Sense and Operations	18	15	39.0	32.5
		Algebra	5	5	11.0	11.0
		Geometry	12	15	25.0	32.5
		Measurement	6	5	14.0	11.0
		Statistics and Probability	5	6	11.0	13.0
6	49	Number Sense and Operations	18	18	37.0	37.0
		Algebra	9	10	19.0	21.0
		Geometry	8	7	16.5	14.0
		Measurement	6	5	11.0	10.0
		Statistics and Probability	8	9	16.5	18.0
7	50	Number Sense and Operations	15	13	30.0	26.0
		Algebra	6	8	12.0	16.0
		Geometry	7	7	14.0	14.0
		Measurement	7	9	14.0	18.0
		Statistics and Probability	15	13	30.0	26.0
8	69	Number Sense and Operations	8	7	11.0	10.0
		Algebra	30	28	44.0	41.0
		Geometry	24	29	35.0	42.0
		Measurement	7	5	10.0	7.0

Tables 3a–3f present Grades 3–8 Mathematics Test item maps with the item type indicator, the answer key, the maximum number of points obtainable from each item, the current strand, and the performance indicator.

Table 3a. NYSTP Mathematics 2009 Operational Test Map, Grade 3

Question	Type	Points	Strand	Content Performance Indicator	Answer Key
Book 1					
1	Multiple Choice	1	Number Sense and Operations	3.N07 Use 1 as the identity element for multiplication	B
2	Multiple Choice	1	Measurement	3.M07 Count and represent combined coins and dollars, using currency symbols ($0.00)	C
3	Multiple Choice	1	Number Sense and Operations	3.N03 Compare and order numbers to 1,000	B
4	Multiple Choice	1	Number Sense and Operations	3.N04 Understand the place value structure of the base ten number system: 10 ones = 1 ten 10 tens = 1 hundred 10 hundreds = 1 thousand	B
5	Multiple Choice	1	Geometry	3.G01 Define and use correct terminology when referring to shapes (circle, triangle, square, rectangle, rhombus, trapezoid, and hexagon)	A
6	Multiple Choice	1	Number Sense and Operations	3.N18 Use a variety of strategies to add and subtract 3-digit numbers (with and without regrouping)	D
7	Multiple Choice	1	Number Sense and Operations	3.N21 Use the area model, tables, patterns, arrays, and doubling to provide meaning for multiplication	A
8	Multiple Choice	1	Number Sense and Operations	3.N13 Recognize fractional numbers as equal parts of a whole	B
9	Multiple Choice	1	Algebra	3.A02 Describe and extend numeric (+, -) and geometric patterns	B
10	Multiple Choice	1	Measurement	3.M07 Count and represent combined coins and dollars, using currency symbols ($0.00)	D
11	Multiple Choice	1	Number Sense and Operations	3.N06 Use and explain the commutative property of addition and multiplication	A
12	Multiple Choice	1	Number Sense and Operations	3.N10 Develop an understanding of fractions as part of a whole unit and as parts of a collection	C
13	Multiple Choice	1	Algebra	3.A01 Use the symbols <, >, and = (with and without the use of a number line) to compare whole numbers and unit fractions (1/2,1/3,1/4,1/5,1/6, and 1/10)	D

(Continued on next page)

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

Question	Type	Points	Strand	Content Performance Indicator	Answer Key
Book 1 (continued)					
14	Multiple Choice	1	Measurement	3.M02 Use a ruler/yardstick to measure to the nearest standard unit (whole and 1/2 inches, whole feet, and whole yards)	B
15	Multiple Choice	1	Geometry	3.G04 Identify the faces on a three-dimensional shape as two-dimensional shapes	D
16	Multiple Choice	1	Number Sense and Operations	3.N25 Estimate numbers up to 500	C
17	Multiple Choice	1	Measurement	3.M09 Tell time to the minute, using digital and analog clocks	D
18	Multiple Choice	1	Geometry	3.G05 Identify and construct lines of symmetry	D
19	Multiple Choice	1	Measurement	3.M01 Select tools and units (customary) appropriate for the length measured	A
20	Multiple Choice	1	Statistics and Probability	3.S07 Read and interpret data in bar graphs and pictographs	C
21	Multiple Choice	1	Number Sense and Operations	3.N08 Use the zero property of multiplication	A
22	Multiple Choice	1	Algebra	3.A02 Describe and extend numeric (+, -) and geometric patterns	D
23	Multiple Choice	1	Number Sense and Operations	3.N16 Identify odd and even numbers	D
24	Multiple Choice	1	Number Sense and Operations	3.N19 Develop fluency with single-digit multiplication facts	A
25	Multiple Choice	1	Statistics and Probability	3.S08 Formulate conclusions and make predictions from graphs	B
Book 2					
26	Short Response	2	Number Sense and Operations	3.N19 Develop fluency with single-digit multiplication facts	n/a
27	Short Response	2	Number Sense and Operations	3.N18 Use a variety of strategies to add and subtract 3-digit numbers (with and without regrouping)	n/a
28	Short Response	2	Geometry	3.G01 Define and use correct terminology when referring to shapes (circle, triangle, square, rectangle, rhombus, trapezoid, and hexagon)	n/a
29	Short Response	2	Statistics and Probability	3.S07 Read and interpret data in bar graphs and pictographs	n/a
30	Extended Response	3	Algebra	3.A02 Describe and extend numeric (+, -) and geometric patterns	n/a
31	Extended Response	3	Statistics and Probability	3.S05 Display data in pictographs and bar graphs	n/a

Table 3b. NYSTP Mathematics 2009 Operational Test Map, Grade 4

Question	Type	Points	Strand	Content Performance Indicator	Answer Key
Book 1					
1	Multiple Choice	1	Number Sense and Operations	4.N03 Compare and order numbers to 10,000	C
2	Multiple Choice	1	Number Sense and Operations	3.N19 Develop fluency with single-digit multiplication facts	B
3	Multiple Choice	1	Number Sense and Operations	4.N04 Understand the place value structure of the base ten number system: 10 ones = 1 ten 10 tens = 1 hundred 10 hundreds = 1 thousand 10 thousands = 1 ten thousand	C
4	Multiple Choice	1	Geometry	4.G03 Find perimeter of polygons by adding sides	D
5	Multiple Choice	1	Number Sense and Operations	4.N14 Use a variety of strategies to add and subtract numbers up to 10,000	D
6	Multiple Choice	1	Geometry	4.G02 Identify points and line segments when drawing a plane figure	C
7	Multiple Choice	1	Algebra	4.A03 Find the value or values that will make an open sentence true, if it contains < or >	A
8	Multiple Choice	1	Number Sense and Operations	4.N06 Understand, use, and explain the associative property of multiplication	C
9	Multiple Choice	1	Number Sense and Operations	4.N18 Use a variety of strategies to multiply two-digit numbers by one-digit numbers (with and without regrouping)	D
10	Multiple Choice	1	Number Sense and Operations	4.N21 Use a variety of strategies to divide two-digit dividends by one-digit divisors (with and without remainders)	C
11	Multiple Choice	1	Measurement	4.M02 Use a ruler to measure to the nearest standard unit (whole, 1/2 and 1/4 inches, whole feet, whole yards, whole centimeters, and whole meters)	B
12	Multiple Choice	1	Number Sense and Operations	4.N02 Read and write whole numbers to 10,000	D
13	Multiple Choice	1	Number Sense and Operations	4.N20 Develop fluency in multiplying and dividing multiples of 10 and 100 up to 1,000	B
14	Multiple Choice	1	Number Sense and Operations	3.N25 Estimate numbers up to 500	C
15	Multiple Choice	1	Statistics and Probability	4.S06 Formulate conclusions and make predictions from graphs	C
16	Multiple Choice	1	Geometry	3.G02 Identify congruent and similar figures	B
17	Multiple Choice	1	Number Sense and Operations	4.N17 Use multiplication and division as inverse operations to solve problems	B

(Continued on next page)

Table 3b. NYSTP Mathematics 2009 Operational Test Map, Grade 4 (cont.)

Question	Type	Points	Strand	Content Performance Indicator	Answer Key
Book 1 (continued)					
18	Multiple Choice	1	Measurement	4.M10 Calculate elapsed time in days and weeks, using a calendar	D
19	Multiple Choice	1	Algebra	4.A05 Analyze a pattern or a whole-number function and state the rule, given a table or an input/output box	B
20	Multiple Choice	1	Number Sense and Operations	3.N14 Explore equivalent fractions (1/2, 1/3, 1/4)	B
21	Multiple Choice	1	Algebra	4.A02 Use the symbols <, >, =, and ≠ (with and without the use of a number line) to compare whole numbers and unit fractions and decimals (up to hundredths)	A
22	Multiple Choice	1	Geometry	4.G01 Identify and name polygons, recognizing that their names are related to the number of sides and angles (triangle, quadrilateral, pentagon, hexagon, and octagon)	A
23	Multiple Choice	1	Statistics and Probability	4.S06 Formulate conclusions and make predictions from graphs	D
24	Multiple Choice	1	Algebra	4.A04 Describe, extend, and make generalizations about numeric (+,-,×,÷) and geometric patterns	C
25	Multiple Choice	1	Number Sense and Operations	4.N27 Check reasonableness of an answer by using estimation	A
26	Multiple Choice	1	Number Sense and Operations	4.N26 Round numbers less than 1,000 to the nearest tens and hundreds	B
27	Multiple Choice	1	Number Sense and Operations	4.N15 Select appropriate computational and operational methods to solve problems	D
28	Multiple Choice	1	Measurement	4.M09 Calculate elapsed time in hours and half hours, not crossing A.M./P.M.	C
29	Multiple Choice	1	Number Sense and Operations	4.N22 Interpret the meaning of remainders	B
30	Multiple Choice	1	Measurement	4.M06 Select tools and units appropriate to the capacity being measured (milliliters and liters)	A
Book 2					
31	Short Response	2	Number Sense and Operations	4.N22 Interpret the meaning of remainders	n/a
32	Short Response	2	Number Sense and Operations	4.N20 Develop fluency in multiplying and dividing multiples of 10 and 100 up to 1,000	n/a
33	Short Response	2	Number Sense and Operations	4.N16 Understand various meanings of multiplication and division	n/a

(Continued on next page)

Table 3b. NYSTP Mathematics 2009 Operational Test Map, Grade 4 (cont.)

Question	Type	Points	Strand	Content Performance Indicator	Answer Key
Book 2 (continued)					
34	Short Response	2	Algebra	4.A01 Evaluate and express relationships using open sentences with one operation	n/a
35	Short Response	2	Number Sense and Operations	4.N14 Use a variety of strategies to add and subtract numbers up to 10,000	n/a
36	Short Response	2	Geometry	4.G01 Identify and name polygons, recognizing that their names are related to the number of sides and angles (triangle, quadrilateral, pentagon, hexagon, and octagon)	n/a
37	Short Response	2	Measurement	4.M08 Make change, using combined coins and dollar amounts	n/a
38	Extended Response	3	Statistics and Probability	4.S05 Develop and make predictions that are based on data	n/a
39	Extended Response	3	Algebra	4.A05 Analyze a pattern or a whole-number function and state the rule, given a table or an input/output box	n/a
Book 3					
40	Short Response	2	Number Sense and Operations	4.N14 Use a variety of strategies to add and subtract numbers up to 10,000	n/a
41	Short Response	2	Measurement	4.M08 Make change, using combined coins and dollar amounts	n/a
42	Short Response	2	Number Sense and Operations	4.N18 Use a variety of strategies to multiply two-digit numbers by one-digit numbers (with and without regrouping)	n/a
43	Short Response	2	Algebra	4.A02 Use the symbols $<$, $>$, $=$, and $\neq$ (with and without the use of a number line) to compare whole numbers and unit fractions and decimals (up to hundredths)	n/a
44	Short Response	2	Measurement	4.M03 Know and understand equivalent standard units of length: 12 inches = 1 foot 3 feet = 1 yard	n/a
45	Short Response	2	Number Sense and Operations	4.N17 Use multiplication and division as inverse operations to solve problems	n/a
46	Short Response	2	Statistics and Probability	4.S03 Represent data using tables, bar graphs, and pictographs	n/a
47	Extended Response	3	Geometry	4.G03 Find perimeter of polygons by adding sides 4.G04 Find the area of a rectangle by counting the number of squares needed to cover the rectangle	n/a
48	Extended Response	3	Number Sense and Operations	3.N20 Use a variety of strategies to solve multiplication problems with factors up to 12 x 12	n/a

Table 3c. NYSTP Mathematics 2009 Operational Test Map, Grade 5

Question	Type	Points	Strand	Content Performance Indicator	Answer Key
Book 1					
1	Multiple Choice	1	Measurement	5.M01 Use a ruler to measure to the nearest inch, 1/2, 1/4, and 1/8 inch	D
2	Multiple Choice	1	Number Sense and Operations	5.N01 Read and write whole numbers to millions	C
3	Multiple Choice	1	Geometry	5.G01 Calculate the perimeter of regular and irregular polygons	D
4	Multiple Choice	1	Algebra	4.A02 Use the symbols <, >, =, and ≠ (with and without the use of a number line) to compare whole numbers and unit fractions and decimals (up to hundredths)	A
5	Multiple Choice	1	Number Sense and Operations	5.N08 Read, write, and order decimals to thousandths	C
6	Multiple Choice	1	Geometry	5.G11 Identify and draw lines of symmetry of basic geometric shapes	D
7	Multiple Choice	1	Algebra	4.A02 Use the symbols <, >, =, and ≠ (with and without the use of a number line) to compare whole numbers and unit fractions and decimals (up to hundredths)	D
8	Multiple Choice	1	Statistics and Probability	4.S04 Read and interpret line graphs	B
9	Multiple Choice	1	Number Sense and Operations	5.N01 Read and write whole numbers to millions	D
10	Multiple Choice	1	Geometry	5.G04 Classify quadrilaterals by properties of their angles and sides	C
11	Multiple Choice	1	Number Sense and Operations	4.N10 Develop an understanding of decimals as part of a whole	A
12	Multiple Choice	1	Number Sense and Operations	5.N16 Use a variety of strategies to multiply three-digit by three-digit numbers	D
13	Multiple Choice	1	Geometry	5.G06 Classify triangles by properties of their angles and sides	A
14	Multiple Choice	1	Algebra	5.A06 Evaluate the perimeter formula for given input values	D
15	Multiple Choice	1	Geometry	5.G03 Identify the ratio of corresponding sides of similar triangles	A
16	Multiple Choice	1	Measurement	5.M05 Convert measurement within a given system	D
17	Multiple Choice	1	Statistics and Probability	4.S04 Read and interpret line graphs	D
18	Multiple Choice	1	Number Sense and Operations	5.N05 Compare and order fractions including unlike denominators (with and without the use of a number line) Note: Commonly used fractions such as those that might be indicated on ruler, measuring cup, etc.	A

(Continued on next page)

Table 3c. NYSTP Mathematics 2009 Operational Test Map, Grade 5 (cont.)

Question	Type	Points	Strand	Content Performance Indicator	Answer Key
Book 1 (continued)					
19	Multiple Choice	1	Number Sense and Operations	5.N23 Use a variety of strategies to add, subtract, multiply, and divide decimals to thousandths	C
20	Multiple Choice	1	Geometry	5.G05 Know that the sum of the interior angles of a quadrilateral is 360 degrees	A
21	Multiple Choice	1	Measurement	5.M07 Calculate elapsed time in hours and minutes	D
22	Multiple Choice	1	Number Sense and Operations	4.N23 Add and subtract proper fractions with common denominators	A
23	Multiple Choice	1	Number Sense and Operations	5.N17 Use a variety of strategies to divide three-digit numbers by one- and two-digit numbers	C
24	Multiple Choice	1	Geometry	5.G09 Identify pairs of congruent triangles	D
25	Multiple Choice	1	Number Sense and Operations	5.N13 Calculate multiples of a whole number and the least common multiple of two numbers	B
26	Multiple Choice	1	Statistics and Probability	5.S03 Calculate the mean for a given set of data and use to describe a set of data	A
Book 2					
27	Short Response	2	Geometry	5.G08 Find a missing angle when given two angles of a triangle	n/a
28	Short Response	2	Number Sense and Operations	5.N23 Use a variety of strategies to add, subtract, multiply, and divide decimals to thousandths	n/a
29	Short Response	2	Measurement	5.M08 Measure and draw angles using a protractor	n/a
30	Short Response	2	Algebra	5.A08 Create algebraic or geometric patterns using concrete objects or visual drawings (e.g., rotate and shade geometric shapes)	n/a
31	Extended Response	3	Geometry	5.G04 Classify quadrilaterals by properties of their angles and sides	n/a
32	Extended Response	3	Number Sense and Operations	5.N08 Read, write, and order decimals to thousandths	n/a
33	Extended Response	3	Statistics and Probability	5.S04 Formulate conclusions and make predictions from graphs	n/a
34	Extended Response	3	Geometry	5.G01 Calculate the perimeter of regular and irregular polygons	n/a

Table 3d. NYSTP Mathematics 2009 Operational Test Map, Grade 6

Question	Type	Points	Strand	Content Performance Indicator	Answer Key
Book 1					
1	Multiple Choice	1	Measurement	6.M03 Identify equivalent customary units of capacity (cups to pints, pints to quarts, and quarts to gallons)	C
2	Multiple Choice	1	Statistics and Probability	6.S06 Determine the range for a given set of data	A
3	Multiple Choice	1	Geometry	5.G13 Plot points to form basic geometric shapes (identify and classify)	C
4	Multiple Choice	1	Number Sense and Operations	6.N24 Represent exponential form as repeated multiplication	D
5	Multiple Choice	1	Algebra	6.A06 Evaluate formulas for given input values (circumference, area, volume, distance, temperature, interest, etc.)	C
6	Multiple Choice	1	Algebra	5.A02 Translate simple verbal expressions into algebraic expressions	C
7	Multiple Choice	1	Number Sense and Operations	6.N16 Add and subtract fractions with unlike denominators	A
8	Multiple Choice	1	Statistics and Probability	6.S05 Determine the mean, mode and median for a given set of data	B
9	Multiple Choice	1	Number Sense and Operations	6.N22 Evaluate numerical expressions using order of operations (may include exponents of two and three)	B
10	Multiple Choice	1	Geometry	6.G01 Calculate the length of corresponding sides of similar triangles, using proportional reasoning	C
11	Multiple Choice	1	Algebra	5.A04 Solve simple one-step equations using basic whole-number facts	D
12	Multiple Choice	1	Algebra	5.A03 Substitute assigned values into variable expressions and evaluate using order of operations	B
13	Multiple Choice	1	Number Sense and Operations	6.N15 Order rational numbers (including positive and negative)	B
14	Multiple Choice	1	Algebra	5.A02 Translate simple verbal expressions into algebraic expressions	C
15	Multiple Choice	1	Number Sense and Operations	6.N18 Add, subtract, multiply, and divide mixed numbers with unlike denominators	D

(Continued on next page)

Table 3d. NYSTP Mathematics 2009 Operational Test Map, Grade 6 (cont.)

Question	Type	Points	Strand	Content Performance Indicator	Answer Key
Book 1 (continued)					
16	Multiple Choice	1	Geometry	5.G14 Calculate perimeter of basic geometric shapes drawn on a coordinate plane (rectangles and shapes composed of rectangles having sides with integer lengths and parallel to the axes)	B
17	Multiple Choice	1	Geometry	6.G06 Understand the relationship between the diameter and radius of a circle	D
18	Multiple Choice	1	Measurement	6.M05 Identify equivalent metric units of capacity (milliliter to liter and liter to milliliter)	D
19	Multiple Choice	1	Number Sense and Operations	6.N13 Define absolute value and determine the absolute value of rational numbers (including positive and negative)	A
20	Multiple Choice	1	Statistics and Probability	6.S07 Read and interpret graphs	D
21	Multiple Choice	1	Number Sense and Operations	6.N11 Read, write, and identify percents of a whole (0% to 100%)	C
22	Multiple Choice	1	Statistics and Probability	5.S07 Create a sample space and determine the probability of a single event, given a simple experiment (e.g., rolling a number cube)	D
23	Multiple Choice	1	Geometry	6.G04 Determine the volume of rectangular prisms by counting cubes and develop the formula	D
24	Multiple Choice	1	Number Sense and Operations	6.N14 Locate rational numbers on a number line (including positive and negative)	B
25	Multiple Choice	1	Measurement	6.M03 Identify equivalent customary units of capacity (cups to pints, pints to quarts, and quarts to gallons)	A
Book 2					
26	Short Response	2	Number Sense and Operations	6.N23 Represent repeated multiplication in exponential form	n/a
27	Short Response	2	Geometry	6.G05 Identify radius, diameter, chords and central angles of a circle	n/a
28	Short Response	2	Number Sense and Operations	6.N21 Find multiple representations of rational numbers (fractions, decimals, and percents 0 to 100)	n/a
29	Short Response	2	Algebra	6.A02 Use substitution to evaluate algebraic expressions (may include exponents of one, two, and three)	n/a

Continued on next page)

Table 3d. NYSTP Mathematics 2009 Operational Test Map, Grade 6 (cont.)

Question	Type	Points	Strand	Content Performance Indicator	Answer Key
Book 2 (continued)					
30	Short Response	2	Statistics and Probability	5.S07 Create a sample space and determine the probability of a single event, given a simple experiment (e.g., rolling a number cube)	n/a
31	Short Response	2	Measurement	6.M01 Measure capacity and calculate volume of a rectangular prism	n/a
32	Extended Response	3	Number Sense and Operations	6.N02 Define and identify the commutative and associative properties of addition and multiplication	n/a
33	Extended Response	3	Number Sense and Operations	6.N12 Solve percent problems involving percent, rate, and base	n/a
34	Extended Response	3	Statistics and Probability	6.S08 Justify predictions made from data	n/a
35	Extended Response	3	Algebra	5.A05 Solve and explain simple one-step equations using inverse operations involving whole numbers	n/a

Table 3e. NYSTP Mathematics 2009 Operational Test Map, Grade 7

Question	Type	Points	Strand	Content Performance Indicator	Answer Key
Book 1					
1	Multiple Choice	1	Number Sense and Operations	7.N01 Distinguish between the various subsets of real numbers (counting/natural numbers, whole numbers, integers, rational numbers, and irrational numbers)	A
2	Multiple Choice	1	Number Sense and Operations	7.N07 Compare numbers written in scientific notation	C
3	Multiple Choice	1	Algebra	7.A06 Evaluate formulas for given input values (surface area, rate, and density problems)	C
4	Multiple Choice	1	Statistics and Probability	7.S06 Read and interpret data represented graphically (pictograph, bar graph, histogram, line graph, double line/bar graphs or circle graph)	D
5	Multiple Choice	1	Measurement	7.M04 Convert mass within a given system	C

Continued on next page)

Table 3e. NYSTP Mathematics 2009 Operational Test Map, Grade 7 (cont.)

Question	Type	Points	Strand	Content Performance Indicator	Answer Key
Book 1 (continued)					
6	Multiple Choice	1	Geometry	7.G03 Identify the two-dimensional shapes that make up the faces and bases of three-dimensional shapes (prisms, cylinders, cones, and pyramids)	C
7	Multiple Choice	1	Statistics and Probability	7.S09 Determine the validity of sampling methods to predict outcomes	D
8	Multiple Choice	1	Number Sense and Operations	7.N06 Translate numbers from scientific notation into standard form	B
9	Multiple Choice	1	Measurement	7.M09 Determine the tool and technique to measure with an appropriate level of precision: mass	A
10	Multiple Choice	1	Number Sense and Operations	7.N12 Add, subtract, multiply, and divide integers	D
11	Multiple Choice	1	Measurement	7.M02 Convert capacities and volumes within a given system	C
12	Multiple Choice	1	Statistics and Probability	7.S08 Interpret data to provide the basis for predictions and to establish experimental probabilities	C
13	Multiple Choice	1	Algebra	7.A06 Evaluate formulas for given input values (surface area, rate, and density problems)	B
14	Multiple Choice	1	Number Sense and Operations	7.N11 Simplify expressions using order of operations (Note: Expressions may include absolute value and/or integral exponents greater than 0)	D
15	Multiple Choice	1	Statistics and Probability	7.S06 Read and interpret data represented graphically (pictograph, bar graph, histogram, line graph, double line/bar graphs or circle graph)	C
16	Multiple Choice	1	Number Sense and Operations	7.N15 Recognize and state the value of the square root of a perfect square (up to 225)	B
17	Multiple Choice	1	Geometry	6.G11 Calculate the area of basic polygons drawn on a coordinate plane (rectangles and shapes composed of rectangles having sides with integer lengths)	B
18	Multiple Choice	1	Statistics and Probability	7.S04 Calculate the range for a given set of data	C
19	Multiple Choice	1	Measurement	7.M11 Estimate surface area	B
20	Multiple Choice	1	Number Sense and Operations	7.N08 Find the common factors and greatest common factor of two or more numbers	C

Continued on next page)

Table 3e. NYSTP Mathematics 2009 Operational Test Map, Grade 7 (cont.)

Question	Type	Points	Strand	Content Performance Indicator	Answer Key
Book 1 (continued)					
21	Multiple Choice	1	Statistics and Probability	7.S12 Compare actual results to predicted results	D
22	Multiple Choice	1	Algebra	7.A01 Translate two-step verbal expressions into algebraic expressions	A
23	Multiple Choice	1	Number Sense and Operations	7.N12 Add, subtract, multiply, and divide integers	B
24	Multiple Choice	1	Measurement	7.M02 Convert capacities and volumes within a given system	D
25	Multiple Choice	1	Statistics and Probability	7.S12 Compare actual results to predicted results	A
26	Multiple Choice	1	Geometry	7.G01 Calculate the radius or diameter, given the circumference or area of a circle	C
27	Multiple Choice	1	Statistics and Probability	6.S11 Determine the number of possible outcomes for a compound event by using the fundamental counting principle and use this to determine the probabilities of events when the outcomes have equal probability	C
28	Multiple Choice	1	Measurement	7.M03 Identify customary and metric units of mass	D
29	Multiple Choice	1	Algebra	6.A03 Translate two-step verbal sentences into algebraic equations	B
30	Multiple Choice	1	Geometry	7.G03 Identify the two-dimensional shapes that make up the faces and bases of three-dimensional shapes (prisms, cylinders, cones, and pyramids)	B
Book 2					
31	Short Response	2	Statistics and Probability	6.S09 List possible outcomes for compound events	n/a
32	Short Response	2	Algebra	6.A05 Solve simple proportions within context	n/a
33	Short Response	2	Algebra	6.A02 Use substitution to evaluate algebraic expressions (may include exponents of one, two and three)	n/a
34	Short Response	2	Number Sense and Operations	7.N05 Write numbers in scientific notation	n/a
35	Extended Response	3	Number Sense and Operations	7.N13 Add and subtract two integers (with and without the use of a number line)	n/a
36	Extended Response	3	Measurement	7.M02 Convert capacities and volumes within a given system	n/a
37	Extended Response	3	Geometry	7.G07 Find a missing angle when given angles of a quadrilateral	n/a
38	Extended Response	3	Statistics and Probability	6.S02 Record data in a frequency table	n/a

Table 3f. NYSTP Mathematics 2009 Operational Test Map, Grade 8

Question	Type	Points	Strand	Content Performance Indicator	Answer Key
Book 1					
1	Multiple Choice	1	Algebra	7.A02 Add and subtract monomials with exponents of one	C
2	Multiple Choice	1	Geometry	8.G05 Calculate the missing angle measurements when given two parallel lines cut by a transversal	B
3	Multiple Choice	1	Geometry	7.G05 Identify the right angle, hypotenuse, and legs of a right triangle	B
4	Multiple Choice	1	Number Sense and Operations	8.N01 Develop and apply the laws of exponents for multiplication and division	C
5	Multiple Choice	1	Geometry	8.G05 Calculate the missing angle measurements when given two parallel lines cut by a transversal	B
6	Multiple Choice	1	Algebra	7.A04 Solve multi-step equations by combining like terms, using the distributive property, or moving variables to one side of the equation	C
7	Multiple Choice	1	Algebra	8.A03 Describe a situation involving relationships that matches a given graph	D
8	Multiple Choice	1	Geometry	7.G08 Use the Pythagorean Theorem to determine the unknown length of a side of a right triangle	D
9	Multiple Choice	1	Geometry	8.G04 Determine angle pair relationships when given two parallel lines cut by a transversal	C
10	Multiple Choice	1	Algebra	8.A08 Multiply a binomial by a monomial or binomial (integer coefficients)	D
11	Multiple Choice	1	Algebra	8.A03 Describe a situation involving relationships that matches a given graph	C
12	Multiple Choice	1	Number Sense and Operations	8.N05 Estimate a percent of quantity, given an application	B
13	Multiple Choice	1	Geometry	8.G04 Determine angle pair relationships when given two parallel lines cut by a transversal	C
14	Multiple Choice	1	Algebra	8.A02 Write verbal expressions that match given mathematical expressions	B
15	Multiple Choice	1	Geometry	8.G01 Identify pairs of vertical angles as congruent	A
16	Multiple Choice	1	Algebra	8.A09 Divide a polynomial by a monomial (integer coefficients)	B
17	Multiple Choice	1	Geometry	8.G06 Calculate the missing angle measurements when given two intersecting lines and an angle	A

Continued on next page)

Table 3f. NYSTP Mathematics 2009 Operational Test Map, Grade 8 (cont.)

Question	Type	Points	Strand	Content Performance Indicator	Answer Key
Book 1 (continued)					
18	Multiple Choice	1	Measurement	7.M01 Calculate distance using a map scale	D
19	Multiple Choice	1	Algebra	8.A04 Create a graph given a description or an expression for a situation involving a linear or nonlinear relationship	D
20	Multiple Choice	1	Geometry	8.G06 Calculate the missing angle measurements when given two intersecting lines and an angle	C
21	Multiple Choice	1	Algebra	8.A09 Divide a polynomial by a monomial (integer coefficients)	A
22	Multiple Choice	1	Measurement	7.M01 Calculate distance using a map scale	C
23	Multiple Choice	1	Geometry	8.G05 Calculate the missing angle measurements when given two parallel lines cut by a transversal	D
24	Multiple Choice	1	Geometry	8.G03 Calculate the missing angle in a supplementary or complementary pair	C
25	Multiple Choice	1	Measurement	7.M01 Calculate distance using a map scale	C
26	Multiple Choice	1	Geometry	7.G05 Identify the right angle, hypotenuse, and legs of a right triangle	A
27	Multiple Choice	1	Geometry	7.G08 Use the Pythagorean Theorem to determine the unknown length of a side of a right triangle	B
Book 2					
28	Short Response	2	Geometry	8.G06 Calculate the missing angle measurements when given two intersecting lines and an angle	n/a
29	Short Response	2	Geometry	8.G05 Calculate the missing angle measurements when given two parallel lines cut by a transversal	n/a
30	Short Response	2	Algebra	7.A10 Write an equation to represent a function from a table of values	n/a
31	Short Response	2	Number Sense and Operations	8.N04 Apply percents to: tax, percent increase/decrease, simple interest, sale price, commission, interest rates, and gratuities	n/a
32	Extended Response	3	Algebra	7.A04 Solve multi-step equations by combining like terms, using the distributive property, or moving variables to one side of the equation	n/a
33	Extended Response	3	Geometry	8.G09 Draw the image of a figure under a reflection over a given line	n/a

Continued on next page)

Table 3f. NYSTP Mathematics 2009 Operational Test Map, Grade 8 (cont.)

Question	Type	Points	Strand	Content Performance Indicator	Answer Key
Book 3					
34	Short Response	2	Geometry	8.G05 Calculate the missing angle measurements when given two parallel lines cut by a transversal	n/a
35	Short Response	2	Algebra	7.A08 Create algebraic patterns using charts/tables, graphs, equations, and expressions	n/a
36	Short Response	2	Algebra	8.A06 Multiply and divide monomials	n/a
37	Short Response	2	Measurement	7.M06 Compare unit prices	n/a
38	Short Response	2	Algebra	8.A07 Add and subtract polynomials (integer coefficients)	n/a
39	Short Response	2	Geometry	8.G05 Calculate the missing angle measurements when given two parallel lines cut by a transversal	n/a
40	Short Response	2	Algebra	8.A03 Describe a situation involving relationships that matches a given graph	n/a
41	Short Response	2	Geometry	8.G01 Identify pairs of vertical angles as congruent	n/a
42	Extended Response	3	Geometry	8.G10 Draw the image of a figure under a translation	n/a
43	Extended Response	3	Algebra	8.A04 Create a graph given a description or an expression for a situation involving a linear or nonlinear relationship	n/a
44	Extended Response	3	Number Sense and Operations	8.N01 Develop and apply the laws of exponents for multiplication and division	n/a
45	Extended Response	3	Algebra	7.A04 Solve multi-step equations by combining like terms, using the distributive property, or moving variables to one side of the equation	n/a

Table 4. NYSTP Mathematics 2009 Strand Coverage

Grade	Strand	MC Item #	SR Item #	ER Item #	Total Items
3	Number Sense and Operations	1, 3, 4, 6, 7, 8, 11, 12, 16, 21, 23, 24	26, 27	n/a	14
	Algebra	9, 13, 22	n/a	30	4
	Geometry	5, 15, 18	28	n/a	4
	Measurement	2, 10, 14, 17, 19	n/a	n/a	5
	Statistics and Probability	20, 25	29	31	4
4	Number Sense and Operations	1, 2, 3, 5, 8, 9, 10, 12, 13, 14, 17, 20, 25, 26, 27, 29	31, 32, 33, 35, 40, 42, 45	48	24
	Algebra	7, 19, 21, 24	34, 43	39	7
	Geometry	4, 6, 16, 22	36	47	6
	Measurement	11, 18, 28, 30	37, 41, 44	n/a	7
	Statistics and Probability	15, 23	46	38	4
5	Number Sense and Operations	2, 5, 9, 11, 12, 18, 19, 22, 23, 25	28	32	12
	Algebra	4, 7, 14	30	n/a	4
	Geometry	3, 6, 10, 13, 15, 20, 24	27	31, 34	10
	Measurement	1, 16, 21	29	n/a	4
	Statistics and Probability	8, 17, 26	n/a	33	4
6	Number Sense and Operations	4, 7, 9, 13, 15, 19, 21, 24	26, 28	32, 33	12
	Algebra	5, 6, 11, 12, 14	29	35	7
	Geometry	3, 10, 16, 17, 23	27	n/a	6
	Measurement	1, 18, 25	31	n/a	4
	Statistics and Probability	2, 8, 20, 22	30	34	6
7	Number Sense and Operations	1, 2, 8, 10, 14, 16, 20, 23	34	35	10
	Algebra	3, 13, 22, 29	32, 33	n/a	6
	Geometry	6, 17, 26, 30	n/a	37	5
	Measurement	5, 9, 11, 19, 24, 28	n/a	36	7
	Statistics and Probability	4, 7, 12, 15, 18, 21, 25, 27	31	38	10

(Continued on next page)

Table 4. NYSTP Mathematics 2009 Strand Coverage (cont.)

Grade	Strand	MC Item #	SR Item #	ER Item #	Total Items
8	Number Sense and Operations	4, 12	31	44	4
	Algebra	1, 6, 7, 10, 11, 14, 16, 19, 21	30, 35, 36, 38, 40	32, 43, 45	17
	Geometry	2, 3, 5, 8, 9, 13, 15, 17, 20, 23, 24, 26, 27	28, 29, 34, 39, 41	33, 42	20
	Measurement	18, 22, 25	37	n/a	4

New York State Educator's Involvement in Test Development

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

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

The selected participants must be certified and have both teaching and testing experience. The majority of participants are classroom teachers, but specialists such as reading coaches, literacy coaches, as well as special education and bilingual instructors participate. Some participants are also recommended by principals, the Staff and Curriculum Development Network (SCDN), professional organizations, Big Five Cities, etc. Other criteria are also considered, such as gender, ethnicity, geographic location, and type of school (urban, suburban, and rural). As recruitment forms are received a file of participants is maintained and is routinely updated with current participant information and the addition of possible future participants. 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 August 2004, CTB/McGraw-Hill facilitated specifications meetings in Albany, New York, during which committees of state educators, along with NYSED staff, reviewed the strands and performance indicators to make the following determinations:

- which performance indicators were to be assessed
- which item types were to be used for the assessable performance indicators (For example, some performance indicators lend themselves more easily to assessment by CR items than others.)

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

The committees were composed of teachers from around the state selected for their grade-level expertise, were grouped by grade band (i.e., 3/4, 5/6, 7/8), and met for four days. The committees were composed of approximately 10 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. In January 2005, a second specifications meeting was held again with New York State educators from around the state in order to review changes made to the New York State Mathematics Learning Standards and all the items were revisited before field testing to certify alignment.

Item Development

Based on the decisions made during the item specifications meetings, the content-lead editors at CTB/McGraw-Hill 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 assignment. Writers were familiarized with the New York State Testing Program and the test specifications. They were also provided with sample test items, a style guide, and a document outlining the criteria for acceptable items (see Appendix A) to help them in their writing process.

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 meetings, committees composed of New York State educators were selected for their content and grade-level expertise for item review. Each committee was composed of approximately 10 participants per grade band. The committee members were provided with the items, the New York State Learning Standards, and the test specifications, and 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 distracters (MC items)

- the conciseness, preciseness, clarity, and readability 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 items into field test (FT) forms and submitted the FT forms to NYSED for their review and approval. The FTs were administered to students across New York State during the week of March 19, 2008. In addition, CTB/McGraw-Hill, in conjunction with NYSED's input and approval, developed a combined *Teacher's Directions and School Administrator's Manual* so that the FTs were administered in a uniform manner to all participating students.

After administration of the FTs, rangefinding sessions were conducted in April 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 8 to 10 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 scoring 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 Grades 3–8 Mathematics OP Tests were administered in March 2009. The test items were selected from the pool of field-tested items, 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 B). Item selection for the NYSTP Grades 3–8 Mathematics Tests was based on the classical and 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. 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 B).

The NYSED staff (including content and research experts) traveled to CTB/McGraw-Hill in Monterey, CA, in August 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 approval by NYSED, the tests were produced and administered in March 2009.

In addition to the test books, CTB/McGraw-Hill 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/3-8mathsam-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 mathematics standard setting in Albany, July 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 proficiency cut score setting, please refer to the *Bookmark Standard Setting Technical Report 2006 for Grades 3, 4, 5, 6, 7, and 8 Mathematics* and *NYS Measurement Review Technical Report 2006 for Mathematics*.

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 mathematics as defined by the New York State Mathematics 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 analyses of test content indicate 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 2–4 in Section II). The test development process requires specific attention to content representation and the balance within each test form. New York State educators were involved in test constructions in various test development stages. For example, during the item review process, they reviewed FTs for their alignment with the test blueprint. Educators also participated in a process of establishing scoring rubrics (during rangefinding sessions) for CR items. Section II, "Test Design and Development," contains more information specific to the item review process. An independent study of alignment between the New York State curriculum and the New York State Grades 3–8 Mathematics 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 Mathematics Tests is supported by several types of evidence that can be obtained from the mathematics 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 populations, the reliability coefficients (Cronbach's alpha) ranged from 0.88–0.94, and for most subgroups, the reliability coefficients are greater than 0.80 (the exception was for Grade 5 students who took the Korean version). Overall, high internal consistency of the New York State Mathematics Tests provides 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 that 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 all items in Grades 3 and 7 Mathematics Tests displayed good item-model fit. Two items in Grade 4, two items in Grade 5, one item in Grade 6, and five items in Grade 8 were flagged for poor fit. The fact that only a few items were deemed to have unacceptable fit across grades of the Mathematics Tests 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 Mathematics Tests were related. What relates the questions is most parsimoniously claimed to be the common ability acquired by students studying the content area. 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 related to what the questions were designed to have in common, i.e., mathematics ability.

To demonstrate the common factor (ability) underlying student responses to mathematics 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 Mathematics 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, demonstrating essential unidimensionality of the trait measured by each test.

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 five 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 Mathematics Tests exhibited first principle components accounting for more than 20 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 5.

Table 5. Factor Analysis Results for Mathematics Tests (Total Population)

Grade	Initial Eigenvalues			
	Component	Total	% of Variance	Cumulative %
3	**1**	**7.60**	**24.50**	**24.50**
	2	1.29	4.16	28.66
	3	1.05	3.38	32.04
4	**1**	**12.63**	**26.31**	**26.31**
	2	1.38	2.87	29.18
	3	1.18	2.46	31.64
	4	1.01	2.11	33.75
5	**1**	**8.36**	**24.60**	**24.60**
	2	1.16	3.41	28.00
6	**1**	**9.43**	**26.95**	**26.95**
	2	1.29	3.68	30.63
	3	1.18	3.36	33.99
7	**1**	**8.76**	**23.05**	**23.05**
	2	1.56	4.11	27.16
	3	1.29	3.40	30.56
	4	1.06	2.79	33.35
	5	1.03	2.72	36.07
8	**1**	**13.81**	**30.68**	**30.68**
	2	1.77	3.94	34.62
	3	1.31	2.92	37.53
	4	1.15	2.56	40.09
	5	1.00	2.23	42.33

This evidence supports the claim that there is a construct ability underlying the items/tasks in each mathematics 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 mathematics 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 mathematics tests for the analyzed subgroups. Factor analysis results for ELL, SWD, SUA, ELL/SUA and SWD/SUA classifications are provided in Table C1 of Appendix C. 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, translation, and socioeconomic status (SES) bias. All materials were written and reviewed to conform to the 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 Mathematics 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 FT materials was reviewed by at least these same people.

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

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

In the fourth procedure, statistical methods were used to identify items exhibiting possible DIF. Although items flagged for DIF in the FT stage were closely examined for content bias and avoided during the OP test construction, DIF analyses were conducted again on OP test data. Three methods were employed to evaluate the amount of DIF in all test items: standardized mean difference, Mantel-Haenszel (see Section V, "Operational Test Data Collection and Classical Analysis"), and Linn-Harnisch (see Section VI, "IRT Scaling and Equating"). Although several items in each grade were flagged for DIF, typically the amount of DIF present was not large and 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 Scoring Site Operations Manual (2009) located at http://www.emsc.nysed.gov/3-8/archived.htm#scoring.

Test Administration

NYSTP Grades 3–8 Mathematics Tests were administered at the classroom level, during March 2009. The testing window for Grades 3, 4, and 5 was March 3–7, 2009. The testing window for Grades 6, 7, and 8 was March 6–12. The makeup test administration window was March 10–14 for Grades 3–5 and from March 13–19 for Grades 6–8. The makeup test administration window 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 oversight 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 enforcing the accuracy of scoring. The titles for administrators, trainers, and facilitators varied per scoring model chosen. At the regional level, oversight was conducted by a site coordinator. A scoring leader trained the scoring committee members and monitored sessions, and a table facilitator assisted in monitoring sessions. At the districtwide level, a school district administrator oversaw OP scoring. A district mathematics leader trained and monitored sessions, and a school mathematics leader assisted in monitoring 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 Mathematics 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 first readers 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 first readers 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 first readers 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 first readers 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 CR items was based primarily on the scoring guides, which were created by CTB/McGraw-Hill handscoring and content development specialists with guidance from NYSED and New York State teachers during rangefinding sessions. The CTB/McGraw-Hill mathematics 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 April 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. Handscoring and content-development specialists created scoring guides based on rangefinding decisions and conferences with NYSED. 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 Mathematics Frequently Asked Questions (FAQs) document, and a DVD, which 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 mathematics 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 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 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 allows 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 test responses. After training, each scoring committee member was deemed prepared and verified as ready to score the 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, mathematics FAQs, and DVD, they called the New York State Helpline. This call center was established to aid teachers and administrators during OP scoring. The helpline staff consisted of trained CTB/McGraw-Hill handscoring personnel who answered questions by phone, fax, or e-mail. 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' OP test 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 at the beginning of April 2009 for Grades 3, 4, and 5, and in the end of April 2009 for Grades 6, 7, and 8. These data were used for all data analysis. Phase 2 involved submitting of "straggler files" to CTB/McGraw-Hill in late-April 2009. The straggler files contained less than 2% of the total population cases and due to late submission were excluded from research data analyses. Nonpublic school data were excluded from all data analyses.

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. CTB/McGraw-Hill research performed data cleaning to the delivered data and excluded some student cases in order to obtain a sample of the utmost integrity. It should be noted that the two 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. In addition, Grade 8 students who were administered an incorrect version of the Grade 8 test were excluded from Grade 8 data files (refer to "Item Rescoring and Replacing" sub-section for details). A list of the data cleaning procedures conducted by research and accompanying case counts is presented in Tables 6a–6f.

Table 6a. NYSTP Mathematics Data Cleaning, Grade 3

Exclusion Rule	# Deleted	# Cases Remain
Initial N	0	200061
Out of grade	88	199973
No grade	1	199972
Duplicate record	0	199972
Non-public and out-of-district schools	2513	197459
Missing values for ALL items on OP form	1	197458
Out-of-range CR scores	0	197458
Remove Math Large Print	266	197192

Table 6b. NYSTP Mathematics Data Cleaning, Grade 4

Exclusion Rule	# Deleted	# Cases Remain
Initial N	0	205278
Out of grade	74	205204
No grade	0	205204
Duplicate record	0	205204
Non-public and out-of-district schools	10282	194922
Missing values for ALL items on OP form	0	194922
Out-of-range CR scores	0	194922
Remove Math Large Print	164	194758

Table 6c. NYSTP Mathematics Data Cleaning, Grade 5

Exclusion Rule	# Deleted	# Cases Remain
Initial N	0	199271
Out of grade	58	199213
No grade	2	199211
Duplicate record	0	199211
Non-public and out-of-district schools	2453	196758
Missing values for ALL items on OP form	1	196757
Out-of-range CR scores	0	196757
Remove Math Large Print	141	196616

Table 6d. NYSTP Mathematics Data Cleaning, Grade 6

Exclusion Rule	# Deleted	# Cases Remain
Initial N	0	202559
Out of grade	61	202498
No grade	0	202498
Duplicate record	0	202498
Non-public and out-of-district schools	5728	196770
Missing values for ALL items on OP form	1	196769
Out-of-range CR scores	0	196769

Table 6e. NYSTP Mathematics Data Cleaning, Grade 7

Exclusion Rule	# Deleted	# Cases Remain
Initial N	0	202871
Out of grade	96	202775
No grade	2	202773
Duplicate record	0	202773
Non-public and out-of-district schools	2601	200172
Missing values for ALL items on OP form	1	200171
Out-of-range CR scores	0	200171

Table 6f. NYSTP Mathematics Data Cleaning, Grade 8

Exclusion Rule	# Deleted	# Cases Remain
Initial N	0	218005
Out of grade	141	217864
No grade	3	217861
Duplicate record	0	217861
Non-public and out-of-district schools	12784	205077
Missing values for ALL items on OP form	4	205073
Out-of-range CR scores	0	205073

Classical Analysis and Calibration Sample Characteristics

The demographic characteristics of students in the classical analysis and calibration sample data sets are presented in the proceeding tables. 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 variables as it was found that the New York State population is fairly evenly split by gender categories.

Table 7a. Grade 3 Sample Characteristics (N = 197192)

Demographic Category		N-count	% of Total N-count
NRC	NYC	69154	35.2
	Big cities	8230	4.2
	Urban/Suburban	16240	8.3
	Rural	11488	5.8
	Average needs	58374	29.7
	Low needs	29660	15.7
	Charter	3435	1.8
Ethnicity	Asian	16158	8.2
	Black	36703	18.6
	Hispanic	42126	21.4
	American Indian	931	0.5
	Multi-Racial	657	0.3
	White	100519	51.0
	Unknown	98	0.1
ELL	No	181245	91.9
	Yes	15947	8.1
SWD	No	170929	86.7
	Yes	26263	13.3
SUA	No	151380	76.8
	Yes	45812	23.2

Table 7b. Grade 4 Sample Characteristics (N = 194758)

Demographic Category		N-count	% of Total N-count
NRC	NYC	67646	34.8
	Big cities	7979	4.1
	Urban/Suburban	15883	8.2
	Rural	11426	5.9
	Average needs	58897	30.3
	Low needs	29459	15.2
	Charter	2868	1.5
Ethnicity	Asian	14932	7. 7
	Black	36668	18.8
	Hispanic	41429	21.3
	American Indian	904	0.5
	Multi-Racial	494	0.3
	White	100247	51.5
	Unknown	84	0.0
ELL	No	181519	93.2
	Yes	13239	6.8

(Continued on next page)

Table 7b. Grade 4 Sample Characteristics (N = 194758) (cont.)

Demographic Category		N-count	% of Total N-count
SWD	No	166942	85.7
	Yes	27816	14.3
SUA	No	148981	76.5
	Yes	45777	23.5

Table 7c. Grade 5 Sample Characteristics (N = 196616)

Demographic Category		N-count	% of Total N-count
NRC	NYC	67625	34.5
	Big cities	7592	3.9
	Urban/Suburban	15572	8.0
	Rural	11416	5.8
	Average needs	59937	30.6
	Low needs	30261	15.5
	Charter	3514	1.8
Ethnicity	Asian	14916	7.6
	Black	37195	18.9
	Hispanic	41345	21.0
	American Indian	946	0.5
	Multi-Racial	504	0.2
	White	101607	51.7
	Unknown	103	0.1
ELL	No	185516	94.4
	Yes	11100	5.7
SWD	No	167072	85.0
	Yes	29544	15.0
SUA	No	150178	76.4
	Yes	46438	23.6

Table 7d. Grade 6 Sample Characteristics (N = 196769)

Demographic Category		N-count	% of Total N-count
NRC	NYC	67488	34.4
	Big cities	7596	3.9
	Urban/Suburban	15138	7.7
	Rural	11359	5.8
	Average needs	60984	31.1
	Low needs	30285	15.5
	Charter	3165	1.6

(Continued on next page)

Table 7d. Grade 6 Sample Characteristics (N = 196769) (cont.)

Demographic Category		N-count	% of Total N-count
Ethnicity	Asian	15077	7.7
	Black	37432	19.0
	Hispanic	40487	20.6
	American Indian	892	0.5
	Multi-Racial	444	0.2
	White	102363	52.0
	Unknown	74	0.0
ELL	No	187746	95.4
	Yes	9023	4.6
SWD	No	166989	84.9
	Yes	29780	15.1
SUA	No	154027	78.3
	Yes	42742	21.7

Table 7e. Grade 7 Sample Characteristics (N = 200171)

Demographic Category		N-count	% of Total N-count
NRC	NYC	68788	34.5
	Big cities	7778	4.0
	Urban/Suburban	15423	7.7
	Rural	11994	6.0
	Average needs	62195	31.2
	Low needs	30614	15.4
	Charter	2443	1.2
Ethnicity	Asian	15108	7.6
	Black	37530	18.8
	Hispanic	40878	20.4
	American Indian	928	0.5
	Multi-Racial	396	0.2
	White	105250	52.6
	Unknown	81	0.0
ELL	No	192287	96.1
	Yes	7884	3.9
SWD	No	170034	84.9
	Yes	30137	15.1
SUA	No	158602	79.2
	Yes	41569	20.8

Table 7f. Grade 8 Sample Characteristics (N = 205073)

Demographic Category		N-count	% of Total N-count
NRC	NYC	70677	34.7
	Big cities	7731	3.8
	Urban/Suburban	15647	7.7
	Rural	12252	6.0
	Average needs	64311	31.5
	Low needs	31243	15.3
	Charter	2140	1.1
Ethnicity	Asian	15317	7.5
	Black	38670	18.9
	Hispanic	41672	20.3
	American Indian	1000	0.5
	Multi-Racial	319	0.2
	White	108030	52.7
	Unknown	65	0.0
ELL	No	197075	96.1
	Yes	7998	3.9
SWD	No	175575	85.6
	Yes	29498	14.4
SUA	No	163597	79.8
	Yes	41476	20.2

Classical Data Analysis

Classical data analysis of the Grades 3–8 Mathematics 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 in which 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 Rescoring

One item in Grade 7 Spanish language version was rescored during the data analysis. In item 29 of the Spanish edition only of the Grade 7 Mathematics Test, the phrase *less than* was translated as *less*. As a result, students who used this edition might have written the wrong algebraic expression or might have been unable to answer this question. To adjust for this,

any student who used the Spanish edition, either exclusively or in conjunction with the English edition, was given credit for this item.

Item Difficulty and Response Distribution

Item difficulty and response distribution tables (Table 8a–8f) 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 selecting 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" and "% Sel" 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 for each MC item or the average percent 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 tests ranged from 0.40 to 0.97. For Grade 3, the item p-values were between 0.70 and 0.97 with a mean of 0.86. For Grade 4, the item p-values were between 0.45 and 0.96 with a mean of 0.77. For Grade 5, the item p-values were between 0.49 and 0.95 with a mean of 0.78. For Grade 6, the item p-values were between 0.52 and 0.96 with a mean of 0.72. For Grade 7, the item p-values were between 0.41 and 0.97 with a mean of 0.73. For Grade 8, the item p-values were between 0.40 and 0.93 with a mean of 0.74. These statistics are also provided in Table 8a–8f, along with other classical test summary statistics.

Table 8a. 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	197192	0.93	0.02	0.03	4.86	*93.05	1.15	0.89	-0.25	*0.39	-0.25	-0.19	0.39
2	197192	0.88	0.05	0.04	3.27	4.66	*88.43	3.56	-0.23	-0.24	*0.44	-0.24	0.44
3	197192	0.92	0.04	0.04	2.27	*92.30	2.92	2.43	-0.23	*0.42	-0.22	-0.26	0.42
4	197192	0.73	0.12	0.07	4.64	*72.75	9.13	13.29	-0.16	*0.53	-0.22	-0.41	0.53
5	197192	0.97	0.04	0.07	*97.45	0.78	0.67	0.99	*0.13	-0.06	-0.08	-0.07	0.13
6	197192	0.82	0.11	0.10	3.52	5.24	8.54	*82.49	-0.26	-0.20	-0.20	*0.39	0.39
7	197192	0.92	0.05	0.03	*92.46	4.17	1.85	1.44	*0.45	-0.29	-0.21	-0.25	0.45
8	197192	0.92	0.04	0.02	4.41	*92.42	0.94	2.17	-0.32	*0.46	-0.18	-0.27	0.46

(Continued on next page)

Table 8a. P-values, Scored Response Distributions, and Point Biserials, Grade 3 (cont.)

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
9	197192	0.84	0.08	0.04	4.46	*83.78	5.92	5.72	-0.27	*0.53	-0.2	-0.39	0.53
10	197192	0.90	0.08	0.08	2.69	2.84	4.48	*89.83	-0.26	-0.22	-0.27	*0.45	0.45
11	197192	0.88	0.06	0.03	*87.52	5.19	4.79	2.40	*0.42	-0.23	-0.23	-0.24	0.42
12	197192	0.92	0.05	0.04	2.35	4.51	*91.57	1.48	-0.29	-0.27	*0.44	-0.16	0.44
13	197192	0.87	0.10	0.08	5.46	5.52	1.88	*86.96	-0.30	-0.21	-0.20	*0.43	0.43
14	197192	0.92	0.06	0.05	1.16	*91.94	3.95	2.84	-0.15	*0.34	-0.19	-0.23	0.34
15	197192	0.97	0.07	0.05	0.64	0.48	1.53	*97.24	-0.13	-0.12	-0.15	*0.24	0.24
16	197192	0.70	0.16	0.04	8.04	14.43	*69.74	7.58	-0.26	-0.19	*0.45	-0.24	0.45
17	197192	0.77	0.07	0.03	0.76	2.08	19.76	*77.31	-0.16	-0.2	-0.34	*0.43	0.43
18	197192	0.77	0.05	0.04	1.61	3.01	18.20	*77.09	-0.20	-0.15	-0.30	*0.4	0.40
19	197192	0.93	0.08	0.03	*92.97	2.24	2.07	2.61	*0.27	-0.18	-0.17	-0.11	0.27
20	197192	0.88	0.13	0.06	7.37	0.97	*88.29	3.18	-0.43	-0.17	*0.55	-0.24	0.55
21	197192	0.95	0.10	0.04	*94.57	1.57	0.63	3.09	*0.37	-0.17	-0.16	-0.28	0.37
22	197192	0.88	0.12	0.06	2.99	4.86	4.44	*87.53	-0.27	-0.32	-0.21	*0.49	0.49
23	197192	0.86	0.16	0.07	6.62	3.90	3.15	*86.11	-0.24	-0.29	-0.28	*0.48	0.48
24	197192	0.80	0.29	0.21	*79.92	7.10	6.33	6.15	*0.36	-0.20	-0.20	-0.17	0.36
25	197192	0.84	0.59	0.02	6.82	*83.74	4.03	4.80	-0.36	*0.51	-0.18	-0.26	0.51
26	197192	0.70	0.06	26.15	6.68	67.12							
27	197192	0.76	0.10	16.20	16.11	67.59							
28	197192	0.86	0.07	4.13	19.10	76.69							
29	197192	0.76	0.09	16.59	15.40	67.92							
30	197192	0.79	0.12	9.66	7.90	17.86	64.46						
31	197192	0.87	0.12	1.32	2.91	27.43	68.23						

Table 8b. 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	194758	0.85	0.02	0.03	6.68	4.56	*85.28	3.43	-0.27	-0.17	*0.38	-0.17	0.38
2	194758	0.94	0.04	0.02	2.08	*93.52	2.60	1.74	-0.24	*0.35	-0.19	-0.16	0.35
3	194758	0.85	0.03	0.03	4.54	1.52	*85.26	8.62	-0.15	-0.09	*0.36	-0.31	0.36
4	194758	0.96	0.04	0.03	2.12	0.64	1.01	*96.17	-0.26	-0.14	-0.17	*0.35	0.35
5	194758	0.93	0.05	0.03	1.31	2.84	2.30	*93.47	-0.21	-0.19	-0.17	*0.33	0.33
6	194758	0.89	0.04	0.02	7.23	2.55	*89.26	0.91	-0.12	-0.13	*0.19	-0.10	0.19
7	194758	0.84	0.05	0.04	*83.89	7.19	4.37	4.45	*0.32	-0.13	-0.18	-0.23	0.32
8	194758	0.88	0.05	0.02	5.32	3.70	*88.27	2.64	-0.20	-0.22	*0.37	-0.19	0.37
9	194758	0.85	0.06	0.03	8.46	3.07	2.98	*85.41	-0.43	-0.23	-0.15	*0.53	0.53
10	194758	0.83	0.09	0.02	5.67	6.12	*83.41	4.68	-0.28	-0.18	*0.47	-0.31	0.47
11	194758	0.62	0.04	0.04	3.55	*62.07	25.44	8.85	-0.17	*0.38	-0.28	-0.11	0.38
12	194758	0.82	0.05	0.05	4.25	9.43	3.75	*82.47	-0.33	-0.18	-0.19	*0.41	0.41
13	194758	0.82	0.07	0.02	13.24	*81.54	4.09	1.04	-0.49	*0.51	-0.07	-0.14	0.51
14	194758	0.74	0.12	0.03	6.71	13.24	*73.95	5.96	-0.22	-0.27	*0.49	-0.29	0.49
15	194758	0.78	0.07	0.03	9.74	7.02	*78.05	5.09	-0.34	-0.16	*0.46	-0.22	0.46

(Continued on next page)

Table 8b. P-values, Scored Response Distributions, and Point Biserials, Grade 4 (cont.)

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
16	194758	0.86	0.07	0.02	7.92	*86.14	3.04	2.81	-0.18	*0.31	-0.19	-0.16	0.31
17	194758	0.86	0.08	0.03	6.46	*86.13	3.33	3.96	-0.34	*0.52	-0.21	-0.29	0.52
18	194758	0.85	0.08	0.05	6.44	2.92	5.70	*84.80	-0.34	-0.23	-0.25	*0.50	0.50
19	194758	0.75	0.11	0.03	4.91	*75.37	8.61	10.98	-0.24	*0.43	-0.25	-0.20	0.43
20	194758	0.61	0.08	0.03	10.33	*60.77	12.64	16.15	-0.24	*0.48	-0.07	-0.37	0.48
21	194758	0.84	0.10	0.04	*83.93	4.10	8.29	3.55	*0.46	-0.26	-0.23	-0.27	0.46
22	194758	0.65	0.08	0.07	*64.84	3.21	4.72	27.08	*0.48	-0.11	-0.16	-0.39	0.48
23	194758	0.60	0.11	0.06	12.24	6.12	21.77	*59.70	-0.12	-0.32	-0.05	*0.28	0.28
24	194758	0.92	0.12	0.03	2.00	3.12	*92.43	2.29	-0.23	-0.26	*0.43	-0.23	0.43
25	194758	0.68	0.15	0.03	*68.26	5.01	24.07	2.47	*0.26	-0.22	-0.09	-0.20	0.26
26	194758	0.79	0.17	0.04	15.37	*78.52	2.82	3.08	-0.37	*0.53	-0.25	-0.22	0.53
27	194758	0.45	0.20	0.06	21.70	22.77	9.97	*45.30	-0.09	0.00	-0.22	*0.22	0.22
28	194758	0.87	0.21	0.03	6.94	3.19	*86.51	3.12	-0.19	-0.22	*0.38	-0.22	0.38
29	194758	0.60	0.36	0.04	3.71	*60.20	10.45	25.24	-0.15	*0.53	-0.25	-0.35	0.53
30	194758	0.65	0.57	0.03	*65.25	8.06	11.99	14.11	*0.41	-0.17	-0.19	-0.24	0.41
31	194758	0.69	0.05	22.11	18.15	59.69							
32	194758	0.84	0.05	7.67	16.65	75.62							
33	194758	0.77	0.10	17.70	9.48	72.72							
34	194758	0.80	0.20	14.17	10.62	75.01							
35	194758	0.60	0.15	27.66	25.12	47.06							
36	194758	0.71	0.07	12.92	32.03	54.99							
37	194758	0.72	0.12	14.23	27.23	58.41							
38	194758	0.65	0.14	11.53	20.77	28.34	39.21						
39	194758	0.68	0.24	6.22	10.80	56.34	26.40						
40	194758	0.84	0.05	9.03	12.88	78.05							
41	194758	0.73	0.07	13.61	26.04	60.28							
42	194758	0.78	0.08	14.24	14.87	70.81							
43	194758	0.87	0.09	4.32	17.15	78.44							
44	194758	0.63	0.15	30.06	13.77	56.02							
45	194758	0.81	0.16	11.51	14.26	74.08							
46	194758	0.88	0.07	2.31	19.41	78.21							
47	194758	0.83	0.08	4.59	13.51	11.42	70.40						
48	194758	0.73	0.08	9.39	6.95	37.67	45.91						

Table 8c. 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	196616	0.93	0.01	0.01	0.63	1.03	4.86	*93.46	-0.08	-0.16	-0.28	*0.33	0.33
2	196616	0.95	0.02	0.01	0.87	1.62	*95.49	2.00	-0.16	-0.18	*0.27	-0.12	0.27
3	196616	0.91	0.03	0.04	2.02	3.97	2.65	*91.30	-0.25	-0.29	-0.12	*0.4	0.40
4	196616	0.85	0.03	0.03	*85.33	10.38	2.22	2.00	*0.26	-0.17	-0.15	-0.11	0.26
5	196616	0.81	0.03	0.04	2.10	5.30	*80.67	11.86	-0.1	-0.09	*0.46	-0.46	0.46
6	196616	0.84	0.03	0.03	1.91	12.77	1.21	*84.05	-0.12	-0.34	-0.15	*0.39	0.39

(Continued on next page)

Table 8c. P-values, Scored Response Distributions, and Point Biserials, Grade 5 (cont.)

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
7	196616	0.77	0.06	0.04	7.56	6.01	9.26	*77.07	-0.23	-0.19	-0.31	*0.47	0.47
8	196616	0.91	0.03	0.02	1.67	*91.45	3.82	3.00	-0.15	*0.37	-0.24	-0.22	0.37
9	196616	0.91	0.02	0.02	0.88	3.10	4.49	*91.49	-0.17	-0.21	-0.19	*0.33	0.33
10	196616	0.65	0.03	0.03	10.10	15.40	*64.91	9.52	-0.20	-0.17	*0.43	-0.29	0.43
11	196616	0.86	0.04	0.03	*85.94	3.17	8.85	1.98	*0.38	-0.19	-0.29	-0.10	0.38
12	196616	0.78	0.09	0.03	5.62	8.39	7.47	*78.41	-0.34	-0.26	-0.23	*0.51	0.51
13	196616	0.90	0.04	0.03	*90.01	3.57	3.22	3.14	*0.42	-0.21	-0.26	-0.23	0.42
14	196616	0.80	0.07	0.04	4.50	10.91	4.29	*80.19	-0.28	-0.36	-0.18	*0.52	0.52
15	196616	0.69	0.15	0.03	*69.26	9.45	10.11	11.01	*0.38	-0.15	-0.19	-0.22	0.38
16	196616	0.88	0.05	0.03	2.28	4.29	4.89	*88.46	-0.21	-0.19	-0.25	*0.39	0.39
17	196616	0.79	0.04	0.03	5.04	1.99	14.31	*78.59	-0.21	-0.21	-0.17	*0.33	0.33
18	196616	0.72	0.05	0.02	*71.95	1.97	24.07	1.94	*0.44	-0.15	-0.38	-0.11	0.44
19	196616	0.81	0.06	0.02	5.45	6.91	*81.34	6.22	-0.20	-0.18	*0.40	-0.25	0.40
20	196616	0.73	0.10	0.05	*73.34	10.02	7.73	8.78	*0.47	-0.22	-0.21	-0.29	0.47
21	196616	0.70	0.09	0.05	2.73	21.01	6.02	*70.09	-0.19	-0.33	-0.22	*0.48	0.48
22	196616	0.66	0.12	0.04	*66.15	2.18	13.38	18.13	*0.60	-0.15	-0.22	-0.48	0.60
23	196616	0.72	0.24	0.04	6.11	9.74	*72.47	11.40	-0.22	-0.24	*0.50	-0.31	0.50
24	196616	0.83	0.14	0.06	2.02	2.12	12.63	*83.04	-0.19	-0.15	-0.19	*0.31	0.31
25	196616	0.72	0.29	0.03	10.70	*71.89	8.02	9.08	-0.30	*0.39	-0.12	-0.16	0.39
26	196616	0.77	0.63	0.02	*77.18	7.71	9.03	5.43	*0.44	-0.20	-0.27	-0.22	0.44
27	196616	0.69	0.12	24.33	12.27	63.28							
28	196616	0.74	0.09	12.20	26.98	60.73							
29	196616	0.74	0.09	13.70	23.83	62.39							
30	196616	0.88	0.07	6.67	10.63	82.63							
31	196616	0.63	0.09	10.84	15.57	45.96	27.53						
32	196616	0.49	0.11	10.23	43.26	34.54	11.86						
33	196616	0.60	0.13	16.88	19.82	29.27	33.91						
34	196616	0.77	0.38	8.23	17.72	8.95	64.71						

Table 8d. 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	Pbis Option 1	Pbis Option 2	Pbis Option 3	Pbis Option 4	Pbis Key
1	196769	0.96	0.02	0.01	1.14	1.72	*96.01	1.09	-0.16	-0.23	*0.31	-0.12	0.31
2	196769	0.70	0.09	0.02	*70.27	10.68	8.61	10.32	*0.49	-0.29	-0.11	-0.33	0.49
3	196769	0.84	0.03	0.01	9.39	4.71	*83.91	1.95	-0.29	-0.18	*0.39	-0.16	0.39
4	196769	0.96	0.02	0.03	1.69	0.39	2.27	*95.60	-0.23	-0.13	-0.20	*0.33	0.33
5	196769	0.78	0.10	0.01	10.89	9.59	*78.15	1.26	-0.21	-0.37	*0.47	-0.15	0.47
6	196769	0.82	0.04	0.02	5.08	2.42	*81.58	10.87	-0.28	-0.19	*0.38	-0.18	0.38
7	196769	0.65	0.10	0.01	*65.08	20.22	10.74	3.85	*0.59	-0.41	-0.23	-0.22	0.59
8	196769	0.79	0.11	0.01	4.61	*79.38	4.79	11.09	-0.14	*0.36	-0.22	-0.21	0.36
9	196769	0.68	0.10	0.01	7.31	*68.21	16.48	7.88	-0.33	*0.44	-0.19	-0.18	0.44
10	196769	0.57	0.12	0.02	27.28	9.70	*57.25	5.63	-0.34	-0.19	*0.49	-0.13	0.49
11	196769	0.81	0.05	0.03	11.94	2.61	4.03	*81.34	-0.32	-0.27	-0.20	*0.48	0.48
12	196769	0.72	0.08	0.01	7.65	*72.02	10.48	9.76	-0.11	*0.49	-0.43	-0.20	0.49

(Continued on next page)

Table 8d. P-values, Scored Response Distributions, and Point Biserials, Grade 6 (cont.)

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
13	196769	0.67	0.05	0.02	14.46	*66.51	4.13	14.83	-0.05	*0.37	-0.22	-0.33	0.37
14	196769	0.77	0.06	0.01	4.98	11.11	*77.22	6.61	-0.21	-0.21	*0.36	-0.16	0.36
15	196769	0.68	0.07	0.02	3.80	19.84	7.85	*68.42	-0.22	-0.47	-0.20	*0.61	0.61
16	196769	0.64	0.08	0.02	2.13	*64.31	30.17	3.28	-0.20	*0.41	-0.31	-0.14	0.41
17	196769	0.80	0.10	0.02	12.01	3.00	4.54	*80.33	-0.33	-0.26	-0.27	*0.53	0.53
18	196769	0.78	0.13	0.02	12.21	4.83	4.34	*78.47	-0.18	-0.17	-0.14	*0.31	0.31
19	196769	0.78	0.07	0.02	*77.71	15.13	2.57	4.50	*0.36	-0.20	-0.20	-0.22	0.36
20	196769	0.74	0.14	0.02	4.41	8.26	13.28	*73.89	-0.25	-0.21	-0.14	*0.36	0.36
21	196769	0.61	0.16	0.03	3.92	17.30	*60.81	17.77	-0.23	-0.35	*0.54	-0.21	0.54
22	196769	0.76	0.19	0.03	3.41	11.98	8.59	*75.81	-0.20	-0.25	-0.29	*0.47	0.47
23	196769	0.62	0.19	0.03	3.83	14.50	19.83	*61.61	-0.14	-0.22	-0.34	*0.50	0.50
24	196769	0.67	0.28	0.02	19.44	*66.87	8.97	4.42	-0.16	*0.38	-0.23	-0.24	0.38
25	196769	0.73	0.36	0.02	*73.25	1.84	1.83	22.71	*0.36	-0.17	-0.24	-0.24	0.36
26	196769	0.84	0.16	8.99	13.79	77.05							
27	196769	0.55	0.27	19.43	51.48	28.82							
28	196769	0.54	0.19	40.06	11.55	48.20							
29	196769	0.74	0.26	14.60	21.28	63.85							
30	196769	0.70	0.17	18.14	22.55	59.14							
31	196769	0.65	0.24	30.07	8.75	60.94							
32	196769	0.61	0.28	14.91	16.99	38.29	29.53						
33	196769	0.52	0.28	30.36	20.50	9.76	39.10						
34	196769	0.72	0.10	2.23	10.06	56.01	31.61						
35	196769	0.85	0.36	5.30	5.39	16.20	72.75						

Table 8e. 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	200171	0.90	0.07	0.01	*89.71	2.53	3.53	4.15	*0.33	-0.20	-0.15	-0.20	0.33
2	200171	0.88	0.03	0.01	2.39	9.3	*87.50	0.78	-0.11	-0.34	*0.39	-0.13	0.39
3	200171	0.80	0.12	0.01	4.88	5.56	*79.92	9.50	-0.23	-0.31	*0.52	-0.29	0.52
4	200171	0.97	0.03	0.01	0.37	1.00	1.14	*97.44	-0.12	-0.1	-0.14	*0.21	0.21
5	200171	0.78	0.05	0.01	1.54	2.11	*77.96	18.34	-0.16	-0.17	*0.36	-0.27	0.36
6	200171	0.74	0.03	0.02	10.44	12.24	*74.48	2.79	-0.18	-0.04	*0.22	-0.14	0.22
7	200171	0.61	0.09	0.03	5.02	28.90	4.72	*61.23	-0.14	-0.25	-0.18	*0.38	0.38
8	200171	0.83	0.06	0.01	5.63	*82.89	4.60	6.80	-0.26	*0.43	-0.28	-0.16	0.43
9	200171	0.93	0.03	0.01	*93.16	1.84	2.41	2.54	*0.28	-0.15	-0.19	-0.13	0.28
10	200171	0.69	0.09	0.02	2.88	7.98	20.34	*68.70	-0.19	-0.27	-0.14	*0.35	0.35
11	200171	0.78	0.03	0.01	1.72	2.03	*77.71	18.50	-0.18	-0.18	*0.38	-0.27	0.38
12	200171	0.88	0.05	0.01	4.97	5.35	*87.83	1.80	-0.23	-0.24	*0.40	-0.18	0.40
13	200171	0.53	0.25	0.02	15.18	*53.04	16.03	15.48	-0.31	*0.33	-0.27	0.13	0.33
14	200171	0.64	0.07	0.02	24.54	2.33	9.20	*63.85	0.13	-0.16	-0.3	*0.12	0.12
15	200171	0.91	0.04	0.01	1.34	4.30	*91.11	3.19	-0.13	-0.07	*0.16	-0.09	0.16
16	200171	0.84	0.18	0.01	4.80	*83.78	7.08	4.16	-0.27	*0.45	-0.22	-0.26	0.45
17	200171	0.91	0.05	0.01	3.91	*90.62	3.31	2.09	-0.13	*0.25	-0.15	-0.13	0.25
18	200171	0.82	0.07	0.01	3.32	7.89	*82.21	6.48	-0.24	-0.21	*0.48	-0.34	0.48

(Continued on next page)

Table 8e. P-values, Scored Response Distributions, and Point Biserials, Grade 7 (cont.)

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
19	200171	0.67	0.17	0.01	10.28	*66.67	12.23	10.64	-0.24	*0.45	-0.25	-0.18	0.45
20	200171	0.55	0.13	0.01	4.24	28.31	*54.84	12.47	-0.08	-0.09	*0.30	-0.28	0.30
21	200171	0.85	0.05	0.04	4.12	3.69	7.08	*85.00	-0.20	-0.18	-0.26	*0.40	0.40
22	200171	0.82	0.07	0.02	*82.08	3.64	12.49	1.70	*0.50	-0.16	-0.41	-0.16	0.50
23	200171	0.78	0.07	0.02	5.99	*78.30	5.07	10.54	-0.25	*0.50	-0.30	-0.25	0.50
24	200171	0.52	0.17	0.03	4.57	24.67	18.21	*52.34	-0.15	-0.25	-0.20	*0.44	0.44
25	200171	0.54	0.12	0.02	*53.97	24.39	10.85	10.65	*0.40	-0.13	-0.26	-0.20	0.40
26	200171	0.72	0.11	0.02	10.25	13.59	*72.20	3.83	-0.24	-0.25	*0.40	-0.10	0.40
27	200171	0.73	0.13	0.02	16.53	8.41	*72.80	2.11	-0.47	-0.19	*0.54	-0.09	0.54
28	200171	0.66	0.12	0.03	11.93	9.50	11.96	*66.47	-0.23	-0.14	-0.28	*0.44	0.44
29	200171	0.66	0.12	0.01	28.37	*66.43	2.68	2.38	-0.37	*0.48	-0.22	-0.17	0.48
30	200171	0.89	0.14	0.01	4.93	*88.79	1.58	4.55	-0.17	*0.26	-0.15	-0.11	0.26
31	200171	0.81	0.23	13.38	11.12	75.28							
32	200171	0.51	0.64	33.73	29.43	36.20							
33	200171	0.72	0.31	16.63	22.26	60.79							
34	200171	0.47	1.12	37.13	29.97	31.78							
35	200171	0.67	0.19	17.37	15.51	14.34	52.59						
36	200171	0.41	0.30	41.63	17.95	15.65	24.47						
37	200171	0.66	0.55	25.97	4.90	11.70	56.89						
38	200171	0.75	0.44	13.92	4.79	23.51	57.34						

Table 8f. 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	Pbis Option 1	Pbis Option 2	Pbis Option 3	Pbis Option 4	Pbis Key
1	205073	0.91	0.05	0.01	3.99	2.81	*91.09	2.06	-0.13	-0.15	*0.27	-0.17	0.27
2	205073	0.88	0.05	0.01	2.10	*88.27	7.41	2.16	-0.22	*0.43	-0.30	-0.17	0.43
3	205073	0.72	0.06	0.01	19.34	*71.94	3.36	5.29	-0.30	*0.45	-0.16	-0.25	0.45
4	205073	0.79	0.07	0.01	4.12	5.68	*79.19	10.94	-0.14	-0.24	*0.39	-0.23	0.39
5	205073	0.87	0.07	0.01	2.77	*86.65	9.05	1.45	-0.20	*0.46	-0.34	-0.22	0.46
6	205073	0.68	0.14	0.02	9.04	15.70	*67.82	7.28	-0.21	-0.30	*0.53	-0.31	0.53
7	205073	0.67	0.06	0.01	24.07	7.16	1.86	*66.83	-0.38	-0.17	-0.15	*0.48	0.48
8	205073	0.73	0.13	0.02	10.49	10.72	5.44	*73.21	-0.36	-0.27	-0.22	*0.55	0.55
9	205073	0.75	0.06	0.01	9.21	10.65	*75.16	4.91	-0.29	-0.25	*0.48	-0.22	0.48
10	205073	0.70	0.07	0.01	9.92	3.16	17.11	*69.73	-0.30	-0.24	-0.36	*0.58	0.58
11	205073	0.89	0.05	0.01	3.74	3.11	*88.87	4.21	-0.25	-0.20	*0.43	-0.26	0.43
12	205073	0.60	0.11	0.01	10.90	*60.11	21.97	6.90	-0.10	*0.35	-0.19	-0.24	0.35
13	205073	0.85	0.07	0.01	4.25	5.03	*85.28	5.37	-0.23	-0.27	*0.48	-0.28	0.48
14	205073	0.67	0.07	0.01	4.72	*66.59	12.51	16.11	-0.25	*0.51	-0.35	-0.19	0.51
15	205073	0.79	0.07	0.02	*79.34	10.86	4.22	5.48	*0.33	-0.16	-0.20	-0.18	0.33
16	205073	0.64	0.13	0.02	6.29	*64.32	13.22	16.01	-0.20	*0.56	-0.37	-0.25	0.56

(Continued on next page)

Table 8f. P-values, Scored Response Distributions, and Point Biserials, Grade 8 (cont.)

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
17	205073	0.88	0.07	0.02	*87.53	7.59	3.51	1.28	*0.49	-0.33	-0.28	-0.19	0.49
18	205073	0.89	0.05	0.02	0.69	7.21	2.93	*89.10	-0.14	-0.31	-0.16	*0.38	0.38
19	205073	0.76	0.08	0.01	1.86	9.71	12.25	*76.08	-0.21	-0.16	-0.30	*0.41	0.41
20	205073	0.84	0.07	0.01	1.60	12.88	*84.43	1.01	-0.19	-0.29	*0.39	-0.19	0.39
21	205073	0.78	0.12	0.02	*77.86	7.61	10.53	3.86	*0.55	-0.31	-0.30	-0.27	0.55
22	205073	0.93	0.09	0.01	2.43	2.62	*93.14	1.71	-0.21	-0.29	*0.4	-0.15	0.40
23	205073	0.40	0.09	0.02	42.60	5.92	11.85	*39.52	-0.15	-0.35	-0.22	*0.46	0.46
24	205073	0.88	0.09	0.01	5.11	5.10	*87.73	1.97	-0.21	-0.33	*0.47	-0.23	0.47
25	205073	0.92	0.11	0.01	3.71	2.79	*92.06	1.32	-0.27	-0.27	*0.42	-0.16	0.42
26	205073	0.83	0.14	0.01	*83.08	3.31	3.64	9.82	*0.33	-0.24	-0.14	-0.17	0.33
27	205073	0.75	0.23	0.01	11.07	*74.80	5.48	8.41	-0.30	*0.48	-0.15	-0.28	0.48
28	205073	0.87	0.38	10.44	4.71	84.47							
29	205073	0.77	0.51	12.59	20.25	66.64							
30	205073	0.57	0.57	19.08	46.80	33.54							
31	205073	0.62	0.45	26.69	20.81	52.05							
32	205073	0.64	0.83	25.22	8.93	11.08	53.94						
33	205073	0.73	0.39	11.08	11.18	25.17	52.18						
34	205073	0.80	0.38	9.88	20.14	69.60							
35	205073	0.73	1.24	10.22	30.63	57.92							
36	205073	0.76	0.89	15.37	15.71	68.03							
37	205073	0.67	0.80	16.88	31.15	51.17							
38	205073	0.53	0.67	34.82	22.10	42.41							
39	205073	0.62	0.79	22.92	28.24	48.04							
40	205073	0.57	1.31	21.34	39.73	37.61							
41	205073	0.65	0.51	15.88	37.32	46.28							
42	205073	0.72	0.41	12.98	11.64	20.51	54.47						
43	205073	0.61	0.46	8.69	17.59	53.40	19.86						
44	205073	0.64	1.05	18.57	18.86	12.64	48.87						
45	205073	0.61	1.08	26.36	13.06	9.38	50.12						

Point-Biserial Correlation Coefficients

Point biserial statistics are used to examine item-test correlations or item discrimination. In the Tables 8a–8f, 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. Point biserials for the correct answer option should be equal to or greater than 0.15, which would indicate 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 indicates that students who scored lower on the overall test had a tendency to pick a distractor. Grade 3 item 5 and Grade 7 item 14 were the only items flagged for having a point biserial for the correct answer below 0.15. Point biserials for correct answer options (pbis*) on the tests ranged from 0.12–0.61. For Grade 3, the pbis* were between 0.13 and 0.55. For Grade 4, the pbis* were between 0.19 and 0.53. For Grade 5, the pbis* were between 0.26 and 0.60. For Grade 6, pbis* were between 0.31 and 0.61. For Grade 7, the pbis* were between 0.12 and 0.54. For Grade 8, the pbis* were between 0.27 and 0.58.

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 choice).The proportions of students selecting incorrect responses while responding to MC items are provided in Tables 8a–8f. Distribution of student responses across answer choices was evaluated. It is expected that the proportion of students selecting the correct answer will be higher than proportions of students selecting any other answer choice. This was true for all New York State mathematics items.

As mentioned in the "Point Biserial Correlations Coefficients" subsection, items were flagged if the point biserial of any distractor was positive. One Grade 7 item was flagged for positive point biserial values on distractor (incorrect) answer options (item 14, 0.13).

Test Statistics and Reliability Coefficients

Test statistics including raw-score mean and standard deviation are presented in Table 9. Reliability coefficients provide measures of internal consistency that range from zero to one. Two reliability coefficients: Cronbach's alpha and Feldt-Raju were computed for the Grades 3–8 Mathematics 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 0.88–0.94. Feldt-Raju reliability coefficients ranged from 0.90–0.95. The lowest reliability was observed for the Grade 3 test, but as that test has the lowest number of score points it is reasonable that its reliability would not be as high as the other grades' tests. The highest reliability was observed for the Grade 8 test. All reliabilities exceeded 0.85, across statistics, which is a good indication that the NYSTP 3–8 Mathematics Tests are acceptably reliable. High reliability indicates that scores are consistent and not unduly influenced by random error. (For more information on test reliability and standard error of measurement, see Section VIII, "Reliability and Standard Error of Measurement.")

Table 9. NYSTP Mathematics 2009 Test Form Statistics and Reliability

Grade	Max RS	RS Mean	RS SD	P-value Mean	Minimum P-value	Maximum P-value	Cronbach's Alpha	Feldt-Raju Alpha
3	39	32.93	6.54	0.84	0.70	0.97	0.88	0.90
4	70	53.62	13.37	0.77	0.45	0.96	0.94	0.94
5	46	34.53	8.60	0.75	0.49	0.95	0.89	0.91
6	49	34.72	10.32	0.71	0.52	0.96	0.91	0.92
7	50	35.31	10.23	0.71	0.41	0.97	0.90	0.92
8	69	49.17	15.54	0.71	0.40	0.93	0.94	0.95

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 did 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 8a–8f show the omit rates for items on the Grades 3–8 Mathematics 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 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), test language (focal group: Spanish; reference group: English) and English language learners (focal group: English language learners; reference group: Non-English language learners). All cases in clean data sets were used to compute DIF statistics. Table 10 shows the number of students in each focal and reference group.

Table 10. NYSTP Mathematics 2009 Classical DIF Sample N-Counts

Grade	Ethnicity				Gender		Needs Resource Category		Test Language	
	Black	Hispanic	Asian	White	Female	Male	High	Low	Spanish	English
3	36703	42126	16158	102205	96496	100696	105112	88034	3442	193228
4	36668	41429	14932	101729	95075	99683	102934	88356	2906	191308
5	37195	41345	14916	103160	96669	99947	102205	90198	2818	193243
6	37432	40487	15077	103773	96129	100640	101581	91269	2984	193066
7	37530	40878	15108	106655	97556	102615	103983	92809	3283	196012
8	38670	41672	15317	109414	100768	104305	106307	95554	3155	201088

Table 11 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 11. Number of Items Flagged by SMD and Mantel-Haenszel DIF Methods

Grade	Number of Flagged Items
3	1
4	6
5	6
6	6
7	14
8	10

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 D.

Section VI: IRT Scaling and Equating

IRT Models and Rationale for Use

Item response theory (IRT) allows comparisons between items and examinees, 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 & 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 model (2PPC) (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 (Lord & Novick, 1968; Lord, 1980) 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 Mathematics Tests. It should be noted that the scaling was done on approximately 98% of the New York State school student population. Exclusion of some cases during the data cleaning process had a very small effect on parameter estimation. The exclusion rules are described in details in final OP test technical reports. As shown in Tables 12 through 14, the 2009 samples were comparable to 2008 populations in terms of needs resource category (NRC), student ethnicity, proportions of English language learner proficiency, proportions of students with disabilities, and proportions of students using testing accommodations.

Table 12. 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	35.60	35.07	35.69	34.73
Big cities	4.16	4.17	3.93	4.10
Urban/Suburban	8.11	8.24	7.96	8.16
Rural	5.81	5.83	5.79	5.87
Average needs	29.73	29.60	30.01	30.24
Low needs	14.96	15.04	15.28	15.13
Charter	1.52	1.74	1.21	1.47
Missing	0.11	0.31	0.13	0.31
ETHNICITY				
Asian	7.49	8.19	7.37	7.67
Black	19.11	18.61	19.22	18.83
Hispanics	21.44	21.36	21.22	21.27
American Indian	0.49	0.47	0.48	0.46
Multi-Racial	0.12	0.33	0.10	0.25
White	51.30	50.98	51.57	51.47
Unknown	0.03	0.05	0.05	0.04
ELL STATUS				
No	92.13	91.91	93.32	93.2
Yes	7.87	8.09	6.68	6.80
DISABILITY				
No	86.45	86.68	85.13	85.72
Yes	13.55	13.32	14.87	14.28
ACCOMMODATIONS				
No	79.23	76.77	78.37	76.50
Yes	20.77	23.23	21.63	23.50

Table 13. 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	35.21	34.39	34.68	34.30
Big cities	3.83	3.86	3.80	3.86
Urban/Suburban	7.71	7.92	7.68	7.69
Rural	5.70	5.81	5.82	5.77
Average needs	30.31	30.48	30.79	30.99
Low needs	15.44	15.39	15.69	15.39
Charter	1.65	1.79	1.37	1.61
Missing	0.14	0.36	0.16	0.38
ETHNICITY				
Asian	7.54	7.59	7.45	7.66
Black	19.29	18.92	18.84	19.02
Hispanics	20.79	21.03	20.53	20.58
American Indian	0.46	0.48	0.45	0.45
Multi-Racial	0.09	0.26	0.09	0.23
White	51.81	51.68	52.60	52.02
Unknown	0.04	0.05	0.04	0.04
ELL STATUS				
No	94.69	94.35	95.70	95.41
Yes	5.31	5.65	4.30	4.59
DISABILITY				
No	84.93	84.97	84.88	84.87
Yes	15.07	15.03	15.12	15.13
ACCOMMODATIONS				
No	79.11	76.38	80.40	78.28
Yes	20.89	23.62	19.60	21.72

Table 14. 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	34.70	34.36	34.71	34.98
Big cities	3.96	3.89	4.00	3.83
Urban/Suburban	7.60	7.70	7.64	7.74
Rural	5.99	5.99	6.22	6.06
Average needs	31.16	31.07	31.43	31.83
Low needs	15.28	15.29	15.01	15.46
Charter	1.10	1.22	0.67	1.06
Missing	0.21	0.47	0.32	0.53
ETHNICITY				
Asian	7.18	7.55	7.06	7.47
Black	19.22	18.75	19.26	18.86
Hispanics	20.36	20.42	19.97	20.32
American Indian	0.49	0.46	0.50	0.49
Multi-Racial	0.07	0.20	0.06	0.16
White	52.66	52.58	53.13	52.68
Unknown	0.03	0.04	0.03	0.03
ELL STATUS				
No	96.08	96.06	96.57	96.10
Yes	3.92	3.94	3.43	3.90
DISABILITY				
No	85.43	84.94	85.65	85.62
Yes	14.57	15.06	14.35	14.38
ACCOMMODATIONS				
No	80.86	79.23	81.24	79.78
Yes	19.14	20.77	18.76	20.22

.

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 EM (expectation-maximization) algorithm (Bock & Aitkin, 1981; Thissen, 1982). Simulation studies have compared PARDUX with MULTILOG (Thissen, 1991), PARSCALE (Muraki & Bock, 1991), and BIGSTEPS (Wright & 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 Mathematics Tests item calibrations did not incur any problems. The number of estimation cycles was set to 50 with convergence criterion of 0.001 for all grades. The maximum value of *a*-parameter was set to 3.4, and range for *b*-parameter 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 always been used for calibration of NYS test data. The estimated *a*- and *b*-parameters were in the original theta metric and all the items were well within the prescribed parameter ranges. The *c*-parameter was not estimated in the 2009 calibration process but was fixed and remained unchanged for anchor items between FT and OP administration. It should be noted that there were a number of items with the default value for the *c*-parameter on the OP test. When the PARDUX program encounters difficulty estimating the *c*-parameter, it assigns a default *c*-parameter value of 0.2000. These default values of *c*-parameter were obtained during the FT calibration and remained unchanged between FT and OP administrations. Table 15 presents a summary of calibration results. For the Grades 3–8 Mathematics Tests, all of the calibration estimation results are reasonable.

Table 15. NYSTP Mathematics 2009 Calibration Results

Grade	Largest *a*-parameter	Lowest and highest *b*-parameters		# Items with Default *c*-parameters	Theta Mean	Theta Standard Deviation	# Students
3	2.336	-4.124	-0.946	15	0.37	1.714	197192
4	2.280	-4.284	0.869	15	0.06	1.158	194758
5	2.659	-3.528	-0.183	7	0.07	1.199	196616
6	2.967	-4.195	0.207	8	0.05	1.197	196769
7	2.231	-4.315	0.415	14	0.05	1.207	200171
8	2.501	-3.550	1.745	10	0.05	1.171	205073

Note that *a*- and *b*-parameters are based on OP calibrations and default *c*-parameter values are based on FT calibrations

Item-Model Fit

Item fit statistics discern the appropriateness of using an item in the 3PL or 2PPC model. A 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

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

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 degree 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 still will increase with sample size, all else being equal. To use this standardized statistic to flag items for potential misfit, 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, which 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_{Q1} was greater than the value of Z_{Q1} critical. If the obtained Z_{Q1} was less than Z_{Q1} critical the items were rated as having acceptable fit. It should be noted that most items in the NYSTP 2009 Mathematics Tests demonstrated a good model fit, further supporting use of the chosen models. No items in Grades 3 or 7 exhibited poor item-model fit statistics. The following items exhibited misfit: Grade 4 items 39 (Z_{Q1}= 1289.67, Z_{Q1} critical= 514.63) and 48 (Z_{Q1}= 563.59, Z_{Q1} critical= 514.64), Grade 5 items 31 (Z_{Q1}= 809.61, Z_{Q1} critical= 517.46) and 33 (Z_{Q1}= 715.48, Z_{Q1} critical= 517.46), Grade 6 item 35 (Z_{Q1}= 541.20, Z_{Q1} critical= 517.85), and Grade 8 items 23 (Z_{Q1}= 1223.18, Z_{Q1} critical= 540.45), 30 (Z_{Q1}= 614.71, Z_{Q1} critical= 540.45), 34 (Z_{Q1}= 745.42, Z_{Q1} critical= 540.45), 35 (Z_{Q1}= 858.36, Z_{Q1} critical= 540.45), and 44 (Z_{Q1}= 818.19, Z_{Q1} critical= 540.45). Fit statistics and status for all items in the Grades 3–8 Mathematics Tests are presented in Tables 16 to 21.

Table 16. Mathematics Grade 3 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	Z_{Q1}	Z_{Q1} critical	Fit OK?
1	3PL	333.14	7	170139	87.16	453.704	Y
2	3PL	229.69	7	170139	59.52	453.704	Y
3	3PL	113.78	7	170139	28.54	453.704	Y
4	3PL	781.23	7	170139	206.92	453.704	Y
5	3PL	230.09	7	170139	59.62	453.704	Y
6	3PL	971.37	7	170139	257.74	453.704	Y
7	3PL	86.55	7	170139	21.26	453.704	Y
8	3PL	152.90	7	170139	38.99	453.704	Y
9	3PL	662.65	7	170139	175.23	453.704	Y
10	3PL	153.27	7	170139	39.09	453.704	Y
11	3PL	203.25	7	170139	52.45	453.704	Y
12	3PL	157.79	7	170139	40.30	453.704	Y
13	3PL	255.56	7	170139	66.43	453.704	Y
14	3PL	302.84	7	170139	79.07	453.704	Y
15	3PL	122.75	7	170139	30.94	453.704	Y
16	3PL	1427.33	7	170139	379.60	453.704	Y
17	3PL	771.86	7	170139	204.42	453.704	Y
18	3PL	1386.86	7	170139	368.78	453.704	Y
19	3PL	294.03	7	170139	76.71	453.704	Y
20	3PL	275.73	7	170139	71.82	453.704	Y
21	3PL	100.63	7	170139	25.02	453.704	Y
22	3PL	227.54	7	170139	58.94	453.704	Y
23	3PL	217.18	7	170139	56.17	453.704	Y
24	3PL	1281.90	7	170139	340.73	453.704	Y
25	3PL	659.53	7	170139	174.40	453.704	Y
26	2PPC	944.62	17	170139	159.09	453.704	Y
27	2PPC	1410.46	17	170139	238.98	453.704	Y
28	2PPC	2204.15	17	170139	375.09	453.704	Y
29	2PPC	1499.07	17	170139	254.17	453.704	Y
30	2PPC	2637.00	26	170139	362.08	453.704	Y
31	2PPC	1986.36	26	170139	271.85	453.704	Y

Table 17. Mathematics Grade 4 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	Z_{Q1}	Z_{Q1} critical	Fit OK?
1	3PL	158.91	7	192988	40.60	514.6347	Y
2	3PL	39.31	7	192988	8.64	514.6347	Y
3	3PL	25.55	7	192988	4.96	514.6347	Y
4	3PL	53.07	7	192988	12.31	514.6347	Y
5	3PL	131.53	7	192988	33.28	514.6347	Y
6	3PL	681.82	7	192988	180.35	514.6347	Y
7	3PL	20.49	7	192988	3.61	514.6347	Y
8	3PL	248.18	7	192988	64.46	514.6347	Y
9	3PL	33.41	7	192988	7.06	514.6347	Y
10	3PL	44.56	7	192988	10.04	514.6347	Y
11	3PL	47.91	7	192988	10.93	514.6347	Y
12	3PL	341.95	7	192988	89.52	514.6347	Y
13	3PL	48.56	7	192988	11.11	514.6347	Y
14	3PL	39.03	7	192988	8.56	514.6347	Y
15	3PL	60.53	7	192988	14.31	514.6347	Y
16	3PL	102.82	7	192988	25.61	514.6347	Y
17	3PL	50.15	7	192988	11.53	514.6347	Y
18	3PL	93.25	7	192988	23.05	514.6347	Y
19	3PL	90.40	7	192988	22.29	514.6347	Y
20	3PL	198.71	7	192988	51.24	514.6347	Y
21	3PL	105.14	7	192988	26.23	514.6347	Y
22	3PL	300.95	7	192988	78.56	514.6347	Y
23	3PL	100.12	7	192988	24.89	514.6347	Y
24	3PL	73.66	7	192988	17.82	514.6347	Y
25	3PL	338.43	7	192988	88.58	514.6347	Y
26	3PL	113.26	7	192988	28.40	514.6347	Y
27	3PL	1558.27	7	192988	414.59	514.6347	Y
28	3PL	84.08	7	192988	20.60	514.6347	Y
29	3PL	155.70	7	192988	39.74	514.6347	Y
30	3PL	95.74	7	192988	23.72	514.6347	Y
31	2PPC	700.65	17	192988	117.24	514.6347	Y
32	2PPC	210.63	17	192988	33.21	514.6347	Y
33	2PPC	417.39	17	192988	68.67	514.6347	Y
34	2PPC	443.48	17	192988	73.14	514.6347	Y
35	2PPC	2220.80	17	192988	377.95	514.6347	Y
36	2PPC	697.16	17	192988	116.65	514.6347	Y
37	2PPC	370.68	17	192988	60.65	514.6347	Y
38	2PPC	1839.29	26	192988	251.46	514.6347	Y
39	2PPC	9325.98	26	192988	1289.67	514.6347	N
40	2PPC	218.26	17	192988	34.52	514.6347	Y
41	2PPC	500.56	17	192988	82.93	514.6347	Y
42	2PPC	728.06	17	192988	121.95	514.6347	Y
43	2PPC	116.57	17	192988	17.08	514.6347	Y
44	2PPC	320.74	17	192988	52.09	514.6347	Y

(Continued on next page)

Table 17. Mathematics Grade 4 Item Fit Statistics (cont.)

Item	Model	Chi Square	DF	Total N	Z_{Q1}	Z_{Q1} critical	Fit OK?
45	2PPC	425.32	17	192988	70.03	514.6347	Y
46	2PPC	485.55	17	192988	80.36	514.6347	Y
47	2PPC	1072.95	26	192988	145.19	514.6347	Y
48	2PPC	4090.09	26	192988	563.59	514.6347	N

Table 18. Mathematics Grade 5 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	Z_{Q1}	Z_{Q1} critical	Fit OK?
1	3PL	88.78	7	194048	21.86	517.4613	Y
2	3PL	84.17	7	194048	20.63	517.4613	Y
3	3PL	93.44	7	194048	23.10	517.4613	Y
4	3PL	581.75	7	194048	153.61	517.4613	Y
5	3PL	61.92	7	194048	14.68	517.4613	Y
6	3PL	62.60	7	194048	14.86	517.4613	Y
7	3PL	369.42	7	194048	96.86	517.4613	Y
8	3PL	181.93	7	194048	46.75	517.4613	Y
9	3PL	32.31	7	194048	6.76	517.4613	Y
10	3PL	272.28	7	194048	70.90	517.4613	Y
11	3PL	173.52	7	194048	44.50	517.4613	Y
12	3PL	161.37	7	194048	41.26	517.4613	Y
13	3PL	62.78	7	194048	14.91	517.4613	Y
14	3PL	112.36	7	194048	28.16	517.4613	Y
15	3PL	366.04	7	194048	95.96	517.4613	Y
16	3PL	44.90	7	194048	10.13	517.4613	Y
17	3PL	61.18	7	194048	14.48	517.4613	Y
18	3PL	110.93	7	194048	27.78	517.4613	Y
19	3PL	446.64	7	194048	117.50	517.4613	Y
20	3PL	452.92	7	194048	119.18	517.4613	Y
21	3PL	164.73	7	194048	42.16	517.4613	Y
22	3PL	435.04	7	194048	114.40	517.4613	Y
23	3PL	119.21	7	194048	29.99	517.4613	Y
24	3PL	83.06	7	194048	20.33	517.4613	Y
25	3PL	80.41	7	194048	19.62	517.4613	Y
26	3PL	118.25	7	194048	29.73	517.4613	Y
27	2PPC	402.94	17	194048	66.19	517.4613	Y
28	2PPC	564.96	17	194048	93.97	517.4613	Y
29	2PPC	2646.84	17	194048	451.01	517.4613	Y
30	2PPC	337.56	17	194048	54.98	517.4613	Y
31	2PPC	5864.19	26	194048	809.61	517.4613	N
32	2PPC	2223.64	26	194048	304.76	517.4613	Y
33	2PPC	5185.37	26	194048	715.48	517.4613	N
34	2PPC	1052.38	26	194048	142.33	517.4613	Y

Table 19. Mathematics Grade 6 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	Z_{Q1}	Z_{Q1} critical	Fit OK?
1	3PL	368.45	7	194194	96.60	517.8507	Y
2	3PL	58.47	7	194194	13.76	517.8507	Y
3	3PL	25.68	7	194194	4.99	517.8507	Y
4	3PL	166.41	7	194194	42.60	517.8507	Y
5	3PL	422.33	7	194194	111.00	517.8507	Y
6	3PL	161.83	7	194194	41.38	517.8507	Y
7	3PL	288.45	7	194194	75.22	517.8507	Y
8	3PL	211.77	7	194194	54.73	517.8507	Y
9	3PL	106.55	7	194194	26.60	517.8507	Y
10	3PL	516.13	7	194194	136.07	517.8507	Y
11	3PL	108.32	7	194194	27.08	517.8507	Y
12	3PL	36.16	7	194194	7.79	517.8507	Y
13	3PL	274.36	7	194194	71.46	517.8507	Y
14	3PL	704.52	7	194194	186.42	517.8507	Y
15	3PL	372.37	7	194194	97.65	517.8507	Y
16	3PL	141.87	7	194194	36.05	517.8507	Y
17	3PL	57.14	7	194194	13.40	517.8507	Y
18	3PL	827.79	7	194194	219.37	517.8507	Y
19	3PL	52.03	7	194194	12.03	517.8507	Y
20	3PL	227.01	7	194194	58.80	517.8507	Y
21	3PL	327.62	7	194194	85.69	517.8507	Y
22	3PL	106.96	7	194194	26.71	517.8507	Y
23	3PL	168.52	7	194194	43.17	517.8507	Y
24	3PL	109.10	7	194194	27.29	517.8507	Y
25	3PL	212.71	7	194194	54.98	517.8507	Y
26	2PPC	801.14	17	194194	134.48	517.8507	Y
27	2PPC	542.81	17	194194	90.18	517.8507	Y
28	2PPC	922.51	17	194194	155.29	517.8507	Y
29	2PPC	1142.88	17	194194	193.09	517.8507	Y
30	2PPC	1013.73	17	194194	170.94	517.8507	Y
31	2PPC	621.59	17	194194	103.69	517.8507	Y
32	2PPC	1246.57	26	194194	169.26	517.8507	Y
33	2PPC	2864.30	26	194194	393.60	517.8507	Y
34	2PPC	771.47	26	194194	103.38	517.8507	Y
35	2PPC	3928.66	26	194194	541.20	517.8507	N

Table 20. Mathematics Grade 7 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	Z_{Q1}	Z_{Q1} critical	Fit OK?
1	3PL	125.70	7	196666	31.72	524.4427	Y
2	3PL	136.76	7	196666	34.68	524.4427	Y
3	3PL	305.63	7	196666	79.81	524.4427	Y
4	3PL	236.70	7	196666	61.39	524.4427	Y
5	3PL	1021.83	7	196666	271.22	524.4427	Y
6	3PL	190.19	7	196666	48.96	524.4427	Y
7	3PL	259.15	7	196666	67.39	524.4427	Y
8	3PL	48.55	7	196666	11.10	524.4427	Y
9	3PL	116.16	7	196666	29.18	524.4427	Y
10	3PL	295.47	7	196666	77.10	524.4427	Y
11	3PL	883.26	7	196666	234.19	524.4427	Y
12	3PL	251.41	7	196666	65.32	524.4427	Y
13	3PL	1085.91	7	196666	288.35	524.4427	Y
14	3PL	1784.08	7	196666	474.94	524.4427	Y
15	3PL	234.53	7	196666	60.81	524.4427	Y
16	3PL	41.88	7	196666	9.32	524.4427	Y
17	3PL	420.83	7	196666	110.60	524.4427	Y
18	3PL	82.93	7	196666	20.29	524.4427	Y
19	3PL	149.45	7	196666	38.07	524.4427	Y
20	3PL	705.86	7	196666	186.78	524.4427	Y
21	3PL	174.77	7	196666	44.84	524.4427	Y
22	3PL	332.41	7	196666	86.97	524.4427	Y
23	3PL	51.28	7	196666	11.83	524.4427	Y
24	3PL	130.75	7	196666	33.07	524.4427	Y
25	3PL	1276.28	7	196666	339.23	524.4427	Y
26	3PL	44.73	7	196666	10.08	524.4427	Y
27	3PL	151.53	7	196666	38.63	524.4427	Y
28	3PL	86.67	7	196666	21.29	524.4427	Y
29	3PL	402.63	7	196666	105.74	524.4427	Y
30	3PL	122.29	7	196666	30.81	524.4427	Y
31	2PPC	236.82	17	196666	37.70	524.4427	Y
32	2PPC	1731.61	17	196666	294.05	524.4427	Y
33	2PPC	1215.84	17	196666	205.60	524.4427	Y
34	2PPC	1187.35	17	196666	200.71	524.4427	Y
35	2PPC	424.13	26	196666	55.21	524.4427	Y
36	2PPC	2312.67	26	196666	317.10	524.4427	Y
37	2PPC	734.48	26	196666	98.25	524.4427	Y
38	2PPC	798.84	26	196666	107.17	524.4427	Y

Table 21. Mathematics Grade 8 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	Z_{Q1}	Z_{Q1} critical	Fit OK?
1	3PL	190.34	7	202667	49.00	540.4453	Y
2	3PL	44.05	7	202667	9.90	540.4453	Y
3	3PL	197.45	7	202667	50.90	540.4453	Y
4	3PL	43.41	7	202667	9.73	540.4453	Y
5	3PL	172.81	7	202667	44.32	540.4453	Y
6	3PL	85.78	7	202667	21.06	540.4453	Y
7	3PL	483.15	7	202667	127.26	540.4453	Y
8	3PL	555.02	7	202667	146.46	540.4453	Y
9	3PL	221.57	7	202667	57.35	540.4453	Y
10	3PL	67.89	7	202667	16.27	540.4453	Y
11	3PL	194.57	7	202667	50.13	540.4453	Y
12	3PL	112.96	7	202667	28.32	540.4453	Y
13	3PL	333.51	7	202667	87.26	540.4453	Y
14	3PL	169.91	7	202667	43.54	540.4453	Y
15	3PL	398.20	7	202667	104.55	540.4453	Y
16	3PL	171.12	7	202667	43.86	540.4453	Y
17	3PL	471.15	7	202667	124.05	540.4453	Y
18	3PL	243.65	7	202667	63.25	540.4453	Y
19	3PL	51.47	7	202667	11.88	540.4453	Y
20	3PL	237.53	7	202667	61.61	540.4453	Y
21	3PL	145.85	7	202667	37.11	540.4453	Y
22	3PL	48.04	7	202667	10.97	540.4453	Y
23	3PL	4583.73	7	202667	1223.18	540.4453	N
24	3PL	91.94	7	202667	22.70	540.4453	Y
25	3PL	46.23	7	202667	10.49	540.4453	Y
26	3PL	347.61	7	202667	91.03	540.4453	Y
27	3PL	641.20	7	202667	169.50	540.4453	Y
28	2PPC	232.28	17	202667	36.92	540.4453	Y
29	2PPC	2948.16	17	202667	502.69	540.4453	Y
30	2PPC	3601.33	17	202667	614.71	540.4453	N
31	2PPC	241.24	17	202667	38.46	540.4453	Y
32	2PPC	1941.66	26	202667	265.65	540.4453	Y
33	2PPC	2053.61	26	202667	281.18	540.4453	Y
34	2PPC	4363.49	17	202667	745.42	540.4453	N
35	2PPC	5022.06	17	202667	858.36	540.4453	N
36	2PPC	1170.62	17	202667	197.84	540.4453	Y
37	2PPC	2142.30	17	202667	364.49	540.4453	Y
38	2PPC	850.45	17	202667	142.93	540.4453	Y
39	2PPC	1180.82	17	202667	199.59	540.4453	Y
40	2PPC	249.90	17	202667	39.94	540.4453	Y
41	2PPC	2958.56	17	202667	504.47	540.4453	Y
42	2PPC	430.44	26	202667	56.09	540.4453	Y
43	2PPC	2862.47	26	202667	393.35	540.4453	Y
44	2PPC	5926.04	26	202667	818.19	540.4453	N
45	2PPC	1532.70	26	202667	208.94	540.4453	Y

Local Independence

In using IRT models, one of the assumptions made is that the items are locally independent; that is, student response on one item are not dependent upon their response to another item. Statistically speaking, when a student's ability is accounted for, their responses to each item are statistically independent.

One way to assess the validity of this assumption, and 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 their overall test performance. The Q_3 for binary items was computed as follows:

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

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. The content of the flagged items was examined in order to identify possible sources of the local dependence.

The Q_3 statistics were examined on all of the Grade 3–8 Mathematics Tests and no items were found to be locally dependent in Grade 3–7. In Grade 8, four pairs of items are found to be locally dependent: items 28 and 29 (Q_3 = 0.209), item 29 and 34 (Q_3 = 0.247), items 32 and 45 (Q_3 = 0.302), item 36 and 44 (Q_3 = 0.238). The magnitudes of these statistics were not sufficient to warrant any concern. Anchor items were excluded from Q_3 computation.

Scaling and Equating

The 2009 Grades 3–8 Mathematics assessments were calibrated and equated to the operational scales 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: Grades 3 through 8 Mathematics Field Test Technical Report*, page 44.

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 NYS OP scale).

In the second equating procedure, the 2009 Mathematics Tests OP data were calibrated after the 2009 OP administration. Equated to OP scale FT parameters for all MC items in OP tests were used as anchors to transform the 2009 OP item parameters to NYS 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 & Lord, 1983). In this procedure, new OP parameter estimates were obtained for all items. The *a*-parameters and *b*-parameters were allowed to be estimated freely while *c*-parameters of anchor items were fixed.

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 the calibration process into the final scale score metric. Table 22 presents the 2009 OP transformation parameters for New York State Grades 3–8 Mathematics Tests.

Table 22. NYSTP Mathematics 2009 Final Transformation Constants

Grade	*M*1	*M*2
3	23.2617	685.6043
4	33.1307	688.7083
5	27.9180	685.4723
6	29.5609	679.2296
7	25.2681	679.7096
8	28.4984	674.5153

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. Outlined below, procedures 1 and 2 refer to evaluation of the overall anchor set and procedures 3, 4, and 5 were applied to evaluate individual anchor items.

1. Anchor set input and estimate of TCC alignment. The overall alignment of TCCs for anchor set input and estimate was evaluated to determine the overall stability of anchor item parameters between FT and 2009 OP administration.

2. Correlations of anchor input and estimate of *a*- and *b*-parameters and p-values. Correlations of anchor input and estimate of *a*- and *b*-parameters and p-values was evaluated for magnitude. Ideally, the correlations between anchor input and estimate

for *a*-parameter should be at least 0.80 and for *b*-parameter and p-value should be at least 0.90.

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. The differential item performance was evaluated by examining previous (input/FT) and transformed (estimated/OP) item parameters. The items with an absolute difference of parameters greater than two times the root mean square deviation are 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, and/or 2007) and the 2009 OP administration will be calculated. The p-values will then be 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 perpendicular distance to the line-of-best-fit. The fitted line is chosen so as to minimize the sum of squared perpendicular distances of the points to the line. Items lying more than two standard deviations of the distance away 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 item is flagged for performance differences between the two item administrations.

Table 23 provides a summary of anchor item evaluation and item flags.

Table 23. Mathematics Anchor Evaluation Summary

Grade	Number of Anchors	Anchor Input/Estimate Correlation			Flagged Anchors			
		a-par	*b*-par	p-value	RMSD *a*-par	RMSD *b*-par	Delta	Lord's Chi-square
3	25	0.720	0.925	0.960	3	14, 15	5, 12	
4	30	0.872	0.965	0.976	4	4, 5, 25	12, 25	
5	26	0.778	0.828	0.907	1	1,2	9	1, 2, 9,
6	25	0.843	0.888	0.933	15, 18	11	11	11
7	30	0.942	0.917	0.903	30	28	28	9, 28, 30
8	27	0.824	0.910	0.935	20	25	25	25

In all cases, the overall TCC alignment for anchor set input and estimate parameters was very good (see Figures 1–6). The correlations for input and estimate p-values were over 0.90 for all grades. Correlations for *b*-parameter input and estimates ranged from 0.83 for Grade 5 to 0.97 for Grade 4. Correlations for *a*-parameter input and estimate ranged from 0.72 for Grade 3 to 0.94 for Grade 7. A number of items were flagged by different statistical methods in each grade. If an item was flagged by at least three different methods a test calibration run was conducted without this item and the parameter estimates obtained using a reduced anchor set were compared with the parameter estimates obtained using an intact anchor set. The following items were removed on a trial basis from anchor sets: item 1 in Grade 5, item 11 in Grade 6, item 28 in Grade 7, and item 25 in Grade 8. It was determined that although in each case removal of anchor items from the anchor set had very small effect on individual item parameter estimates, some impact on the scoring tables and student scores was observed for all affected grades.

The equating results and impact data from multiple equating runs for Grades 5, 6, 7, and 8 were carefully evaluated. Because the overall anchor set TCC alignments were very good in equating runs with full anchor sets and the differences in scale score means and standard deviations between the two equating runs were very small, all anchor items were retained for the purpose of OP data equating.

Figure 1. Mathematics Grade 3 Anchor Set TCC Alignment

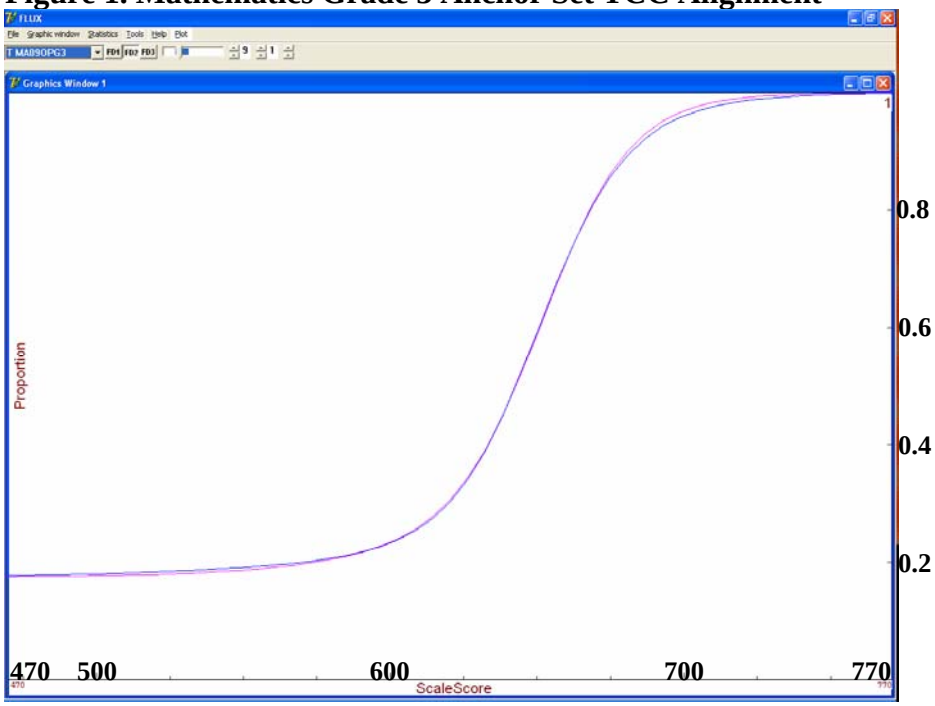


Figure 2. Mathematics Grade 4 Anchor Set TCC Alignment

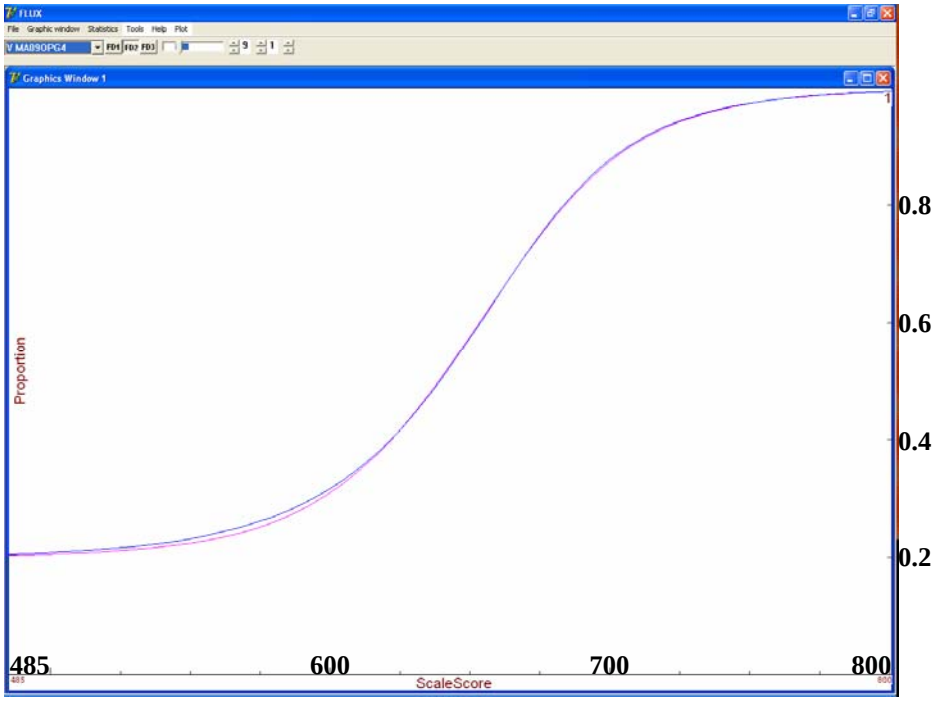


Figure 3. Mathematics Grade 5 Anchor Set TCC Alignment

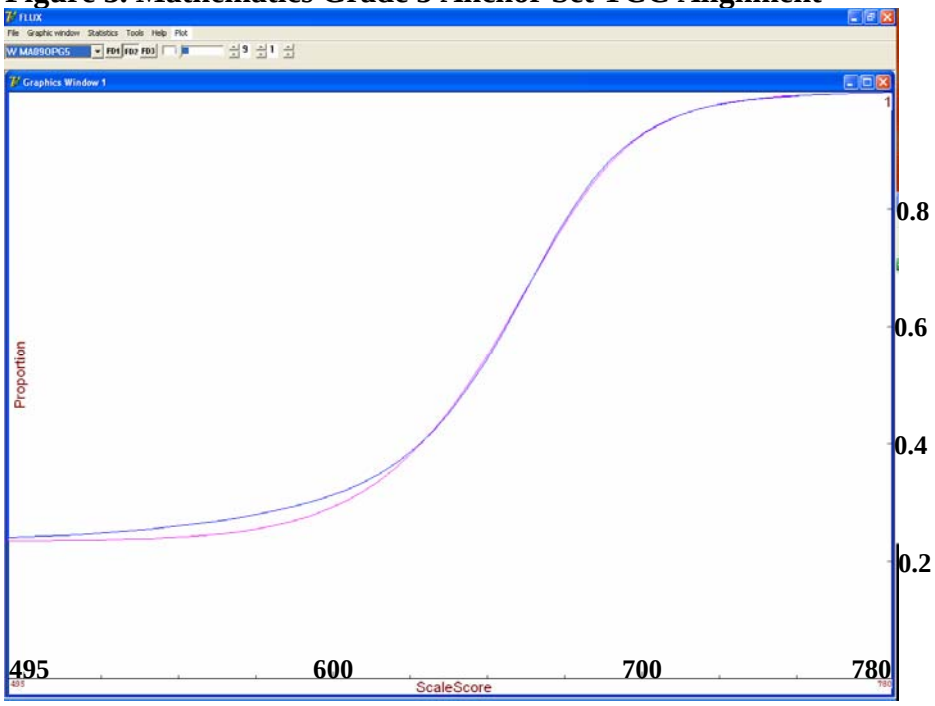


Figure 4. Mathematics Grade 6 Anchor Set TCC Alignment

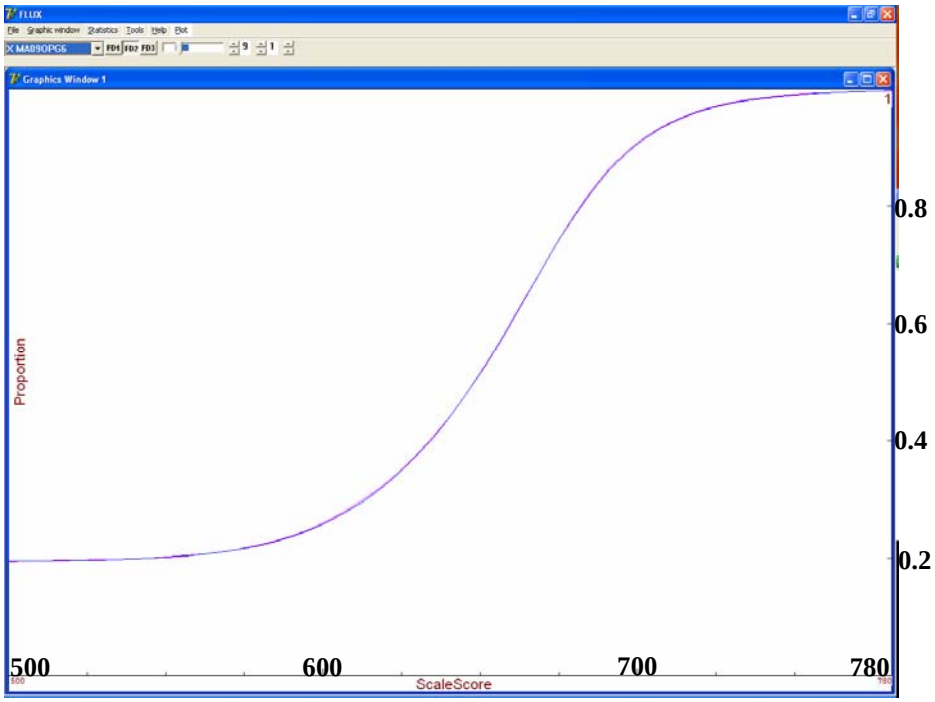


Figure 5. Mathematics Grade 7 Anchor Set TCC Alignment

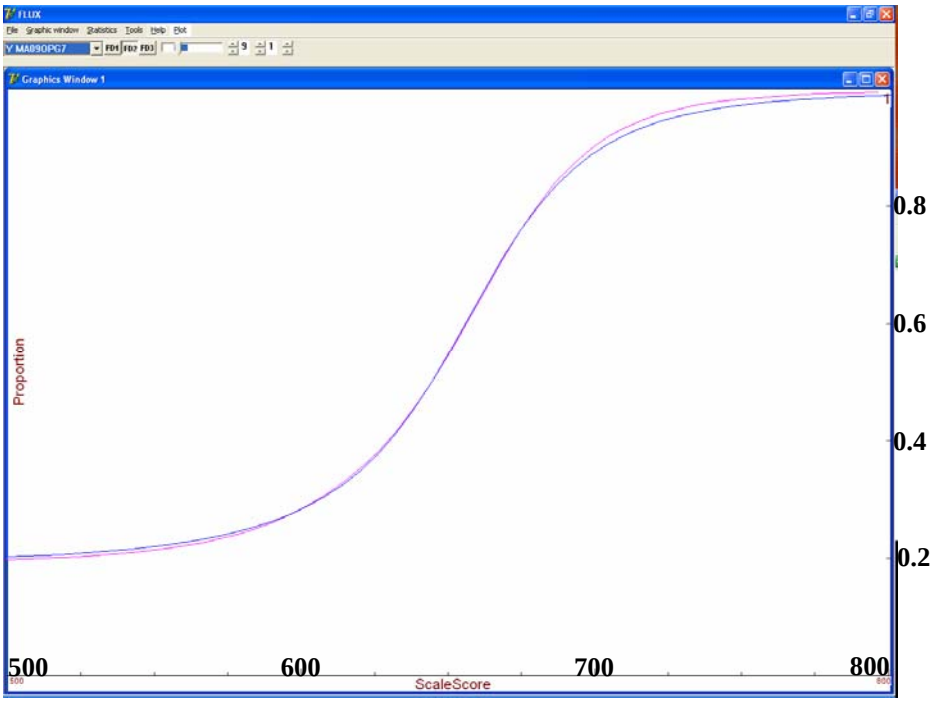
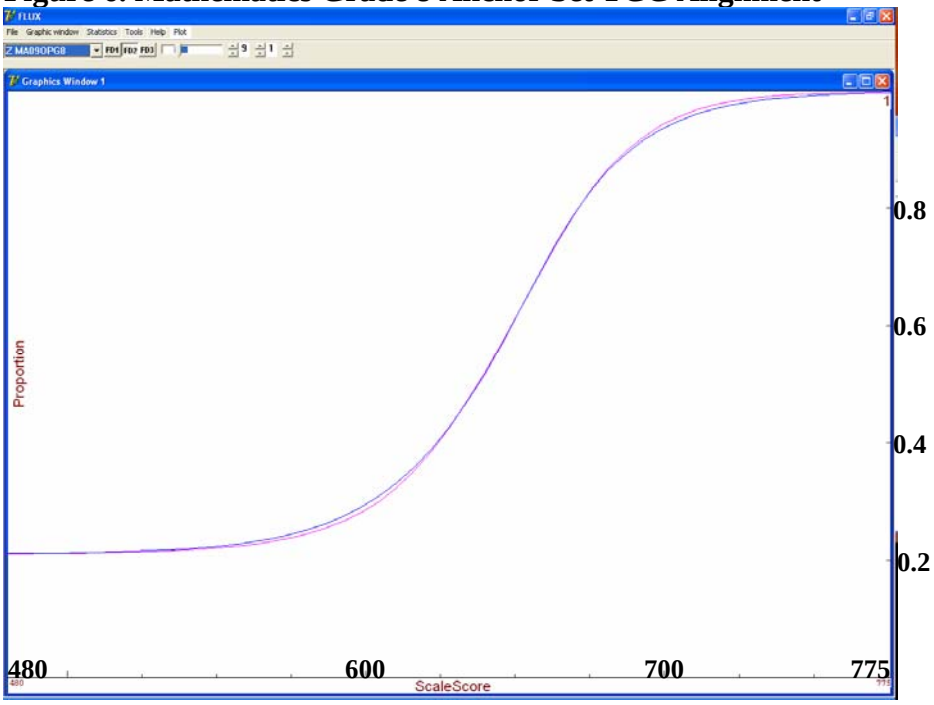


Figure 6. Mathematics Grade 8 Anchor Set TCC Alignment



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 24–29. The parameter estimates are expressed in scale score metric and are defined below:

- a-parameter a 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 24. Grade 3 2009 Operational Item Parameter Estimates

Item	Max Pts	*a*-par/ alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3
1	1	0.04076	638.3604	0.2000	
2	1	0.04308	650.3649	0.2000	
3	1	0.04504	641.1812	0.1286	
4	1	0.05198	669.5657	0.0774	
5	1	0.01960	574.6539	0.2000	
6	1	0.03416	656.6857	0.2000	
7	1	0.05129	644.7859	0.2000	
8	1	0.05224	645.0713	0.2000	
9	1	0.05466	659.6345	0.1564	
10	1	0.04864	649.7283	0.2000	
11	1	0.04064	651.1115	0.2000	
12	1	0.04694	645.8329	0.2251	
13	1	0.04170	652.4905	0.2000	
14	1	0.03480	637.5522	0.2000	
15	1	0.03345	613.0953	0.2000	
16	1	0.03915	671.3913	0.1033	
17	1	0.03929	665.9806	0.2000	
18	1	0.03277	661.4185	0.1257	
19	1	0.02722	625.3511	0.2000	
20	1	0.05908	652.7739	0.0982	
21	1	0.04357	635.3165	0.2000	
22	1	0.05336	655.1657	0.2000	
23	1	0.04615	652.9433	0.0976	
24	1	0.03298	660.2497	0.1966	
25	1	0.04891	658.1097	0.1417	
26	2	0.03577	25.4415	22.2589	
27	2	0.04661	31.1705	30.5023	
28	2	0.03266	20.2467	20.8875	
29	2	0.05511	36.8590	36.2648	
30	3	0.04757	31.5677	31.1694	31.3895
31	3	0.03810	24.0735	23.1658	25.1311

Table 25. Grade 4 2009 Operational Item Parameter Estimates

Item	Max Pts	*a*-par/ alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3
1	1	0.02128	636.3474	0.1904	
2	1	0.02612	615.7037	0.2000	
3	1	0.02100	636.6584	0.2000	
4	1	0.03124	608.0414	0.2000	
5	1	0.02358	611.2455	0.2000	
6	1	0.01206	589.9807	0.2000	
7	1	0.01752	633.9554	0.2000	
8	1	0.02319	631.2784	0.2000	
9	1	0.03730	648.3708	0.1543	
10	1	0.03051	650.7490	0.2000	
11	1	0.02104	682.5303	0.1639	
12	1	0.02253	645.7947	0.2000	
13	1	0.03311	654.8842	0.1769	
14	1	0.03049	665.9245	0.1544	
15	1	0.02744	658.1861	0.1695	
16	1	0.01793	627.8869	0.2000	
17	1	0.03661	647.7330	0.1937	
18	1	0.03671	653.1106	0.2548	
19	1	0.03334	674.3480	0.3323	
20	1	0.03040	683.9712	0.1377	
21	1	0.02838	648.2090	0.2000	
22	1	0.03018	679.9946	0.1684	
23	1	0.01508	689.7354	0.2000	
24	1	0.03279	627.8872	0.2000	
25	1	0.02022	693.5255	0.3975	
26	1	0.03590	660.6686	0.1537	
27	1	0.01758	717.7935	0.1772	
28	1	0.02198	634.4703	0.2000	
29	1	0.04049	684.9492	0.1269	
30	1	0.02415	680.475	0.2000	
31	2	0.03784	25.4521	24.9890	
32	2	0.03662	23.2118	23.5186	
33	2	0.03670	24.9424	23.1838	
34	2	0.03321	22.2587	20.8291	
35	2	0.04497	30.2959	30.7136	
36	2	0.03139	19.9996	21.1538	
37	2	0.04786	31.0196	32.2721	
38	3	0.02667	17.0531	17.8467	18.3292
39	3	0.02089	13.2563	12.5384	15.3373
40	2	0.04626	29.8123	29.7467	

(Continued on next page)

Table 25. Grade 4 2009 Operational Item Parameter Estimates (cont.)

Item	Max Pts	*a*-par/ alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3
41	2	0.04334	28.0281	29.0496	
42	2	0.03917	25.8285	25.3422	
43	2	0.03178	19.3956	20.1595	
44	2	0.04588	31.5437	30.4682	
45	2	0.04097	26.7163	26.4086	
46	2	0.03377	19.8260	21.6115	
47	3	0.01982	11.8943	13.4877	11.8451
48	3	0.02902	19.2690	17.8789	19.9730

Table 26. Grade 5 2009 Operational Item Parameter Estimates

Item	Max Pts	*a*-par/ alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3
1	1	0.03127	624.6389	0.2000	
2	1	0.02731	609.4871	0.2000	
3	1	0.03531	635.2612	0.2000	
4	1	0.02381	657.4570	0.4829	
5	1	0.03418	655.3853	0.1553	
6	1	0.03037	651.1280	0.2507	
7	1	0.03830	664.6898	0.2089	
8	1	0.03197	632.0676	0.2000	
9	1	0.02830	631.9723	0.2995	
10	1	0.04042	682.8077	0.2473	
11	1	0.04068	662.9254	0.4788	
12	1	0.04065	660.8179	0.1476	
13	1	0.04118	647.7031	0.3654	
14	1	0.04335	659.1381	0.1503	
15	1	0.02673	672.3020	0.2000	
16	1	0.03529	648.7734	0.3614	
17	1	0.02162	652.4904	0.2000	
18	1	0.03045	667.1588	0.1582	
19	1	0.02666	651.2480	0.1964	
20	1	0.04215	674.7303	0.3122	
21	1	0.03812	671.9055	0.1618	
22	1	0.05602	675.3091	0.0981	
23	1	0.04126	669.5604	0.1651	
24	1	0.02036	640.7250	0.2000	
25	1	0.02447	663.4974	0.1363	
26	1	0.03878	668.3229	0.2936	

(Continued on next page)

Table 26. Grade 5 2009 Operational Item Parameter Estimates (cont.)

Item	Max Pts	*a*-par/ alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3
27	2	0.04301	29.4116	27.9315	
28	2	0.04320	27.8684	28.8117	
29	2	0.05346	34.8353	35.6800	
30	2	0.02596	16.7509	15.6302	
31	3	0.02465	16.0274	15.6729	17.6289
32	3	0.03587	22.4782	24.8486	26.3505
33	3	0.02764	18.2730	18.4477	19.0873
34	3	0.02737	17.3167	19.1271	16.8678

Table 27. Grade 6 2009 Operational Item Parameter Estimates

Item	Max Pts	*a*-par/ alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3
1	1	0.03385	606.9011	0.2000	
2	1	0.03237	662.1589	0.1251	
3	1	0.02697	639.5656	0.2000	
4	1	0.03673	612.0471	0.2000	
5	1	0.03110	651.2698	0.1560	
6	1	0.02781	650.8179	0.3116	
7	1	0.05370	670.9966	0.1238	
8	1	0.02363	648.4630	0.2405	
9	1	0.03467	671.1152	0.2339	
10	1	0.04703	681.8214	0.1723	
11	1	0.03551	649.5164	0.2000	
12	1	0.03570	662.6204	0.1687	
13	1	0.02278	668.7786	0.2000	
14	1	0.02328	650.3940	0.2000	
15	1	0.05905	667.9041	0.1270	
16	1	0.02901	673.1973	0.1898	
17	1	0.04227	652.6847	0.1764	
18	1	0.02290	656.7939	0.3443	
19	1	0.02173	646.2958	0.1762	
20	1	0.02147	654.8708	0.2000	
21	1	0.04372	675.0790	0.1216	
22	1	0.03236	657.0815	0.1919	
23	1	0.04104	676.9852	0.1831	
24	1	0.02404	668.1420	0.1899	
25	1	0.02111	655.8598	0.2000	
26	2	0.02626	16.8175	16.0300	

(Continued on next page)

Table 27. Grade 6 2009 Operational Item Parameter Estimates (cont.)

Item	Max Pts	*a*-par/ alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3
27	2	0.02078	12.9617	14.8353	
28	2	0.04484	31.2060	29.4265	
29	2	0.05067	32.7999	33.3174	
30	2	0.04002	26.1717	26.3083	
31	2	0.04425	30.5686	28.2938	
32	3	0.03256	21.2440	21.1357	22.7689
33	3	0.03746	25.2418	26.1799	24.6027
34	3	0.02571	15.1527	15.4304	18.1993
35	3	0.04577	29.1111	28.8615	29.4702

Table 28. Grade 7 2009 Operational Item Parameter Estimates

Item	Max Pts	*a*-par/ alpha	*b*-par/ gamma1	c-par/ gamma2	gamma3
1	1	0.02927	630.8110	0.2000	
2	1	0.03550	641.0360	0.2000	
3	1	0.04905	655.9160	0.1387	
4	1	0.03253	601.6734	0.2000	
5	1	0.02754	653.7222	0.2000	
6	1	0.01382	644.7371	0.2000	
7	1	0.02657	673.5528	0.1258	
8	1	0.03764	650.7351	0.2111	
9	1	0.02790	617.9601	0.2000	
10	1	0.02664	667.5739	0.2000	
11	1	0.02761	651.9597	0.1519	
12	1	0.03540	639.1725	0.1707	
13	1	0.02463	689.6318	0.2000	
14	1	0.00820	665.4420	0.2000	
15	1	0.01437	588.4884	0.2000	
16	1	0.04435	653.5253	0.2685	
17	1	0.02171	616.3049	0.2000	
18	1	0.04598	654.2716	0.2110	
19	1	0.05194	677.7239	0.2957	
20	1	0.02217	684.1264	0.1519	
21	1	0.03284	643.1461	0.1707	
22	1	0.04784	653.4097	0.1525	
23	1	0.04531	659.2853	0.2035	
24	1	0.03747	684.3094	0.1309	
25	1	0.03612	684.2833	0.1506	
26	1	0.02960	663.2764	0.2000	

(Continued on next page)

Table 28. Grade 7 2009 Operational Item Parameter Estimates (cont.)

Item	Max Pts	*a*-par/ alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3
27	1	0.05188	665.4622	0.1707	
28	1	0.03762	672.4420	0.2000	
29	1	0.04948	673.6277	0.2009	
30	1	0.02071	620.4391	0.2000	
31	2	0.04508	29.8604	28.5992	
32	2	0.04410	29.7250	30.1448	
33	2	0.04997	32.7587	32.9963	
34	2	0.04730	32.0230	32.5844	
35	3	0.03737	24.7498	25.1987	24.3263
36	3	0.04581	31.5433	31.4793	31.5061
37	3	0.03604	25.5130	23.3855	23.1258
38	3	0.04028	27.4550	25.3220	26.6169

Table 29. Grade 8 2009 Operational Item Parameter Estimates

Item	Max Pts	*a*-par/ alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3
1	1	0.02009	603.4050	0.2000	
2	1	0.03392	630.6582	0.2000	
3	1	0.03092	657.0308	0.1674	
4	1	0.02519	643.4673	0.2000	
5	1	0.03652	635.6517	0.2000	
6	1	0.04547	665.6846	0.1919	
7	1	0.03979	667.0246	0.1939	
8	1	0.05163	659.9311	0.1775	
9	1	0.04156	661.1751	0.3079	
10	1	0.04801	660.7487	0.1186	
11	1	0.03448	628.7515	0.1727	
12	1	0.02117	671.0472	0.1526	
13	1	0.03733	637.9860	0.1667	
14	1	0.04250	667.6299	0.2062	
15	1	0.02034	638.2225	0.2000	
16	1	0.04660	666.9648	0.1261	
17	1	0.04405	639.2159	0.2747	
18	1	0.02857	623.8598	0.2000	
19	1	0.02649	650.0150	0.2000	
20	1	0.03915	654.4691	0.4698	
21	1	0.04580	652.2768	0.1622	
22	1	0.03784	619.3267	0.2000	

(Continued on next page)

Table 29. Grade 8 2009 Operational Item Parameter Estimates (cont.)

Item	Max Pts	*a*-par/ alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3
23	1	0.04695	696.3817	0.2000	
24	1	0.04095	637.6584	0.2653	
25	1	0.04266	629.4398	0.3318	
26	1	0.02135	630.9284	0.2000	
27	1	0.03587	655.0078	0.1782	
28	2	0.05044	33.2457	30.7719	
29	2	0.04278	27.3930	27.6260	
30	2	0.05273	33.8504	36.4056	
31	2	0.04385	29.1647	28.9035	
32	3	0.05505	36.9609	36.4942	36.0003
33	3	0.03657	23.6189	23.4180	24.1418
34	2	0.04096	25.8968	26.2990	
35	2	0.04156	26.1105	27.4419	
36	2	0.05375	34.9687	34.7333	
37	2	0.03192	20.4445	21.1491	
38	2	0.04317	29.1077	28.8671	
39	2	0.04407	28.8465	29.4552	
40	2	0.03520	22.7179	24.0670	
41	2	0.03754	23.8488	25.2818	
42	3	0.03573	23.2484	23.1355	23.3255
43	3	0.03582	22.5440	22.8205	25.6314
44	3	0.03923	25.6214	26.6137	25.4846
45	3	0.05850	38.9860	39.4657	38.3773

Test Characteristic Curves

Test Characteristic Curves (TCCs) provide an overview of the test in IRT scale score metric. The 2008 and 2009 TCCs were generated using final OP item parameters. TCCs are the summation of all the Item Characteristic Curves (ICCs) for items which 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 blue TCCs and SE curves represent 2008 OP test and pink TCCs and SE curves represent 2009 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 Mathematics 2008 and 2009 OP TCCs and SE

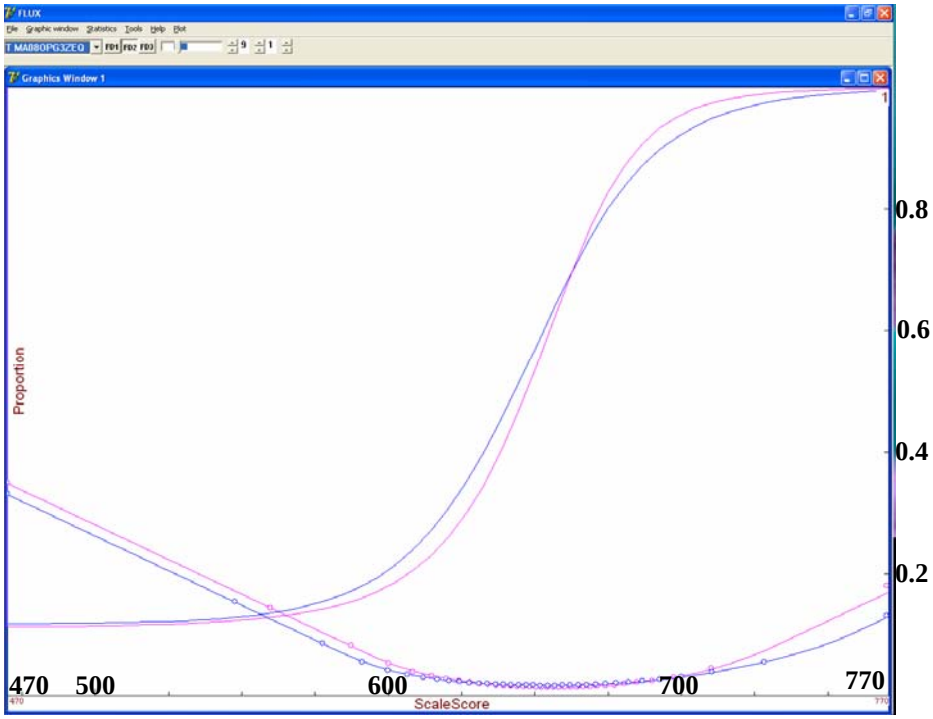


Figure 8. Grade 4 Mathematics 2008 and 2009 OP TCCs and SE

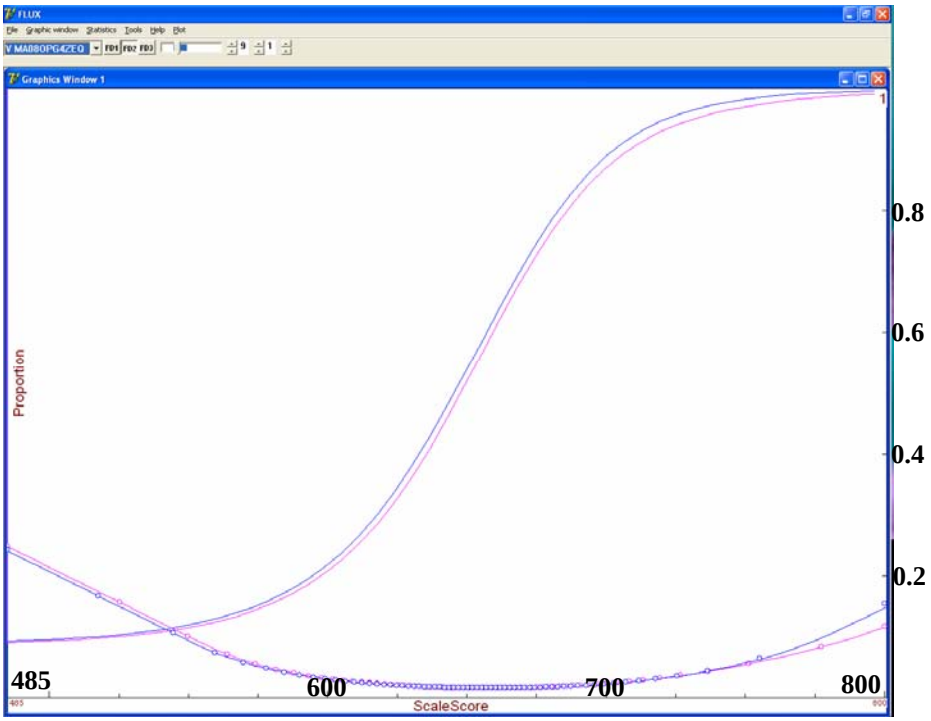


Figure 9. Grade 5 Mathematics 2008 and 2009 OP TCCs and SE

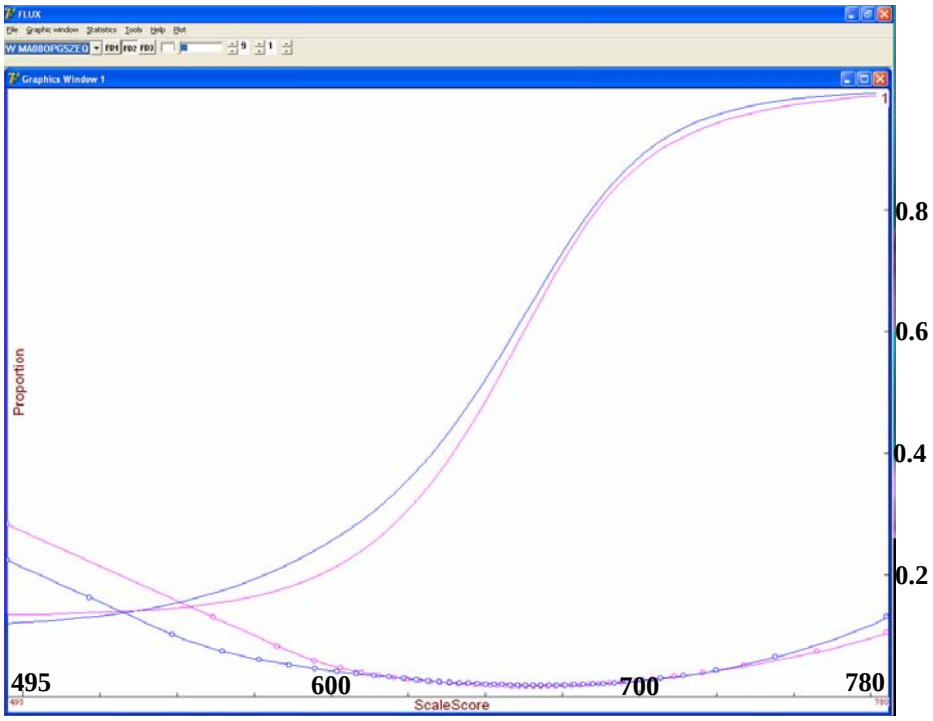


Figure 10. Grade 6 Mathematics 2008 and 2009 OP TCCs and SE

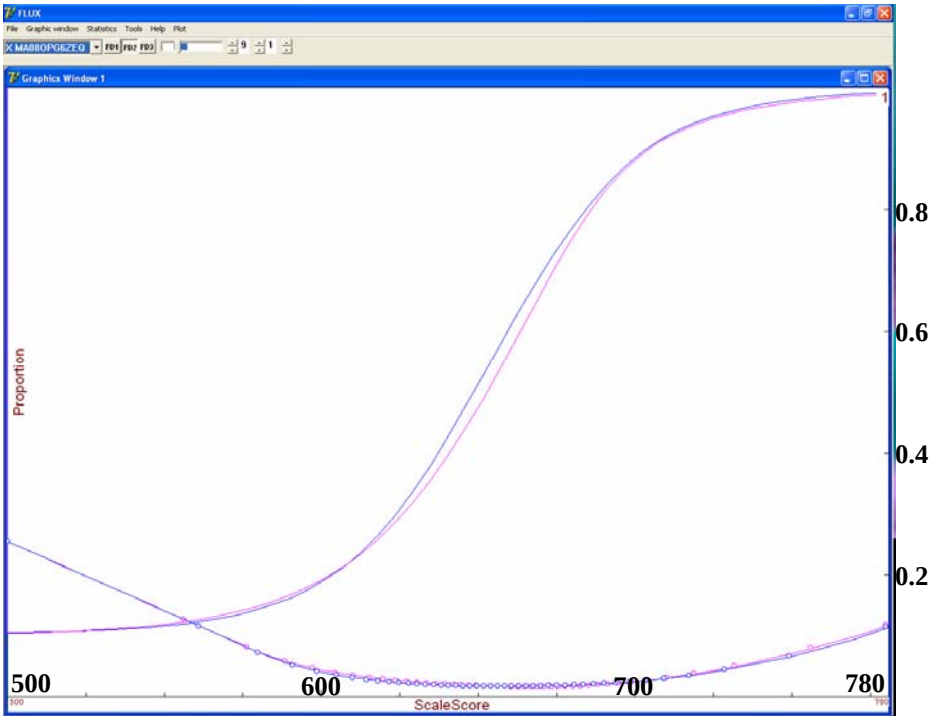


Figure 11. Grade 7 Mathematics 2008 and 2009 OP TCCs and SE

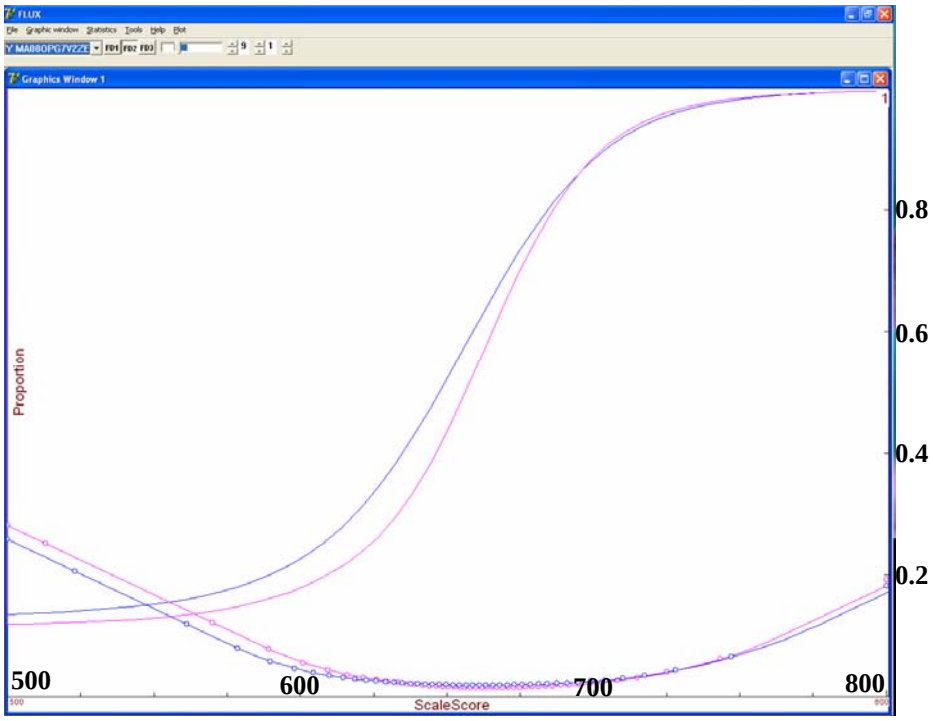
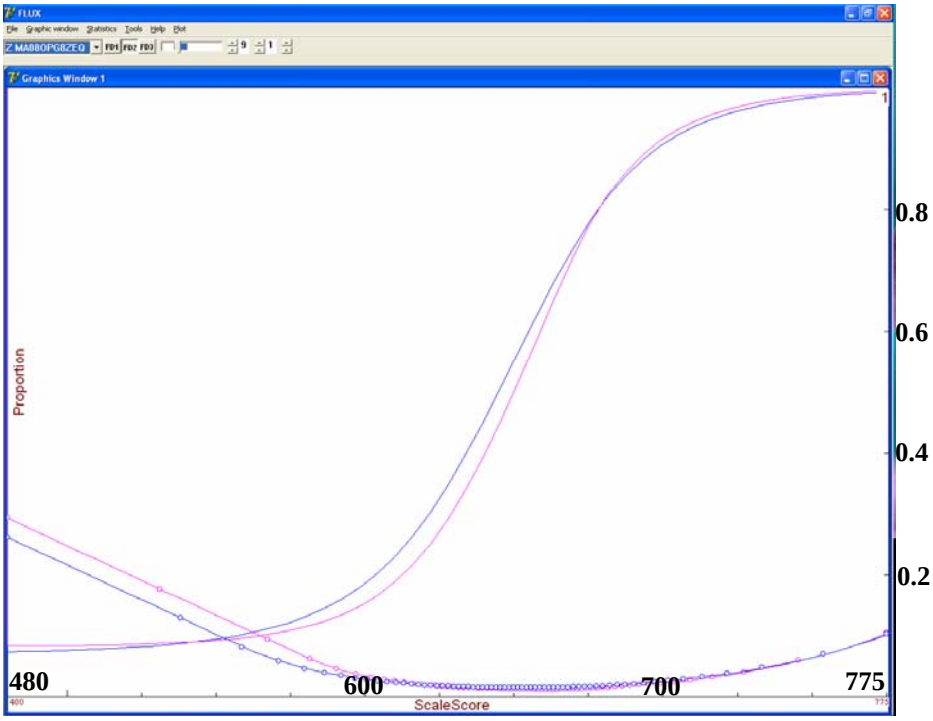


Figure 12. Grade 8 Mathematics 2008 and 2009 OP TCCs and SE



As seen in Figures 8, 10, and 12, very good alignments of the 2008 and 2009 TCCs and SE curves were found for Grades 4, 6 and 8. The TCCs for Grades 3, 5, and 7 were somewhat less well aligned at the lower end of the scale, indicating that the 2009 form tended to be slightly more difficult for lower-ability students. 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 score. That is, two students with the same number of score points on the test will receive the same 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 scale score metric were used to produce raw score-to-scale score conversion tables for the Grades 3–8 Mathematics Tests. An inverse TCC method was employed. The scoring tables were created using CTB/McGraw-Hill's 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 (Yen, 1984). The New York State Mathematics Tests have a maximum raw score ranging from 39 points (Grade 3) to 70 points (Grade 4). 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 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.

Raw Score-to-Scale Score and SEM Conversion Tables

The scale score 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 the standard-based performance index scores (SPIs). Scores on the NYSTP examinations are determined using number correct scoring. raw score-to-scale score conversion tables are presented in this section. The lowest and highest obtainable 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 is inversely related to the amount of information provided by the test at each ability level. The SE is estimated as follows:

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

where

$SE\left(\hat{\theta}\right)$ is the standard error of the scale score (theta), and
$I(\theta)$ is the amount of information provided by the test at a given ability level.

It should be noted that the information is estimated based on thetas in scale score metric; therefore, the SE is also expressed in 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 30. Grade 3 Raw Score-to-Scale Score (with Standard Error)

Raw Score	Scale Score	Standard Error
0	470	153
1	470	153
2	470	153
3	470	153
4	470	153
5	560	63
6	587	36
7	600	23
8	608	17
9	615	14
10	620	12
11	624	11
12	628	10
13	631	9
14	634	8
15	637	8

(Continued on next page)

Table 30. Grade 3 Raw Score-to-Scale Score (with Standard Error) (cont.)

Raw Score	Scale Score	Standard Error
16	639	7
17	642	7
18	644	7
19	646	6
20	648	6
21	650	6
22	652	6
23	654	6
24	656	6
25	658	6
26	660	6
27	662	6
28	664	6
29	667	6
30	669	7
31	671	7
32	674	7
33	677	8
34	681	8
35	685	9
36	690	11
37	697	13
38	710	19
39	770	79

Table 31. Grade 4 Raw Score-to-Scale Score (with Standard Error)

Raw Score	Scale Score	Standard Error
0	485	109
1	485	109
2	485	109
3	485	109
4	485	109
5	485	109
6	485	109
7	525	69
8	550	44
9	564	31
10	574	24
11	582	20
12	588	18

(Continued on next page)

Table 31. Grade 4 Raw Score-to-Scale Score (with Standard Error) (cont.)

Raw Score	Scale Score	Standard Error
13	593	16
14	598	14
15	602	13
16	606	12
17	609	11
18	612	11
19	615	10
20	618	10
21	621	9
22	623	9
23	625	9
24	627	9
25	630	8
26	632	8
27	634	8
28	636	8
29	637	8
30	639	7
31	641	7
32	643	7
33	644	7
34	646	7
35	648	7
36	649	7
37	651	7
38	653	7
39	654	7
40	656	7
41	658	7
42	659	7
43	661	7
44	663	7
45	664	7
46	666	7
47	668	7
48	669	7
49	671	7
50	673	7
51	675	7
52	677	7
53	679	8

(Continued on next page)

Table 31. Grade 4 Raw Score-to-Scale Score (with Standard Error) (cont.)

Raw Score	Scale Score	Standard Error
54	681	8
55	683	8
56	686	8
57	688	8
58	691	9
59	694	9
60	697	9
61	700	10
62	704	11
63	708	11
64	713	12
65	719	14
66	727	16
67	737	19
68	751	24
69	777	37
70	800	51

Table 32. Grade 5 Raw Score-to-Scale Score (with Standard Error)

Raw Score	Scale Score	Standard Error
0	495	124
1	495	124
2	495	124
3	495	124
4	495	124
5	495	124
6	495	124
7	562	57
8	583	36
9	595	26
10	603	20
11	610	17
12	615	15
13	620	14
14	625	12
15	629	11
16	632	11
17	635	10

(Continued on next page)

Table 32. Grade 5 Raw Score-to-Scale Score (with Standard Error) (cont.)

Raw Score	Scale Score	Standard Error
18	638	10
19	641	9
20	644	9
21	647	9
22	649	8
23	652	8
24	654	8
25	656	8
26	659	8
27	661	8
28	663	8
29	666	8
30	668	8
31	670	8
32	673	8
33	675	8
34	678	8
35	681	8
36	684	8
37	687	9
38	690	9
39	694	10
40	699	11
41	704	12
42	711	14
43	720	17
44	734	22
45	758	32
46	780	46

Table 33. Grade 6 Raw Score-to-Scale Score (with Standard Error)

Raw Score	Scale Score	Standard Error
0	500	112
1	500	112
2	500	112
3	500	112
4	500	112
5	500	112
6	556	56
7	576	35
8	588	25
9	597	20
10	604	17
11	610	15
12	615	14
13	620	12
14	624	11
15	627	11
16	630	10
17	634	10
18	637	9
19	639	9
20	642	9
21	645	8
22	647	8
23	649	8
24	652	8
25	654	8
26	656	7
27	658	7
28	660	7
29	663	7
30	665	7
31	667	7
32	669	7
33	671	7
34	673	7
35	675	7
36	678	7
37	680	7
38	683	8
39	685	8

(Continued on next page)

Table 33. Grade 6 Raw Score-to-Scale Score (with Standard Error) (cont.)

Raw Score	Scale Score	Standard Error
40	688	8
41	691	9
42	695	10
43	699	10
44	704	12
45	710	13
46	719	16
47	731	22
48	756	35
49	780	51

Table 34. Grade 7 Raw Score-to-Scale Score (with Standard Error)

Raw Score	Scale Score	Standard Error
0	500	123
1	500	123
2	500	123
3	500	123
4	500	123
5	500	123
6	513	110
7	570	53
8	589	34
9	601	24
10	609	19
11	616	16
12	621	13
13	626	12
14	630	11
15	633	10
16	636	9
17	639	9
18	642	8
19	644	8
20	646	7
21	649	7
22	651	7
23	653	7

(Continued on next page)

Table 34. Grade 7 Raw Score-to-Scale Score (with Standard Error) (cont.)

Raw Score	Scale Score	Standard Error
24	654	7
25	656	7
26	658	7
27	660	6
28	662	6
29	664	6
30	665	6
31	667	6
32	669	6
33	671	7
34	673	7
35	675	7
36	677	7
37	679	7
38	681	7
39	684	7
40	686	8
41	689	8
42	692	8
43	695	9
44	699	9
45	703	10
46	708	12
47	715	14
48	725	17
49	743	27
50	800	84

Table 35. Grade 8 Raw Score-to-Scale Score (with Standard Error)

Raw Score	Scale Score	Standard Error
0	480	128
1	480	128
2	480	128
3	480	128
4	480	128
5	480	128

(Continued on next page)

Table 35. Grade 8 Raw Score-to-Scale Score (with Standard Error) (cont.)

Raw Score	Scale Score	Standard Error
6	531	77
7	567	41
8	581	27
9	590	20
10	597	16
11	602	14
12	606	12
13	610	11
14	613	10
15	616	9
16	619	9
17	621	8
18	624	8
19	626	7
20	628	7
21	630	7
22	631	7
23	633	7
24	635	6
25	636	6
26	638	6
27	640	6
28	641	6
29	642	6
30	644	6
31	645	6
32	647	6
33	648	6
34	649	5
35	651	5
36	652	5
37	653	5
38	655	5
39	656	5
40	657	5
41	658	5
42	660	5
43	661	5
44	662	5
45	663	5
46	665	5
47	666	5

(Continued on next page)

Table 35. Grade 8 Raw Score-to-Scale Score (with Standard Error) (cont.)

Raw Score	Scale Score	Standard Error
48	667	5
49	669	5
50	670	5
51	671	6
52	673	6
53	674	6
54	676	6
55	678	6
56	680	6
57	682	7
58	684	7
59	686	7
60	689	8
61	692	8
62	695	9
63	699	10
64	704	11
65	709	12
66	717	14
67	727	18
68	745	26
69	775	46

Standard Performance Index

The standard performance index (SPI) reported for each objective measured by the Grades 3–8 Mathematics 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 New York State Mathematics 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 36 presents SPI target ranges. The objectives in this table are denoted as follows: 1—Number Sense and Operations, 2—Algebra, 3—Geometry, 4—Measurement, and 5—Statistics and Probability.

Table 36. SPI Target Ranges

Grade	Objective	# Items	Total Points	Level III Cut SPI Target Range
3	1	14	16	42–58
	2	4	6	32–51
	3	4	5	69–77
	4	5	5	56–70
	5	4	7	44–59
4	1	24	33	44–57
	2	7	11	55–66
	3	6	9	58–67
	4	7	10	33–45
	5	4	7	46–55
5	1	12	15	40–53
	2	4	5	54–68
	3	10	15	40–54
	4	4	5	43–60
	5	4	6	39–52
6	1	12	18	33–45
	2	7	10	50–69
	3	6	7	37–49
	4	4	5	46–60
	5	6	9	48–61
7	1	10	13	37–50
	2	6	8	27–42
	3	5	7	42–55
	4	7	9	31–39
	5	10	13	45–62
8	1	4	7	33–44
	2	17	28	35–46
	3	20	29	53–64
	4	4	5	68–75

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 mathematics 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 Number Sense, but has a low level of knowledge in Algebra, 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 (strand) 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 Mathematics 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 given by

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

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

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

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), Spanish language test version and English language learners. Most of the items flagged by IRT DIF were items from the Spanish language versions of the test. Also, as indicated in the classical DIF analysis section, items flagged for DIF do not necessarily display bias. Applying the Linn-Harnisch method revealed that no item was flagged for DIF on the Grade 3 test; three items were flagged on the Grade 4 test; four items were flagged on the Grade 5 test; two items were flagged on the Grade 6 test; seven items were flagged on the Grade 7 test; and five items were flagged on the Grade 8 test, as is shown in Table 37.

Table 37. Number of Items Flagged for DIF by the Linn-Harnisch Method

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

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

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 Mathematics 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). Please note that the reliability statistics in Section V, "Operational Test Data Collections and Classical Analysis," are based upon the classical analysis and calibration sample, whereas the statistics in this section are based on the total student population data.

Reliability for Total Test

The overall test reliability is a very good indication of each test's internal consistency. Included in Table 38 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 mathematics tests.

Table 38. Reliability and Standard Error of Measurement

Grade	N-count	# Items	# RS Points	Cronbach's Alpha	SEM of Cronbach	Feldt-Raju	SEM of Feldt-Raju
3	200058	31	39	0.88	2.25	0.90	2.10
4	197379	48	70	0.94	3.42	0.94	3.22
5	199180	34	46	0.90	2.80	0.91	2.64
6	199605	35	49	0.91	3.07	0.92	2.86
7	204292	38	50	0.90	3.28	0.92	2.98
8	208835	45	69	0.94	3.72	0.95	3.37

All the coefficients for total test reliability were in the range of 0.88–0.94, which indicated high internal consistency. As expected, the lowest reliabilities were found for shortest tests (Grades 3, 5, 6, and 7) and the highest reliabilities are associated with the longer tests (Grades 4 and 8).

Reliability for 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 estimated for the overall test form. Table 39 presents reliabilities for the MC subsets.

Table 39. 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	200058	25	0.86	1.47	0.87	1.45
4	197379	30	0.86	1.95	0.87	1.93
5	199180	26	0.86	1.78	0.87	1.75
6	199605	25	0.87	1.91	0.87	1.89
7	204292	30	0.84	2.06	0.85	2.03
8	208835	27	0.88	1.84	0.89	1.82

Reliability for CR Items

Reliability coefficients were also computed for the subsets of CR items. It should be noted that the Grades 3–8 Mathematics Tests include 6–18 CR items depending on grade level. The results are presented in Table 40.

Table 40. 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	200058	6	14	0.74	1.56	0.76	1.51
4	197379	18	40	0.91	2.61	0.91	2.56
5	199180	8	20	0.79	2.00	0.80	1.97
6	199605	10	24	0.84	2.23	0.85	2.15
7	204292	8	20	0.84	2.23	0.85	2.17
8	208835	18	41	0.93	2.95	0.93	2.82

Note: Results should be interpreted with caution for Grades 3, 5, 6, and 7 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 learner (ELL) status, 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, Bilingual Dictionaries and Glossaries, Translated Edition, Oral Translation, and Responses Written in Native Language. In addition, reliability coefficients were computed for the following subgroups of English language learners:

students taking the English version of the mathematics test and students taking the mathematics tests in each of the five translated languages (Chinese, Haitian Creole, Korean, Russian, and Spanish). As shown in Tables 41a–41f, the estimated reliabilities for subgroups were close in magnitude to the test reliability estimates of the population. Cronbach's alpha reliability coefficients across subgroups were equal to or greater than 0.81, with the exception of Grade 5 Korean subgroups. Feldt-Raju reliability coefficients, which tend to be larger than the Cronbach's alpha estimates for the same group, were all larger than 0.84, with the exception of Grade 5 Korean subgroup. Overall, the New York State Mathematics Tests were found to have very good test internal consistency (reliability) for analyzed subgroups of examinees.

Table 41a. 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	200058	0.88	2.25	0.90	2.10
Gender	Female	97579	0.88	2.24	0.89	2.09
	Male	102479	0.89	2.26	0.90	2.11
Ethnicity	Asian	16338	0.87	1.78	0.88	1.67
	Black	37659	0.89	2.53	0.90	2.37
	Hispanic	43140	0.88	2.44	0.90	2.28
	American Indian	946	0.88	2.47	0.89	2.31
	Multi-Racial	676	0.86	2.23	0.87	2.10
	White	101200	0.87	2.09	0.88	1.96
	Unknown	99	0.85	2.04	0.87	1.91
NRC	New York City	71498	0.89	2.31	0.91	2.15
	Big 4 Cites	8334	0.89	2.70	0.91	2.54
	High Needs Urban/Suburban	16279	0.88	2.44	0.90	2.27
	High Needs Rural	11509	0.87	2.40	0.88	2.25
	Average Needs	58657	0.86	2.16	0.88	2.03
	Low Needs	29687	0.83	1.86	0.85	1.77
	Charter	3476	0.84	2.21	0.85	2.10
SWD	All Codes	27469	0.90	2.75	0.91	2.57
SUA	All Codes	47490	0.90	2.64	0.91	2.46
SWD/ SUA	SUA=504 Plan Codes	23703	0.90	2.79	0.91	2.60
ELL/ SUA	SUA=ELL Codes	17410	0.89	2.60	0.90	2.43
ELL	English	16319	0.89	2.59	0.90	2.42
	Chinese	358	0.87	2.19	0.89	2.05
	Haitian Creole	58	0.90	2.95	0.91	2.74
	Korean	66	0.85	1.82	0.86	1.74
	Russian	58	0.93	2.48	0.94	2.26
	Spanish	3505	0.90	2.68	0.91	2.50
	All Translations	4045	0.90	2.65	0.91	2.46

Table 41b. 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	197379	0.94	3.42	0.94	3.22
Gender	Female	96017	0.93	3.43	0.94	3.24
	Male	101362	0.94	3.40	0.95	3.20
Ethnicity	Asian	15073	0.92	2.79	0.93	2.64
	Black	37635	0.94	3.73	0.94	3.52
	Hispanic	42368	0.94	3.64	0.94	3.44
	American Indian	913	0.93	3.65	0.94	3.46
	Multi-Racial	511	0.93	3.45	0.94	3.25
	White	100795	0.93	3.24	0.93	3.07
	Unknown	84	0.94	3.20	0.95	2.99
NRC	New York City	69845	0.94	3.48	0.95	3.26
	Big 4 Cites	8068	0.94	3.90	0.95	3.67
	High Needs Urban/Suburban	15914	0.93	3.65	0.94	3.45
	High Needs Rural	11454	0.93	3.62	0.93	3.43
	Average Needs	59087	0.93	3.33	0.93	3.16
	Low Needs	29501	0.91	2.95	0.92	2.82
	Charter	2898	0.92	3.47	0.92	3.32
SWD	All Codes	29088	0.94	3.94	0.95	3.69
SUA	All Codes	47345	0.94	3.87	0.95	3.62
SWD/ SUA	SUA=504 Plan Codes	26147	0.94	3.97	0.95	3.71
ELL/ SUA	SUA=ELL Codes	14252	0.93	3.88	0.94	3.66
ELL	English	13521	0.93	3.86	0.94	3.65
	Chinese	356	0.91	3.09	0.92	2.94
	Haitian Creole	82	0.92	4.08	0.93	3.87
	Korean	55	0.92	2.54	0.93	2.36
	Russian	65	0.95	3.76	0.96	3.45
	Spanish	2969	0.94	4.00	0.94	3.77
	All Translations	3527	0.94	3.92	0.95	3.68

Table 41c. 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	199180	0.90	2.80	0.91	2.64
Gender	Female	97620	0.89	2.81	0.90	2.65
	Male	101560	0.90	2.80	0.91	2.63
Ethnicity	Asian	15058	0.89	2.31	0.90	2.18
	Black	38167	0.89	3.03	0.90	2.88
	Hispanic	42248	0.89	2.94	0.90	2.79
	American Indian	955	0.89	2.95	0.90	2.80
	Multi-Racial	520	0.90	2.90	0.91	2.76
	White	102129	0.88	2.68	0.89	2.53
	Unknown	103	0.89	2.59	0.91	2.38
NRC	New York City	69778	0.90	2.83	0.92	2.66
	Big 4 Cites	7667	0.90	3.18	0.91	3.01
	High Needs Urban/Suburban	15618	0.89	2.96	0.90	2.81
	High Needs Rural	11448	0.88	2.97	0.89	2.82
	Average Needs	60097	0.88	2.75	0.89	2.60
	Low Needs	30298	0.86	2.45	0.87	2.33
	Charter	3556	0.87	2.86	0.88	2.74
SWD	All Codes	30811	0.90	3.15	0.91	3.00
SUA	All Codes	47992	0.90	3.10	0.91	2.94
SWD/ SUA	SUA=504 Plan Codes	28263	0.89	3.17	0.90	3.01
ELL/ SUA	SUA=ELL Codes	12149	0.90	3.12	0.91	2.95
ELL	English	11364	0.90	3.10	0.91	2.95
	Chinese	387	0.87	2.44	0.88	2.33
	Haitian Creole	70	0.89	3.15	0.90	3.01
	Korean	53	0.69	1.89	0.71	1.81
	Russian	59	0.91	3.06	0.92	2.87
	Spanish	2864	0.89	3.19	0.90	3.04
	All Translations	3433	0.91	3.13	0.92	2.96

Table 41d. 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	199605	0.91	3.07	0.92	2.86
Gender	Female	97359	0.91	3.06	0.92	2.87
	Male	102246	0.92	3.06	0.93	2.85
Ethnicity	Asian	15322	0.91	2.58	0.92	2.41
	Black	38288	0.91	3.28	0.92	3.08
	Hispanic	41482	0.91	3.23	0.92	3.03
	American Indian	903	0.90	3.24	0.91	3.05
	Multi-Racial	463	0.90	3.10	0.91	2.91
	White	103072	0.90	2.93	0.91	2.76
	Unknown	75	0.91	2.80	0.92	2.55
NRC	New York City	69450	0.92	3.13	0.93	2.91
	Big 4 Cites	7653	0.90	3.38	0.91	3.19
	High Needs Urban/Suburban	15230	0.91	3.20	0.92	3.02
	High Needs Rural	11407	0.89	3.14	0.91	2.97
	Average Needs	61209	0.90	3.00	0.91	2.82
	Low Needs	30691	0.88	2.73	0.90	2.58
	Charter	3196	0.89	3.05	0.90	2.87
SWD	All Codes	30554	0.90	3.33	0.91	3.16
SUA	All Codes	43839	0.91	3.33	0.92	3.14
SWD/ SUA	SUA=504 Plan Codes	27567	0.90	3.34	0.91	3.16
ELL/ SUA	SUA=ELL Codes	9936	0.91	3.37	0.92	3.17
ELL	English	9293	0.90	3.35	0.91	3.17
	Chinese	486	0.89	2.87	0.90	2.69
	Haitian Creole	115	0.91	3.39	0.92	3.17
	Korean	68	0.91	2.64	0.92	2.44
	Russian	61	0.94	3.24	0.95	2.98
	Spanish	3057	0.90	3.39	0.92	3.19
	All Translations	3787	0.92	3.35	0.93	3.13

Table 41e. 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	204292	0.90	3.28	0.92	2.98
Gender	Female	99379	0.89	3.27	0.91	2.98
	Male	104913	0.90	3.27	0.92	2.97
Ethnicity	Asian	15496	0.90	2.86	0.92	2.58
	Black	38510	0.88	3.49	0.90	3.23
	Hispanic	42151	0.89	3.48	0.90	3.21
	American Indian	942	0.89	3.40	0.90	3.15
	Multi-Racial	420	0.91	3.31	0.92	3.01
	White	106690	0.88	3.05	0.90	2.82
	Unknown	83	0.89	3.16	0.91	2.88
NRC	New York City	71034	0.90	3.42	0.92	3.09
	Big 4 Cites	7830	0.89	3.51	0.90	3.27
	High Needs Urban/Suburban	15511	0.88	3.42	0.90	3.17
	High Needs Rural	12047	0.87	3.29	0.89	3.06
	Average Needs	63079	0.87	3.10	0.89	2.88
	Low Needs	31366	0.86	2.82	0.88	2.63
	Charter	2472	0.87	3.35	0.89	3.12
SWD	All Codes	31204	0.88	3.50	0.89	3.28
SUA	All Codes	42979	0.89	3.52	0.90	3.28
SWD/ SUA	SUA=504 Plan Codes	27928	0.88	3.51	0.89	3.28
ELL/ SUA	SUA=ELL Codes	9567	0.88	3.53	0.90	3.29
ELL	English	8163	0.88	3.51	0.89	3.30
	Chinese	579	0.88	3.20	0.90	2.91
	Haitian Creole	148	0.88	3.48	0.89	3.30
	Korean	98	0.81	2.77	0.84	2.58
	Russian	77	0.89	3.50	0.91	3.21
	Spanish	3375	0.87	3.50	0.89	3.27
	All Translations	4277	0.90	3.52	0.91	3.23

Table 41f. 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	208835	0.94	3.72	0.95	3.37
Gender	Female	102402	0.94	3.69	0.95	3.35
	Male	106433	0.95	3.74	0.96	3.37
Ethnicity	Asian	15613	0.94	3.11	0.95	2.81
	Black	39686	0.94	4.01	0.95	3.69
	Hispanic	42887	0.94	3.96	0.95	3.63
	American Indian	1015	0.94	3.91	0.95	3.57
	Multi-Racial	328	0.94	3.84	0.95	3.49
	White	109241	0.93	3.47	0.94	3.19
	Unknown	65	0.95	3.43	0.96	3.16
NRC	New York City	73113	0.95	3.89	0.96	3.51
	Big 4 Cites	7793	0.94	4.06	0.95	3.75
	High Needs Urban/Suburban	15711	0.94	3.91	0.95	3.59
	High Needs Rural	12279	0.93	3.78	0.94	3.49
	Average Needs	65121	0.93	3.50	0.94	3.24
	Low Needs	31562	0.92	3.11	0.93	2.91
	Charter	2162	0.92	3.72	0.93	3.47
SWD	All Codes	30515	0.94	4.04	0.95	3.74
SUA	All Codes	42818	0.94	4.04	0.95	3.72
SWD/ SUA	SUA=504 Plan Codes	27455	0.94	4.06	0.95	3.75
ELL/ SUA	SUA=ELL Codes	9792	0.94	4.05	0.95	3.71
ELL	English	8326	0.94	4.04	0.95	3.74
	Chinese	533	0.94	3.33	0.95	2.98
	Haitian Creole	161	0.94	4.02	0.95	3.72
	Korean	80	0.89	2.70	0.91	2.47
	Russian	75	0.95	3.92	0.96	3.50
	Spanish	3228	0.94	4.06	0.95	3.75
	All Translations	4077	0.95	4.01	0.96	3.65

Standard Error of Measurement

The standard error of measurements (SEMs), as computed from Cronbach's alpha and the Feldt-Raju reliability statistics, are presented in Table 38. SEMs based on Cronbach's alpha ranged from 2.25–3.72, which is reasonably small given the maximum number of score points on mathematics tests. 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 41a–41f. The SEMs associated with all reliability estimates for all subpopulations are in the range of 1.67–4.08, which is acceptably close to those for the entire population. This narrow range indicates that across the Grades 3–8 Mathematics 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 Mathematics 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 H.

Classification consistency and accuracy were estimated using the IRT procedure suggested by Lee, Hanson, and Brennan (2002) and Wang, Kolen, and Harris (2000) and implemented by CTB/McGraw-Hill proprietary software WLCLASS (Kim, 2004). Appendix G 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. Included in Tables 42 and 43 are 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 42 depicts the consistency study results based on the range of performance levels for all grades. Overall, between 78% and 83% 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 from 0.61–0.71.

Table 42. Decision Consistency (All Cuts)

Grade	N-count	Agreement	Inconsistency	Kappa
3	200058	0.7989	0.2011	0.6130
4	197379	0.8172	0.1828	0.6958
5	199180	0.7844	0.2156	0.6370
6	199605	0.7928	0.2072	0.6606
7	204292	0.8079	0.1921	0.6685
8	208835	0.8322	0.1678	0.7071

Table 43 depicts the consistency study results based on two performance levels (passing and not passing), as defined by the Level III cut. Overall, about 94%–97% of the classifications of individual students were estimated to remain stable with a second administration. Kappa coefficients for classification consistency based on one cut ranged from 0.74–0.83.

Table 43. Decision Consistency (Level III Cut)

Grade	N-count	Agreement	Inconsistency	Kappa
3	200058	0.9657	0.0343	0.7442
4	197379	0.9561	0.0439	0.8039
5	199180	0.9452	0.0548	0.7404
6	199605	0.9375	0.0625	0.7803
7	204292	0.9434	0.0566	0.7499
8	208835	0.9453	0.0547	0.8251

Accuracy

The results of classification accuracy are presented in Table 44. 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 83%–88% of the time across all performance levels and approximately 96%–98% of the time in regards to the Level III cut score.

Table 44. 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	200058	**0.8317**	0.1385	0.0298	**0.9761**	0.0098	0.0141
4	197379	**0.8685**	0.0771	0.0544	**0.9692**	0.0141	0.0167
5	199180	**0.8445**	0.0884	0.067	**0.9615**	0.0160	0.0224
6	199605	**0.8511**	0.0865	0.0623	**0.9561**	0.0212	0.0227
7	204292	**0.8579**	0.0932	0.0489	**0.9599**	0.0205	0.0196
8	208835	**0.8764**	0.0803	0.0432	**0.9618**	0.0181	0.0201

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 Mathematics Tests. These include the scale score means, standard deviations, and percentiles and performance level distributions for each grade's population and specific subgroups. Gender, ethnic identification, needs resource category, English language learners, students with disabilities, students using accommodation, and test language variables (Test Language) 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 H.

Scale Score Distribution Summary

Scale score distribution summary tables are presented and discussed in Table 45. First, scale score statistics for total populations of students from public and charter schools are presented. Next, scale score statistics are presented for selected subgroups in each grade level. The statistics for groups with small number counts should be interpreted with caution. Some general observations: Females and Males had very similar achievement patterns; Asian and White students outperformed their peers from other ethnic groups; Low Needs and Average Needs schools (as identified by NRC) outperformed other school types (New York City, Big 4 Cities, Urban/Suburban, Rural, and Charter); students taking the Chinese and Korean translations met or exceeded the population at every reported percentile, whereas the other translation subgroups (Haitian Creole, Spanish, and Russian) were below the population scale score at each percentile; English language learners, taking the mathematics test in English, SWD, and/or SUA status, achieved below the State aggregate (All Students) in every percentile. This pattern of achievement was consistent across all grades. Note that complete scale score frequency distribution tables for the total population of students are located in Appendix H.

Table 45. Mathematics Scale Score Distribution Summary Grades 3–8

Grade	N-count	SS Mean	SS Std Dev	10^{th} %tile	25^{th} %tile	50^{th} %tile	75^{th} %tile	90^{th} %tile
3	200058	692.06	37.02	654	669	685	710	770
4	197379	689.59	38.28	644	668	691	713	737
5	199180	686.32	33.80	647	666	687	704	720
6	199605	679.91	35.21	639	660	680	699	719
7	204292	680.84	32.27	646	662	679	699	715
8	208835	674.99	33.75	636	656	674	695	717

Grade 3

Scale score statistics and N-counts of demographic groups for Grade 3 are presented in Table 46. The population scale score mean was 692.06 with a standard deviation of 37.02. The gender subgroups performed very similarly, with a mean difference of less than two scale score points. Asian, Multi-Racial, and White ethnic subgroups had scale score means that exceeded the State mean scale score on the test, as did students from Low Needs and Average Needs districts. The lowest performing NRC subgroup was the Big 4 Cities, with a mean of 670.92, and the lowest performing ethnic subgroup was Black (mean scale score of 679.25). SWD, SUA, and ELL without testing in an alternate language subgroup scored consistently below the statewide percentile scale score rankings. At the 50th percentile, the scale scores on translated forms range from 652 (Haitian-Creole subgroup) to 697 (Korean subgroup), a difference that exceeds a standard deviation. The subgroup who used the Haitian-Creole translation had a scale score mean of 43 scale score units below the population mean, which was the lowest performing group analyzed. At the 50th percentile, the following groups exceeded the population scale score of 685: Asian (697), White (690), Average Needs (690), Low Needs (697), and students who used the Korean (697) translations.

Table 46. 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	200058	692.06	37.02	654	669	685	710	770
Gender	Female	97579	692.87	36.88	656	669	685	710	770
	Male	102479	691.30	37.15	654	669	685	710	770
Ethnicity	Asian	16338	710.72	39.71	669	685	697	770	770
	Black	37659	679.25	33.54	646	660	674	690	710
	Hispanic	43140	683.21	34.06	650	664	677	697	710
	American Indian	946	683.06	33.05	648	662	677	697	710
	Multi-Racial	676	693.09	35.31	656	671	685	710	770
	White	101200	697.66	36.60	662	674	690	710	770
	Unknown	99	703.78	40.22	664	677	690	710	770
NRC	New York City	71498	689.40	37.53	652	667	685	697	770
	Big 4 Cites	8334	670.92	32.27	639	652	669	685	710
	High Needs Urban/Suburban	16279	685.02	35.22	650	664	681	697	710
	High Needs Rural	11509	685.67	33.39	652	667	681	697	710
	Average Needs	58657	694.88	35.74	660	674	690	710	770
	Low Needs	29687	705.54	36.92	669	681	697	710	770
	Charter	3476	691.39	32.77	658	671	685	697	770
SWD	All Codes	27469	666.36	32.14	631	648	664	681	697
SUA	All Codes	47490	673.59	34.16	639	654	671	690	710
SWD/SUA	SUA=504 Plan Codes	23703	664.21	30.98	631	648	664	681	697
ELL/SUA	SUA=ELL Codes	17410	675.93	31.94	644	658	674	690	710

(Continued on next page)

Table 46. Scale Score Distribution Summary, by Subgroup, Grade 3 (cont.)

Demographic Category (Subgroup)		N-count	SS Mean	SS Std Dev	10th %tile	25th %tile	50th %tile	75th %tile	90th %tile
ELL	English	16319	676.08	32.18	644	658	674	690	710
	Chinese	358	693.38	35.66	658	671	685	710	770
	Haitian Creole	58	648.93	33.61	624	634	652	667	685
	Korean	66	707.58	37.63	671	681	697	710	770
	Russian	58	676.22	43.63	634	656	679	697	710
	Spanish	3505	671.41	33.80	637	654	669	685	710
	All Translations	4045	673.69	35.13	639	656	671	690	710

Grade 4

Scale score statistics and N-counts of demographic groups for Grade 4 are presented in Table 47. The population scale score mean was 689.59 with a standard deviation of 38.28. The gender subgroups performed very similarly, with a mean difference of less than one scale score point. Asian and White students' scale score means exceeded the State mean scale score on the test. Asian students (the highest performing ethnic subgroup) exceeded the State mean by more than one-half of a standard deviation. Black, Hispanics, and American Indian ethnic subgroups had mean scale scores almost one standard deviation below the Asian subgroup. Students from Low Needs and Average Needs districts outperformed the other NRC subgroups. The lowest performing NRC subgroup was the Big 4 Cities, with a mean of 665.15, well more than one-half of a standard deviation below the State mean. SWD, SUA, and ELL without testing in an alternate language subgroup scored consistently below the Statewide percentile scale score rankings. The Haitian-Creole translation subgroup had means over one standard deviation below the population and was the lowest performing group analyzed. ELL who took the mathematics test in English outperformed the total group of students who took translated forms in terms of test mean and reported percentile scores except for Chinese, Korean, and Russian translation subgroups. At the 50th percentile, the following groups exceeded the population scale score of 691: Asian (713), White (694), Low Needs (704), and students who used the Chinese (700) and Korean (713) translations.

Table 47. 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	197379	689.59	38.28	644	668	691	713	737
Gender	Female	96017	689.76	37.75	646	666	688	713	737
	Male	101362	689.42	38.78	644	668	691	713	737
Ethnicity	Asian	15073	714.16	38.47	669	691	713	737	777
	Black	37635	675.45	38.20	632	653	675	697	719
	Hispanic	42368	680.33	37.45	636	659	681	704	727
	American Indian	913	679.50	36.22	637	659	679	700	727
	Multi-Racial	511	688.95	36.69	646	668	688	713	737
	White	100795	695.17	35.52	654	675	694	713	737
	Unknown	84	695.51	40.36	653	675	700	719	737
NRC	New York City	69845	688.34	41.41	641	663	688	713	737
	Big 4 Cites	8068	665.15	38.66	618	643	666	688	713
	High Needs Urban/Suburban	15914	679.97	36.34	637	659	681	700	727
	High Needs Rural	11454	680.28	34.31	641	661	681	700	719
	Average Needs	59087	691.51	34.28	651	671	691	713	737
	Low Needs	29501	704.81	33.64	666	683	704	727	751
	Charter	2898	688.56	33.27	649	668	686	708	727
SWD	All Codes	29088	656.83	40.55	606	636	659	681	704
SUA	All Codes	47345	665.43	40.31	615	643	668	691	713
SWD/SUA	SUA=504 Plan Codes	26147	655.03	40.15	606	634	658	681	700
ELL/SUA	SUA=ELL Codes	14252	668.57	36.41	625	648	669	691	713
ELL	English	13521	668.52	36.18	625	648	669	691	713
	Chinese	356	702.78	34.38	661	680	700	719	751
	Haitian Creole	82	648.88	32.87	606	632	650	669	683
	Korean	55	715.20	29.73	683	700	713	737	751
	Russian	65	671.55	45.82	623	643	673	700	727
	Spanish	2969	659.31	38.60	612	639	661	683	704
	All Translations	3527	664.55	40.83	615	641	666	688	713

Grade 5

Grade 5 demographic groups N-counts and scale score statistics are presented in Table 48. The population scale score mean was 686.32 with a standard deviation of 33.80. The gender subgroups performed very similarly, with a mean difference of less than one scale score point. Asian and White students' scale score means exceeded the State mean scale score on the test. Asian students (the highest performing ethnic subgroup) exceeded the State mean by close to 20 scale score points. American Indian, Black, and Hispanic ethnic subgroups had mean scale scores approximately one standard deviation below the Asian subgroup. Students from Low Needs and Average Needs districts outperformed the other NRC subgroups. The lowest performing NRC subgroup was the Big 4 Cities, with a mean of 662.24, nearly one-half of a standard deviation below the second lowest performing NRC subgroup (High Needs, Urban/Suburban: 677.59) and close to 40 scale score units below the Low Needs subgroup mean. SWD, SUA, and ELL without testing in an alternate language subgroup

scored consistently below the Statewide percentile scale score rankings. The Haitian-Creole translation subgroup, which had a scale score mean (628.96) of more than 57 units below the population mean, was the lowest performing group analyzed. The Korean translation subgroup was the highest performing group analyzed, with a scale score mean of 713.66, more than four-fifths of a standard deviation above the population mean. At the 50th percentile, the following groups exceeded the population scale score of 687: Asian (704), White (690), Average Needs (690), Low Needs (699), and students who used the Chinese (694) and Korean (711) translations.

Table 48. 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	199180	686.32	33.80	647	666	687	704	720
Gender	Female	97620	686.66	33.00	649	668	687	704	720
	Male	101560	685.99	34.54	647	666	687	704	734
Ethnicity	Asian	15058	705.82	33.55	668	687	704	720	758
	Black	38167	671.82	32.59	635	654	673	690	711
	Hispanic	42248	677.39	31.74	641	659	678	699	711
	American Indian	955	677.37	30.73	644	659	678	694	711
	Multi-Racial	520	682.01	35.10	644	663	681	699	720
	White	102129	692.65	31.99	656	673	690	711	734
	Unknown	103	697.31	33.01	659	675	699	720	734
NRC	New York City	69778	683.34	34.99	644	663	684	704	720
	Big 4 Cites	7667	662.24	35.06	625	644	663	684	699
	High Needs Urban/Suburban	15618	677.59	31.26	641	659	678	694	711
	High Needs Rural	11448	678.78	29.91	644	661	678	694	711
	Average Needs	60097	689.67	31.39	654	670	690	704	734
	Low Needs	30298	701.16	30.89	666	681	699	720	734
	Charter	3556	680.61	27.60	647	663	681	694	711
SWD	All Codes	30811	657.57	34.80	620	638	661	678	694
SUA	All Codes	47992	663.72	34.98	625	647	666	684	704
SWD/SUA	SUA=504 Plan Codes	28263	656.46	34.46	620	638	659	678	694
ELL/SUA	SUA=ELL Codes	12149	664.29	34.26	629	647	666	684	704
ELL	English	11364	664.00	33.52	625	647	666	684	699
	Chinese	387	698.39	29.32	663	678	694	720	734
	Haitian Creole	70	628.96	47.86	573	615	638	654	680
	Korean	53	713.66	23.19	690	699	711	734	758
	Russian	59	665.90	38.04	632	649	673	687	704
	Spanish	2864	657.12	34.81	620	641	661	678	694
	All Translations	3433	662.22	37.67	620	641	663	684	704

Grade 6

Grade 6 scale score statistics and N-counts of demographic groups are presented in Table 49. The population scale score mean was 679.91 with a standard deviation of 35.21. The gender

subgroups performed very similarly, with a mean difference of less than three scale score points. Asian and White students' scale score means exceeded the State mean scale score. American Indian, Black, and Hispanic ethnic subgroups had mean scale scores approximately one standard deviation below the Asian subgroup. Students from Low Needs and Average Needs districts outperformed the other NRC subgroups. The lowest performing NRC subgroup was the Big 4 Cities, with a mean of 657.00. New York City, High Needs Urban/Suburban, High Needs Rural, and Charter subgroups had similar scale score means (ranging from approximately 671–676). SWD, SUA, and ELL without testing in an alternate language subgroup scored consistently below the Statewide percentile scale score rankings. The Haitian-Creole translation subgroup, which had a scale score mean (639.50) more than 40 units below the population mean, was the lowest performing group analyzed. Asian students (the highest performing subgroup with a mean of 702.04) exceeded the State mean by over 22 scale score points. At the 50th percentile, the following groups exceeded the population scale score of 680: Asian (699), White (685), Average Needs (683), Low Needs (695), and students who used the Chinese (688) and Korean (697) translations.

Table 49. 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	199605	679.91	35.21	639	660	680	699	719
Gender	Female	97359	681.17	34.17	642	663	680	699	719
	Male	102246	678.71	36.13	637	658	680	699	719
Ethnicity	Asian	15322	702.04	35.79	660	680	699	719	756
	Black	38288	664.37	32.84	627	647	667	685	699
	Hispanic	41482	668.31	33.64	627	649	669	688	704
	American Indian	903	668.84	31.19	630	652	669	688	704
	Multi-Racial	463	679.47	34.87	645	658	678	699	719
	White	103072	687.15	32.70	652	669	685	704	731
	Unknown	75	695.00	35.62	654	669	695	719	731
NRC	New York City	69450	674.60	37.22	630	652	675	695	719
	Big 4 Cites	7653	657.00	33.52	620	639	658	678	695
	High Needs Urban/Suburban	15230	670.60	32.86	634	652	671	691	710
	High Needs Rural	11407	675.88	30.95	642	660	675	691	710
	Average Needs	61209	684.03	31.93	649	667	683	699	719
	Low Needs	30691	696.02	31.99	660	678	695	710	731
	Charter	3196	680.12	27.78	645	663	680	695	710
SWD	All Codes	30554	647.20	35.53	604	627	649	669	685
SUA	All Codes	43839	652.72	36.12	610	634	656	675	691
SWD/SUA	SUA=504 Plan Codes	27567	646.17	35.04	604	627	649	669	685
ELL/SUA	SUA=ELL Codes	9936	652.50	36.10	610	634	654	673	691

(Continued on next page)

Table 49. Scale Score Distribution Summary, by Subgroup, Grade 6 (cont.)

Demographic Category (Subgroup)		N-count	SS Mean	SS Std Dev	10th %tile	25th %tile	50th %tile	75th %tile	90th %tile
ELL	English	9293	651.02	35.06	610	634	654	673	688
	Chinese	486	688.75	30.02	652	671	688	704	719
	Haitian Creole	115	639.50	43.67	588	615	647	669	685
	Korean	68	696.84	31.82	656	677	697	715	756
	Russian	61	658.95	40.81	610	630	667	685	719
	Spanish	3057	647.50	36.82	604	627	652	671	688
	All Translations	3787	653.62	39.25	610	634	656	678	699

Grade 7

N-counts and scale score statistics of demographic groups for Grade 7 are presented in Table 50. The population scale score mean was 680.84 with a standard deviation of 32.27. The gender subgroups performed very similarly, with a mean difference of less than two scale score points. Asian and White ethnic subgroups' scale score means exceeded the State mean scale score. American Indian, Black, and Hispanic ethnic subgroups had mean scale scores between one-quarter and one-half of a standard deviation below the population. The lowest performing NRC subgroup, Big 4 Cities, had a scale score mean of 657.25, while the Low Needs subgroup's scale score mean was 697.26. SWD, SUA, and ELL without testing in an alternate language subgroup scored consistently below the Statewide percentile scale score rankings and had means nearly one standard deviation below the population mean. The Haitian-Creole translation was the lowest performing group analyzed, while the Korean translation subgroup was the highest. At the 50th percentile, the following groups exceeded the population scale score of 679: Asian (699), White (686), Average Needs (684), Low Needs (695), and students who used the Chinese (686) and Korean (695) translations.

Table 50. 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	204292	680.84	32.27	646	662	679	699	715
Gender	Female	99379	681.46	31.35	646	664	679	699	715
	Male	104913	680.24	33.10	644	662	679	699	715
Ethnicity	Asian	15496	700.54	38.00	660	679	699	715	743
	Black	38510	664.87	27.88	633	651	665	681	695
	Hispanic	42151	668.50	28.32	636	653	669	684	699
	American Indian	942	671.72	27.69	639	658	671	689	703
	Multi-Racial	420	679.37	34.29	642	662	677	699	725
	White	106690	688.69	30.05	658	671	686	703	725
	Unknown	83	687.57	37.46	658	669	686	703	725

(Continued on next page)

Table 50. Scale Score Distribution Summary, by Subgroup, Grade 7 (cont.)

Demographic Category (Subgroup)		N-count	SS Mean	SS Std Dev	10th %tile	25th %tile	50th %tile	75th %tile	90th %tile
NRC	New York City	71034	674.01	33.40	639	654	671	692	715
	Big 4 Cites	7830	657.25	30.04	626	642	658	675	689
	High Needs Urban/Suburban	15511	671.52	27.33	642	656	671	686	703
	High Needs Rural	12047	677.46	27.00	649	664	677	692	708
	Average Needs	63079	686.64	28.67	656	671	684	699	715
	Low Needs	31366	697.26	30.84	665	679	695	708	725
	Charter	2472	677.54	26.85	649	662	675	692	708
SWD	All Codes	31204	652.74	29.33	621	639	654	671	684
SUA	All Codes	42979	656.39	29.90	626	642	658	673	689
SWD/SUA	SUA=504 Plan Codes	27928	651.96	28.92	621	639	654	669	681
ELL/SUA	SUA=ELL Codes	9567	654.50	29.35	621	639	656	671	686
ELL	English	8163	651.87	29.83	621	636	654	669	684
	Chinese	579	686.64	29.93	653	669	686	703	725
	Haitian Creole	148	646.53	29.47	616	630	646	665	679
	Korean	98	698.49	24.67	671	681	695	715	725
	Russian	77	663.04	25.31	626	646	664	681	695
	Spanish	3375	652.90	26.57	621	639	654	669	681
	All Translations	4277	658.47	30.12	626	642	658	675	695

Grade 8

Grade 8 scale score statistics and N-counts of demographic groups are presented in Table 51. The population scale score mean was 674.99 with a standard deviation of 33.75. The gender subgroups performed similarly, with a mean difference of less than four scale score points. Asian and White ethnic subgroups' scale score means exceeded the State mean scale score. The Black, Hispanic, and American Indian ethnic subgroups' scale score means were all close to or more than 10 scale score points below the population mean. The lowest performing NRC subgroup, Big 4 Cities, had a scale score mean of 648.12, while the Low Needs subgroup's scale score mean was 692.26, which indicated a large performance discrepancy by school district NRC designation. SWD, SUA, and ELL without testing in an alternate language subgroup scored consistently below the Statewide percentile scale score rankings. At the 50th percentile, the following groups exceeded the population scale score of 674: Female (676), Asian (695), White (682), Average Needs (680), Low Needs (692), and students who used the Chinese (692) and Korean (699) translations.

Table 51. 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	208835	674.99	33.75	636	656	674	695	717
Gender	Female	102402	676.85	33.34	638	657	676	695	717
	Male	106433	673.20	34.04	633	655	673	695	717
Ethnicity	Asian	15613	696.17	34.75	655	674	695	717	745
	Black	39686	658.02	31.28	624	641	658	676	695
	Hispanic	42887	663.17	31.22	628	645	663	682	699
	American Indian	1015	665.79	30.02	630	649	666	682	704
	Multi-Racial	328	671.28	32.60	633	652	669	689	709
	White	109241	682.85	31.12	649	666	682	699	717
	Unknown	65	680.82	36.37	649	666	680	704	727
NRC	New York City	73113	667.72	35.00	630	647	666	689	709
	Big 4 Cites	7793	648.12	33.44	610	630	649	667	686
	High Needs Urban/Suburban	15711	665.62	30.56	630	649	666	684	704
	High Needs Rural	12279	671.44	29.16	640	656	671	689	704
	Average Needs	65121	681.32	29.93	649	665	680	699	717
	Low Needs	31562	692.26	29.58	660	674	692	709	727
	Charter	2162	674.00	27.02	644	657	673	689	709
SWD	All Codes	30515	643.38	34.17	606	626	648	665	680
SUA	All Codes	42818	648.93	34.55	610	631	652	670	686
SWD/SUA	SUA=504 Plan Codes	27455	642.77	33.97	602	626	647	665	678
ELL/SUA	SUA =ELL Codes	9792	652.06	33.51	616	635	653	671	689
ELL	English	8326	649.63	32.62	613	633	652	669	686
	Chinese	533	689.05	31.15	651	670	692	709	727
	Haitian Creole	161	640.41	34.52	602	624	644	661	680
	Korean	80	703.00	28.64	673	682	699	717	745
	Russian	75	659.61	35.34	616	638	665	684	699
	Spanish	3228	646.96	33.74	610	630	651	667	684
	All Translations	4077	653.54	36.98	613	633	655	674	699

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. Table 52 shows the mathematics cut scores used for classification of students to the four performance levels in 2009.

Table 52. Mathematics Grades 3–8 Performance Level Cut Scores

Grade	Level II Cut	Level III Cut	Level IV Cut
3	624	650	703
4	622	650	702
5	619	650	699
6	616	650	696
7	611	650	693
8	616	650	701

Tables 53–59 show the performance level distribution for all examinees from public and charter school with valid scores. Table 53 presents performance level data for total populations of students in Grades 3–8. Tables 54–59 contain performance level data for selected subgroups of students. In general, these summaries reflect the same achievement trends in the scale score summary discussion. Male and Female students performed similarly across grades; however, Females consistently outperformed Males in all grade levels. More White and Asian students were classified in Level III and above, as compared to their peers from other ethnic subgroups. Students from Low and Average Needs districts outperformed students from High Needs districts (New York City, Big 4 Cities, High Needs Urban/Suburban, and High Needs Rural) and Charter schools. The subgroups that took the Korean or Chinese translations outperformed other test translation subgroups. The Level III and above rates for 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, Low Needs, Chinese translation, and Korean translation. Please note that the case counts for the Haitian Creole, Korean, and Russian translation subgroups were very low, and the results might have been heavily influenced by very high and/or very low achieving individual students.

Table 53. Mathematics Test Performance Level Distributions Grades 3–8

Grade	N-count	Percent of New York State Population in Performance Level				
		Level I	Level II	Level III	Level IV	Levels III & IV
3	200058	0.98	5.98	66.06	26.98	93.04
4	197379	3.69	9.00	51.82	35.49	87.31
5	199180	2.16	9.67	52.29	35.89	88.18
6	199605	3.56	13.30	55.02	28.12	83.14
7	204292	1.42	11.16	57.65	29.76	87.41
8	208835	3.47	16.18	61.09	19.27	80.36

Grade 3

Performance level summaries and N-counts of demographic groups for Grade 3 are presented in Table 54. Statewide, 93.04% of third-graders were in Levels III and IV. American Indian, Black, and Hispanic subgroups had a lower percentage of students in Levels III and IV than the rest of the population, but the percentage of Asian, Multi-Racial, and White ethnic subgroups in Levels III and IV exceeded the overall State population. Student achievement varied widely by NRC subgroup, as well. Over 98% of students from Low Needs districts were classified in Levels III and IV; whereas, only about 80% of Big 4 Cities students were in Levels III and IV. Only about four-fifths of SWD, SUA, or those who took translated test forms were classified in Levels III or above; however, the subgroups for Korean and Chinese translations had more than 93% in Levels III and IV with Korean students having the greatest percentage, more than 98%.

Table 54. Performance Level Distributions, by Subgroup, Grade 3

Demographic Category (Subgroup)		N-count	Level I %	Level II %	Level III %	Level IV %	Levels III & IV %
State	All Students	200058	0.98	5.98	66.06	26.98	93.04
Gender	Female	97579	0.73	5.62	65.95	27.70	93.65
	Male	102479	1.23	6.31	66.17	26.29	92.46
Ethnicity	Asian	16338	0.48	2.15	50.21	47.17	97.37
	Black	37659	1.79	11.10	71.58	15.53	87.11
	Hispanic	43140	1.38	8.62	71.86	18.14	90.00
	American Indian	946	0.85	9.41	71.99	17.76	89.75
	Multi-Racial	676	0.30	5.33	68.64	25.74	94.38
	White	101200	0.60	3.54	64.03	31.83	95.86
	Unknown	99	1.01	2.02	60.61	36.36	96.97
NRC	New York City	71498	1.32	7.33	66.36	24.99	91.35
	Big 4 Cites	8334	2.87	17.61	69.17	10.34	79.52
	High Needs Urban/Suburban	16279	1.20	8.63	69.72	20.45	90.17
	High Needs Rural	11509	0.98	6.54	72.72	19.76	92.48
	Average Needs	58657	0.58	4.05	66.85	28.51	95.37
	Low Needs	29687	0.28	1.64	57.82	40.25	98.07
	Charter	3476	0.06	3.80	72.44	23.71	96.15
SWD	All Codes	27469	5.01	20.54	66.34	8.10	74.44
SUA	All Codes	47490	3.43	15.42	68.71	12.44	81.15
SWD/SUA	SUA=504 Plan Codes	23703	5.38	22.06	65.86	6.70	72.55
ELL/SUA	SUA=ELL Codes	17410	2.18	12.27	73.23	12.32	85.55
ELL	ELL status = Y	19796	2.47	12.98	72.51	12.04	84.55
ELL Test Language	English	16319	2.24	12.34	72.91	12.52	85.43
	Chinese	358	0.84	5.59	66.20	27.37	93.58
	Haitian Creole	58	8.62	36.21	53.45	1.72	55.17
	Korean	66	1.52	0.00	56.06	42.42	98.48
	Russian	58	5.17	10.34	62.07	22.41	84.48
	Spanish	3505	3.65	15.89	69.76	10.70	80.46
	All Translations	4045	3.46	14.93	68.88	12.73	81.61

Grade 4

Performance level summaries and N-counts of demographic groups for Grade 4 are presented in Table 55. Statewide, 87.31% of the fourth-grade population was placed in Levels III and IV. Around 5%–7% of American Indian, Black, and Hispanic students were Level I, as compared to only about 1.25% of Asian students and 2.24% of White students. American Indian, Black, and Hispanic ethnic subgroups had percentages of students in Levels III and IV ranging from 77%–83%, but the percentages of the White and Asian subgroups students meeting standards for Levels III and IV (91.77% and 96.01%) exceeded the population. Student achievement also varied widely by NRC subgroup. Almost 96% of students from Low Needs districts were meeting standards for Levels III and IV, but only about 68% Big 4 Cities students were. Only about two-thirds of SWD or SUA status students or those who took translated test forms met or exceeded the Level III cut; however, the Chinese translation subgroup had a very high percentage of students in Levels III and IV (94.66%). 96.36% of students in the Korean translation subgroup were in Levels III and IV. The following subgroups had a higher percentage of students meeting standards for Levels III and IV than the State population: Female, Asian, White, Average Needs, Low Needs, Charter, Chinese translation, and Korean translation.

Table 55. 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	197379	3.69	9.00	51.82	35.49	87.31
Gender	Female	96017	3.24	9.12	52.64	35.00	87.64
	Male	101362	4.12	8.89	51.03	35.96	86.99
Ethnicity	Asian	15073	1.25	2.75	33.46	62.55	96.01
	Black	37635	6.65	15.53	56.48	21.34	77.82
	Hispanic	42368	5.35	12.53	56.54	25.58	82.12
	American Indian	913	4.82	12.49	58.16	24.53	82.69
	Multi-Racial	511	3.33	10.18	53.62	32.88	86.50
	White	100795	2.24	5.98	50.78	41.00	91.77
	Unknown	84	2.38	5.95	44.05	47.62	91.67
NRC	New York City	69845	4.60	10.56	50.21	34.63	84.84
	Big 4 Cites	8068	11.47	20.60	52.76	15.17	67.94
	High Needs Urban/Suburban	15914	4.85	12.54	58.37	24.24	82.61
	High Needs Rural	11454	4.11	11.18	60.82	23.90	84.71
	Average Needs	59087	2.42	7.11	54.04	36.44	90.48
	Low Needs	29501	1.03	3.13	43.50	52.34	95.84
	Charter	2898	1.55	9.04	57.56	31.85	89.41
SWD	All Codes	29088	16.43	22.68	50.67	10.21	60.89
SUA	All Codes	47345	11.96	19.19	53.60	15.25	68.85
SWD/SUA	SUA=504 Plan Codes	26147	17.26	23.48	50.27	9.00	59.26
ELL/SUA	SUA=ELL Codes	14252	8.70	18.37	57.55	15.38	72.93

(Continued on next page)

Table 55. Performance Level Distribution Summary, by Subgroup, Grade 4 (cont.)

Demographic Category (Subgroup)		N-count	Level I %	Level II %	Level III %	Level IV %	Levels III & IV %
ELL	ELL status = Y	16468	9.45	19.03	56.72	14.80	71.51
ELL Test Language	English	13521	8.61	18.38	57.86	15.15	73.01
	Chinese	356	1.12	4.21	46.35	48.31	94.66
	Haitian Creole	82	17.07	32.93	45.12	4.88	50.00
	Korean	55	1.82	1.82	23.64	72.73	96.36
	Russian	65	9.23	21.54	44.62	24.62	69.23
	Spanish	2969	13.71	22.53	53.08	10.68	63.76
	All Translations	3527	12.25	20.58	51.60	15.57	67.17

Grade 5

Performance level summaries and N-counts of demographic groups for Grade 5 are presented in Table 56. Statewide, 88.18% of the fifth-grade population was placed in Levels III and IV. There was little performance differentiation by gender subgroup, with less than 2% difference between each level. However, across ethnic and test translation subgroups, there were marked differences. American Indian, Black, Hispanic, and Multi-Racial ethnic subgroups were below the State average of students meeting standards for Levels III and IV (ranging from 78%–86%), as compared to the percentage of Asian and White students meeting standards for Levels III and IV (96% and 93% respectively). Over 96% of students from Low Needs districts were in Levels III or IV, but only about 68% of the Big 4 Cities students were. Only about 10%–14% of SWD or SUA subgroups were placed in Level IV, compared to the population's 35.89% in Level IV. Less than 14% of students who took translated test forms or who reported ELL with English language test forms were placed in Level IV, except for Russian (15.25%) and the Chinese and Korean translation subgroups that had very high percentages of students in Level IV (49.87% and 75.47%). The following subgroups had a higher percentage of students meeting standards for Levels III and IV than the State population: Female, Asian, White, Average Needs, Low Needs, Charter, Chinese translation, and Korean translation.

Table 56. 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	199180	2.16	9.67	52.29	35.89	88.18
Gender	Female	97620	1.83	9.30	53.00	35.87	88.87
	Male	101560	2.47	10.01	51.61	35.90	87.51
Ethnicity	Asian	15058	1.07	3.10	33.05	62.78	95.83
	Black	38167	4.18	17.41	58.94	19.47	78.41
	Hispanic	42248	2.96	13.79	58.15	25.10	83.25
	American Indian	955	3.04	11.83	61.57	23.56	85.13

(Continued on next page)

Table 56. Performance Level Distribution Summary, by Subgroup, Grade 5 (cont.)

Demographic Category (Subgroup)		N-count	Level I %	Level II %	Level III %	Level IV %	Levels III & IV %
Ethnicity	Multi-Racial	520	3.27	10.96	56.35	29.42	85.77
	White	102129	1.22	6.01	50.12	42.65	92.77
	Unknown	103	0.00	7.77	39.81	52.43	92.23
NRC	New York City	69778	2.81	11.84	51.78	33.57	85.35
	Big 4 Cites	7667	7.79	24.38	54.70	13.13	67.84
	High Needs Urban/Suburban	15618	2.61	13.60	59.25	24.54	83.79
	High Needs Rural	11448	2.20	11.78	61.94	24.08	86.02
	Average Needs	60097	1.26	6.97	53.48	38.28	91.76
	Low Needs	30298	0.59	3.14	42.22	54.05	96.27
	Charter	3556	1.15	10.52	63.41	24.92	88.33
SWD	All Codes	30811	9.59	27.74	53.16	9.51	62.67
SUA	All Codes	47992	7.42	22.97	55.63	13.98	69.61
SWD/SUA	SUA=504 Plan Codes	28263	9.90	28.47	53.10	8.53	61.63
ELL/SUA	SUA=ELL Codes	12149	6.61	22.03	57.66	13.70	71.36
ELL	ELL status = Y	14215	7.20	23.30	56.46	13.04	69.50
ELL Test Language	English	11364	6.56	22.53	57.48	13.43	70.91
	Chinese	387	0.00	4.65	45.48	49.87	95.35
	Haitian Creole	70	27.14	42.86	30.00	0.00	30.00
	Korean	53	0.00	0.00	24.53	75.47	100.00
	Russian	59	8.47	18.64	57.63	15.25	72.88
	Spanish	2864	9.53	26.71	55.52	8.24	63.76
	All Translations	3433	8.65	24.00	53.42	13.92	67.35

Grade 6

Performance level summaries and N-counts of demographic groups for Grade 6 are presented in Table 57. Statewide, 83.14% of the sixth-grade population was placed in Levels III and IV. There was a slight performance differentiation by gender subgroup with less than 3% difference between each level. There were marked differences across ethnic and test translation subgroups. About 5%–7% of American Indian, Black, and Hispanic students were in Level I, as compared to less than 2% of Asian students and White students. American Indian, Black, and Hispanic ethnic subgroups were below the State average of students meeting standards for Levels III and IV (ranging from 70%–76%), as compared to the percentage of Asian and White students meeting standards for Levels III and IV (93.89% and 90.22%). About 95% of students from Low Needs districts were in Levels III or IV, but only about 61% of the Big 4 Cities students were. Only about 5%–8% of SWD and SUA subgroups were placed in Level IV, compared to the population's 28.12% in Level IV. Less than 15% of students who took translated test forms or who reported ELL with English language test forms were placed in Level IV, except for the Chinese and Korean translation subgroups that had very high percentages of students in Level IV (37.04% and 50%). The following subgroups had a higher percentage of students meeting standards for Levels III and IV than the State population: Female, Asian, White, Average Needs, Low Needs, Chinese translation, and Korean translation.

Table 57. 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	199605	3.56	13.30	55.02	28.12	83.14
Gender	Female	97359	2.86	12.47	56.26	28.41	84.67
	Male	102246	4.24	14.08	53.84	27.85	81.68
Ethnicity	Asian	15322	1.31	4.80	38.47	55.42	93.89
	Black	38288	6.56	23.45	57.21	12.78	69.99
	Hispanic	41482	5.65	20.51	57.23	16.61	73.84
	American Indian	903	4.76	19.71	59.69	15.84	75.53
	Multi-Racial	463	2.38	12.74	58.96	25.92	84.88
	White	103072	1.94	7.83	55.72	34.50	90.22
	Unknown	75	1.33	5.33	46.67	46.67	93.30
NRC	New York City	69450	5.19	17.89	52.42	24.50	76.92
	Big 4 Cites	7653	9.04	30.00	52.63	8.32	60.96
	High Needs Urban/Suburban	15230	4.59	18.46	59.25	17.70	76.95
	High Needs Rural	11407	3.07	12.85	64.07	20.01	84.08
	Average Needs	61209	2.06	9.24	58.49	30.21	88.70
	Low Needs	30691	1.04	4.42	48.84	45.70	94.54
	Charter	3196	1.28	11.92	61.98	24.81	86.80
SWD	All Codes	30554	15.99	34.26	44.93	4.82	49.75
SUA	All Codes	43839	13.10	30.64	48.45	7.81	56.25
SWD/SUA	SUA=504 Plan Codes	27567	16.50	35.05	44.29	4.17	48.46
ELL/SUA	SUA=ELL Codes	9936	12.54	32.01	47.35	8.09	55.44
ELL	ELL status = Y	12280	13.77	32.74	46.32	7.17	53.49
ELL Test Language	English	9293	13.16	32.92	47.53	6.39	53.92
	Chinese	486	1.03	8.02	53.91	37.04	90.95
	Haitian Creole	115	26.09	24.35	46.96	2.61	49.57
	Korean	68	0.00	8.82	41.18	50.00	91.18
	Russian	61	14.75	24.59	45.90	14.75	60.66
	Spanish	3057	15.87	33.50	44.49	6.15	50.64
	All Translations	3787	13.97	29.36	45.74	10.93	56.67

Grade 7

Performance level summaries and N-counts of demographic groups for Grade 7 are presented in Table 58. Statewide, 87.41% of the seventh-grade population was placed in Levels III and IV. Overall there was only slight performance differentiation by gender subgroup with only about 2% difference between each level. However, there were marked differences across ethnic and test translation subgroups. Black, Hispanic, and American Indian ethnic subgroups had around 75%–83% of students meeting standards for Levels III and IV, with less than 19% of those students in Level IV, whereas over 94% of Asian students were meeting standards for Levels III and IV (and over 55% were in Level IV.) About 64% of Big 4 Cities students were meeting standards for Levels III and IV, with less than 8% in Level IV, yet over 97% of students from Low Needs districts were meeting standards for Levels III and IV (with about 50% in Level IV). Less than 8% of SWD and SUA subgroups were placed in Level IV, and about 6% were in Level I. Less than 17% of students who took translated test

forms or who reported ELL with English language test forms were placed in Level IV, except for the Chinese and Korean translation subgroups that had very high rates (37.31% and 56.12%). Across all subgroups, the Haitian-Creole translation subgroup had the largest percentage of students placed in Level I (9.46%) and the Korean translation subgroup had the largest percentage of students (98.98%) who met the standards for Levels III and IV. The following subgroups had a higher percentage of students meeting Levels III and IV standards than the State population: Female, Asian, White, High Needs Rural, Average Needs, Low Needs, Chinese translation, and Korean translation.

Table 58. 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	204292	1.42	11.16	57.65	29.76	87.41
Gender	Female	99379	1.16	10.35	58.61	29.88	88.49
	Male	104913	1.68	11.94	56.74	29.64	86.39
Ethnicity	Asian	15496	0.87	4.65	39.06	55.43	94.48
	Black	38510	2.61	22.23	63.47	11.68	75.15
	Hispanic	42151	2.32	18.92	63.80	14.96	78.76
	American Indian	942	1.59	15.61	64.44	18.37	82.80
	Multi-Racial	420	2.14	12.38	56.19	29.29	85.48
	White	106690	0.72	5.01	55.77	38.50	94.27
	Unknown	83	2.41	3.61	53.01	40.96	93.98
NRC	New York City	71034	2.11	17.22	57.82	22.86	80.68
	Big 4 Cites	7830	4.99	30.57	56.73	7.70	64.43
	High Needs Urban/Suburban	15511	1.75	14.94	66.53	16.79	83.32
	High Needs Rural	12047	1.15	9.11	67.42	22.31	89.73
	Average Needs	63079	0.63	5.43	58.97	34.97	93.94
	Low Needs	31366	0.29	2.71	46.56	50.45	97.00
	Charter	2472	0.73	9.83	66.67	22.78	89.44
SWD	All Codes	31204	6.49	34.30	54.43	4.79	59.22
SUA	All Codes	42979	5.58	30.84	56.30	7.28	63.58
SWD/SUA	SUA=504 Plan Codes	27928	6.66	35.32	53.79	4.23	58.02
ELL/SUA	SUA=ELL Codes	9567	5.82	33.67	53.77	6.74	60.51
ELL	ELL status = Y	11644	6.42	35.31	52.13	6.15	58.28
ELL Test Language	English	8163	6.90	35.75	52.24	5.12	57.36
	Chinese	579	0.52	7.43	54.75	37.31	92.06
	Haitian Creole	148	9.46	41.89	44.59	4.05	48.65
	Korean	98	0.00	1.02	42.86	56.12	98.98
	Russian	77	1.30	29.87	51.95	16.88	68.83
	Spanish	3375	5.51	35.79	54.19	4.50	58.70
	All Translations	4277	4.77	31.26	53.64	10.33	63.97

Grade 8

Performance level summaries and N-counts of demographic groups for Grade 8 are presented in Table 59. Statewide, 80.36% of the eighth-grade population was placed in Levels III and

IV. Overall, there was little performance differentiation by gender subgroup, with less than 3% difference between each level percentage. Across ethnic and test translation subgroups, there were marked differences in performance. Around 4%–7% of Black, Hispanic, and American Indian students were in Level I, compared to less than 2% of Asian and White students. American Indian, Black, Hispanic, and Multi-Racial ethnic subgroups had around 63%–74% of students meeting standards for Levels III and IV, respectively, whereas about 92% of Asian students were meeting Levels III and IV standards. About 49% of Big 4 Cities students were in Levels III and IV, yet over 94% of students from Low Needs districts were classified in these proficiency levels. Approximately 11%–16% of SWD, SUA, and ELL students were placed in Level I. Less than 10% of students who took translated test forms or who reported ELL with English language test forms were placed in Level IV, except for the Chinese and Korean translation subgroups that had a very high percentage of students in Level IV (33.58% and 46.25%). Across all subgroups, the Haitian-Creole translation subgroup had the largest percentage of students placed in Level I (16.15%), and the Korean translation subgroup had the largest percentage of students placed in Level IV (46.25%). The following subgroups had a higher percentage of students meeting standards for Levels III and IV than the State population: Female, Asian, White, High Needs Rural, Average Needs, Low Needs, Charter, Chinese translation, and Korean translation.

Table 59. 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	208835	3.47	16.18	61.09	19.27	80.36
Gender	Female	102402	2.89	15.06	61.39	20.66	82.05
	Male	106433	4.03	17.25	60.80	17.92	78.72
Ethnicity	Asian	15613	1.33	6.73	48.88	43.07	91.94
	Black	39686	6.60	30.31	56.34	6.76	63.10
	Hispanic	42887	5.05	25.52	60.12	9.31	69.43
	American Indian	1015	4.04	21.77	63.45	10.74	74.19
	Multi-Racial	328	3.05	18.90	61.59	16.46	78.05
	White	109241	2.01	8.67	64.92	24.40	89.32
	Unknown	65	3.08	9.23	60.00	27.69	87.69
NRC	New York City	73113	4.66	24.27	55.86	15.21	71.07
	Big 4 Cites	7793	12.41	38.59	44.89	4.12	49.01
	High Needs Urban/Suburban	15711	4.23	22.44	62.95	10.39	73.34
	High Needs Rural	12279	3.14	14.37	70.10	12.40	82.50
	Average Needs	65121	1.90	8.90	67.36	21.84	89.20
	Low Needs	31562	0.91	4.35	60.30	34.44	94.74
	Charter	2162	1.20	14.34	70.26	14.20	84.46
SWD	All Codes	30515	15.83	37.82	44.52	1.83	46.35
SUA	All Codes	42818	12.81	34.33	48.90	3.96	52.85

(Continued on next page)

Table 59. Performance Level Distribution Summary, by Subgroup, Grade 8 (cont.)

Demographic Category (Subgroup)		N-count	Level I %	Level II %	Level III %	Level IV %	Levels III & IV %
SWD/SUA	SUA=504 Plan Codes	27455	16.14	38.34	43.86	1.65	45.52
ELL/SUA	SUA=ELL Codes	9792	9.95	35.09	49.38	5.59	54.96
ELL	ELL status = Y	11764	11.06	36.00	47.89	5.05	52.94
ELL Test Language	English	8326	11.04	36.39	48.49	4.08	52.57
	Chinese	533	0.75	8.82	56.85	33.58	90.43
	Haitian Creole	161	16.15	44.72	37.27	1.86	39.13
	Korean	80	0.00	1.25	52.50	46.25	98.75
	Russian	75	9.33	25.33	56.00	9.33	65.33
	Spanish	3228	12.76	37.21	46.81	3.22	50.03
	All Translations	4077	11.01	32.87	48.03	8.09	56.12

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 Mathematics 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 60.

Table 60. Mathematics 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	200058	692.06	37.02	0.98	5.98	66.06	26.98	93.04
	2008	197306	688.36	34.39	2.26	7.80	63.60	26.34	89.94
	2007	200071	684.93	36.64	4.09	10.61	55.97	29.33	85.30
	2006	201908	677.49	37.75	6.35	13.13	55.42	25.11	80.52
4	2009	197379	689.59	38.28	3.69	9.00	51.82	35.49	87.31
	2008	198509	683.13	38.11	4.70	11.37	54.49	29.45	83.93
	2007	199181	679.91	39.85	6.02	13.97	52.52	27.49	80.01
	2006	202695	676.55	40.81	7.41	14.59	52.12	25.88	78.00
5	2009	199180	686.32	33.80	2.16	9.67	52.29	35.89	88.18
	2008	199474	679.65	36.38	3.77	12.93	56.27	27.04	83.31
	2007	203670	673.69	37.93	5.78	18.01	54.10	22.11	76.20
	2006	209200	665.59	39.85	10.29	21.24	49.31	19.16	68.47
6	2009	199605	679.91	35.21	3.56	13.30	55.02	28.12	83.14
	2008	201719	674.85	38.21	5.45	15.04	53.21	26.31	79.52
	2007	205976	667.96	40.34	8.71	19.94	51.33	20.02	71.35
	2006	211376	655.94	40.44	13.32	26.23	47.26	13.19	60.45
7	2009	204292	680.84	32.27	1.42	11.16	57.65	29.76	87.41
	2008	208694	674.60	38.30	3.82	17.15	51.25	27.77	79.02
	2007	213165	662.84	38.16	7.46	26.06	48.13	18.35	66.48
	2006	217225	651.08	40.55	13.19	31.12	43.52	12.17	55.69
8	2009	208835	674.99	33.75	3.47	16.18	61.09	19.27	80.36
	2008	210265	666.44	38.19	7.31	22.69	53.10	16.89	69.99
	2007	215108	656.93	38.62	12.21	28.90	46.97	11.92	58.89
	2006	219294	651.55	41.15	14.98	31.09	43.74	10.18	53.93

As seen in Table 60, an increase in scale score means was observed for all mathematics grades between 2006 and 2009. The least gain was observed for Grades 3 and 4 for which total gain was 15 and 13 scale score points, respectively, between 2006 and 2009 test administrations. The greatest gain in scale score points between 2006 and 2009 test administrations was noted for Grades 6, 7, and 8 (24, 30, and 23 scale score points, respectively).

The variability of scale score distribution decreased steadily across years for mathematics Grades 5, 6, 7, and 8. The scale score standard deviation was around 40 scale score points for those grades in the first test administration year and decreased to around 34 scale score points in 2009. The scale score standard deviation for Grades 3 and 4 only decreased slightly between years 2006 and 2009 (less than 3 scale score points).

Following evaluation of the pattern of means scale score change between the 2006 and 2009 Mathematics Tests administrations, a longitudinal trend of proficiency score distribution was evaluated. The percentage of students classified in Levels III and IV increased each year for each grade but the magnitude of this increase varied depending on the grade level.

An increase of 5% was observed for the percentage of Grade 3 students classified in Levels III and IV between administration years 2006 and 2007, and again between years 2007 and 2008; a 3% increase was observed between years 2008 and 2009, resulting in a total increase of 13% of students classified in Levels III and IV between administration years 2006 and 2008. For Grade 4, an increase of 2% of students classified in Levels III and IV between years 2006 and 2007, 4% between years 2007 and 2008, and 3% between years 2008 and 2009 was noted. The total increase in the percentage of students classified in Levels III and IV for Grade 4 between administration years 2006 and 2009 was approximately 9% (from 78% to 87%). Larger gains in the percentage of students classified in Levels III and IV between administration years 2006 and 2009 were observed for Grades 5, 6, 7, and 8. The Grade 5 proficiency score trend indicated relatively steady increase in the percentage of students classified in Levels III and IV between years 2006 and 2008, with an 8% increase between years 2006 and 2007, approximately 7% increase between years 2007 and 2008, and approximately 5% increase between year 2008 and 2009. Overall, the percentage of Grade 5 students classified in Levels III and IV increased from approximately 68% to 88% between years 2006 and 2009. The Grade 6 trend showed approximately an 11% increase of students classified in Levels III and IV between years 2006 and 2007, an 8% increase between years 2007 and 2008, and a 3% increase between year 2008 and 2009. Overall, the percentage of Grade 6 students classified in Levels III and IV increased by about 22% from 60.5% to 83% between years 2006 and 2009. The Grade 7 proficiency score tend showed the most gain in the four years of Mathematics Tests administrations. It was observed that the percentage of students classified in Levels III and IV increased by approximately 10% between years 2006 and 2007, by approximately 13% between year 2007 and 2008, and by about 8% between years 2008 and 2009. Overall, the percentage of Grade 7 students classified in Levels III and IV increased by approximately 31% from 56% to 87% between years 2006 and 2009. The Grade 8 trend showed approximately a 5% increase in the percentage of students classified in Levels III and IV between years 2006 and 2007, about an 11% increase between years 2007 and 2008, and a 10% increase between years 2008 and 2009. Overall, the percentage of Grade 8 students classified in Levels III and IV increased by about 26% from approximately 54% to 80% between years 2006 and 2009.

In summary, an increase in the mean scale score and the percentage of students classified in Levels III and IV was observed in the third and fourth years of the Mathematics Tests administrations for all grade levels. These changes were not uniform across grades. The least gain was observed for Grades 3 and 4, while the largest increase was noted for Grades 6 and 7. As expected, the mean scale score change was found to be in alignment with the performance levels score trend between years 2007 and 2009.

Appendix A—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 is scorable 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 for 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 B—Psychometric Guidelines for Operational Item Selection

It is primarily up to the content development department to select items for the 2009 OP test. Research staff will provide support, as necessary, and will review the final item selection. Research staff 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% difference the of score point percentage 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.
- Minimize the number of items flagged for DIF (gender, ethnicity, 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 TCCs 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 your MC items cover a wide range of the scale.
- Provide research 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 C—Factor Analysis Results

As described in Section III, "Validity," a principal component factor analysis, was conducted on the Grades 3–8 Mathematics 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 C1 contains a table of eigenvalues and proportion of variance accounted for by extracted factors for these subgroups.

Table C1. Factor Analysis Results for Mathematics Tests (Selected Subpopulations)

Grade	Subgroup	Initial Eigenvalues			
		Component	Total	% of Variance	Cumulative %
3	ELL	**1**	**7.80**	**25.17**	**25.17**
		2	1.36	4.38	29.54
		3	1.04	3.37	32.91
		4	1.02	3.28	36.19
	SWD	**1**	**8.44**	**27.22**	**27.22**
		2	1.29	4.17	31.39
		3	1.05	3.37	34.76
		4	1.00	3.24	38.00
	SUA	**1**	**8.45**	**27.24**	**27.24**
		2	1.30	4.18	31.42
		3	1.06	3.43	34.85
	SWD/SUA	**1**	**8.41**	**27.13**	**27.13**
		2	1.32	4.24	31.37
		3	1.03	3.33	34.70
		4	1.01	3.25	37.96
	ELL/SUA	**1**	**7.85**	**25.33**	**25.33**
		2	1.38	4.46	29.80
		3	1.03	3.31	33.10
		4	1.00	3.23	36.34
4	ELL	**1**	**12.32**	**25.66**	**25.66**
		2	1.46	3.04	28.69
		3	1.20	2.50	31.19
		4	1.06	2.21	33.41
		5	1.03	2.14	35.55
		6	1.00	2.09	37.64
	SWD	**1**	**13.48**	**28.09**	**28.09**
		2	1.41	2.95	31.03
		3	1.21	2.51	33.55
		4	1.08	2.24	35.79
		5	1.02	2.13	37.91

(Continued on next page)

Table C1. Factor Analysis Results for Mathematics Tests (Selected Subpopulations) (cont.)

Grade	Subgroup	Initial Eigenvalues			
		Component	Total	% of Variance	Cumulative %
4	SWD/SUA	**1**	**13.50**	**28.13**	**28.13**
		2	1.41	2.94	31.06
		3	1.20	2.50	33.56
		4	1.08	2.24	35.81
		5	1.03	2.15	37.95
	ELL/SUA	**1**	**12.54**	**26.13**	**26.13**
		2	1.47	3.06	29.19
		3	1.21	2.53	31.72
		4	1.05	2.18	33.90
		5	1.01	2.11	36.01
5	ELL	**1**	**8.21**	**24.16**	**24.16**
		2	1.23	3.63	27.78
		3	1.04	3.07	30.85
		4	1.00	2.95	33.80
	SWD	**1**	**8.15**	**23.97**	**23.97**
		2	1.25	3.68	27.65
		3	1.03	3.04	30.68
		4	1.01	2.96	33.64
	SUA	**1**	**8.49**	**24.98**	**24.98**
		2	1.22	3.59	28.56
		3	1.02	2.99	31.55
	SWD/SUA	**1**	**8.07**	**23.75**	**23.75**
		2	1.26	3.72	27.46
		3	1.03	3.04	30.50
		4	1.01	2.96	33.46
	ELL/SUA	**1**	**8.28**	**24.34**	**24.34**
		2	1.23	3.60	27.94
		3	1.03	3.02	30.96
6	ELL	**1**	**8.53**	**24.37**	**24.37**
		2	1.42	4.07	28.44
		3	1.22	3.49	31.92
	SWD	**1**	**8.48**	**24.22**	**24.22**
		2	1.43	4.08	28.30
		3	1.24	3.54	31.84
	SUA	**1**	**8.98**	**25.66**	**25.66**
		2	1.41	4.01	29.67
		3	1.23	3.50	33.18
	SWD/SUA	**1**	**8.33**	**23.79**	**23.79**
		2	1.44	4.10	27.89
		3	1.25	3.57	31.46

(Continued on next page)

Table C1. Factor Analysis Results for Mathematics Tests (Selected Subpopulations) (cont.)

Grade	Subgroup	Initial Eigenvalues			
		Component	Total	% of Variance	Cumulative %
6	ELL/SUA	**1**	**8.84**	**25.26**	**25.26**
		2	1.43	4.09	29.35
		3	1.22	3.49	32.83
7	ELL	**1**	**7.44**	**19.57**	**19.57**
		2	1.62	4.27	23.84
		3	1.39	3.67	27.51
		4	1.17	3.07	30.58
		5	1.03	2.70	33.28
	SWD	**1**	**7.36**	**19.38**	**19.38**
		2	1.58	4.17	23.54
		3	1.36	3.58	27.12
		4	1.12	2.95	30.07
		5	1.03	2.72	32.79
	SUA	**1**	**7.79**	**20.49**	**20.49**
		2	1.59	4.18	24.66
		3	1.37	3.60	28.26
		4	1.13	2.98	31.24
		5	1.02	2.69	33.92
	SWD/SUA	**1**	**7.30**	**19.20**	**19.20**
		2	1.58	4.16	23.36
		3	1.36	3.58	26.94
		4	1.11	2.93	29.87
		5	1.03	2.71	32.59
	ELL/SUA	**1**	**7.71**	**20.30**	**20.30**
		2	1.63	4.28	24.58
		3	1.37	3.59	28.17
		4	1.15	3.03	31.20
		5	1.02	2.67	33.87
8	ELL	**1**	**12.69**	**28.19**	**28.19**
		2	1.85	4.11	32.30
		3	1.46	3.24	35.54
		4	1.20	2.68	38.21
		5	1.11	2.47	40.68
		6	1.07	2.37	43.05
	SWD	**1**	**12.42**	**27.60**	**27.60**
		2	1.86	4.13	31.73
		3	1.45	3.23	34.96
		4	1.17	2.61	37.57
		5	1.09	2.42	39.99
		6	1.05	2.33	42.32

(Continued on next page)

Table C1. Factor Analysis Results for Mathematics Tests (Selected Subpopulations) (cont.)

Grade	Subgroup	Initial Eigenvalues			
		Component	Total	% of Variance	Cumulative %
8	SUA	**1**	**13.14**	**29.19**	**29.19**
		2	1.82	4.05	33.25
		3	1.43	3.19	36.43
		4	1.17	2.60	39.04
		5	1.06	2.36	41.40
		6	1.05	2.32	43.72
	SWD/SUA	**1**	**12.34**	**27.41**	**27.41**
		2	1.86	4.13	31.55
		3	1.45	3.23	34.78
		4	1.18	2.62	37.39
		5	1.09	2.43	39.82
		6	1.05	2.34	42.16
	ELL/SUA	**1**	**13.22**	**29.37**	**29.37**
		2	1.82	4.03	33.40
		3	1.43	3.19	36.59
		4	1.19	2.64	39.22
		5	1.12	2.49	41.71
		6	1.06	2.35	44.06

Appendix D—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 D1 shows items flagged by the SMD and Mantel-Haenszel methods, and Table D2 presents items flagged by the Linn-Harnisch method. Note that positive values of SMD and Delta in Table D1 indicate DIF in favor of a focal group and negative values of SMD and Delta indicate DIF against a focal group.

Table D1. NYSTP Mathematics 2009 Classical DIF Item Flags

Grade	Item #	Subgroup	DIF	SMD	Mantel-Haenszel	Delta
3	4	Spanish	In Favor	-0.10	No Flag	No Flag
4	18	Hispanic	Against	No Flag	1659.60	-1.77
4	18	Asian	Against	No Flag	497.82	-1.86
4	18	ELL	Against	-0.11	1133.34	-1.64
4	18	Spanish	Against	-0.10	No Flag	No Flag
4	34	Black	Against	-0.10	No Flag	No Flag
4	34	Hispanic	Against	-0.13	No Flag	No Flag
4	34	ELL	Against	-0.14	No Flag	No Flag
4	34	Spanish	Against	-0.11	No Flag	No Flag
4	39	Black	In Favor	0.16	No Flag	No Flag
4	39	Hispanic	In Favor	0.13	No Flag	No Flag
4	39	Asian	In Favor	0.11	No Flag	No Flag
4	39	High needs	In Favor	0.11	No Flag	No Flag
4	44	Black	Against	-0.10	No Flag	No Flag
4	47	Hispanic	In Favor	0.10	No Flag	No Flag
4	48	Spanish	In Favor	0.10	No Flag	No Flag
5	5	Spanish	In Favor	0.12	No Flag	No Flag
5	27	Black	In Favor	0.12	No Flag	No Flag
5	27	Spanish	In Favor	0.10	No Flag	No Flag
5	29	Spanish	In Favor	0.11	No Flag	No Flag
5	31	Asian	Against	-0.12	No Flag	No Flag
5	33	Black	Against	-0.10	No Flag	No Flag
5	33	Hispanic	Against	-0.10	No Flag	No Flag
5	33	Asian	Against	-0.16	No Flag	No Flag
5	34	Black	In Favor	0.12	No Flag	No Flag
5	34	Asian	In Favor	0.14	No Flag	No Flag
5	34	High needs	In Favor	0.11	No Flag	No Flag
5	34	ELL	In Favor	0.13	No Flag	No Flag
5	34	Spanish	In Favor	0.11	No Flag	No Flag

(Continued on next page)

Table D1. NYSTP Mathematics 2009 Classical DIF Item Flags (cont.)

Grade	Item #	Subgroup	DIF	SMD	Mantel-Haenszel	Delta
6	27	Black	In Favor	0.10	No Flag	No Flag
6	27	Hispanic	In Favor	0.11	No Flag	No Flag
6	30	Asian	Against	-0.10	No Flag	No Flag
6	30	ELL	Against	-0.17	No Flag	No Flag
6	30	Spanish	Against	-0.10	No Flag	No Flag
6	31	Black	In Favor	0.12	No Flag	No Flag
6	31	Asian	In Favor	0.10	No Flag	No Flag
6	32	Black	In Favor	0.13	No Flag	No Flag
6	32	Female	In Favor	0.10	No Flag	No Flag
6	33	ELL	In Favor	0.10	No Flag	No Flag
6	33	Spanish	In Favor	0.10	No Flag	No Flag
6	34	Black	Against	-0.10	No Flag	No Flag
6	34	Hispanic	Against	-0.10	No Flag	No Flag
6	34	High needs	Against	-0.10	No Flag	No Flag
6	34	Spanish	Against	-0.12	No Flag	No Flag
7	1	Spanish	Against	-0.10	No Flag	No Flag
7	4	ELL	Against	No Flag	331.99	-1.67
7	4	Spanish	Against	No Flag	128.60	-1.68
7	5	Female	Against	-0.12	5065.89	-2.02
7	9	ELL	Against	No Flag	666.60	-1.62
7	11	Female	Against	-0.12	4823.55	-1.97
7	16	Asian	In Favor	No Flag	321.63	1.68
7	21	Spanish	Against	-0.12	273.46	-1.59
7	23	Female	Against	No Flag	2465.14	-1.52
7	29	ELL	In Favor	0.17	1284.91	1.80
7	29	Spanish	In Favor	0.60	No Flag	No Flag
7	31	Black	Against	-0.13	No Flag	No Flag
7	31	Hispanic	Against	-0.16	No Flag	No Flag
7	31	High needs	Against	-0.11	No Flag	No Flag
7	31	ELL	Against	-0.24	No Flag	No Flag
7	31	Spanish	Against	-0.19	No Flag	No Flag
7	32	Female	In Favor	0.12	No Flag	No Flag
7	32	ELL	In Favor	0.11	No Flag	No Flag
7	32	Spanish	In Favor	0.14	No Flag	No Flag
7	35	Hispanic	Against	-0.13	No Flag	No Flag
7	35	Female	Against	-0.18	No Flag	No Flag
7	35	ELL	Against	-0.16	No Flag	No Flag
7	35	Spanish	Against	-0.22	No Flag	No Flag
7	37	Black	In Favor	0.14	No Flag	No Flag

(Continued on next page)

Table D1. NYSTP Mathematics 2009 Classical DIF Item Flags (cont.)

Grade	Item #	Subgroup	DIF	SMD	Mantel-Haenszel	Delta
7	37	Hispanic	In Favor	0.14	No Flag	No Flag
7	37	Female	In Favor	0.10	No Flag	No Flag
7	37	ELL	In Favor	0.15	No Flag	No Flag
7	37	Spanish	In Favor	0.11	No Flag	No Flag
7	38	Black	Against	-0.16	No Flag	No Flag
7	38	Hispanic	Against	-0.12	No Flag	No Flag
7	38	Female	In Favor	0.13	No Flag	No Flag
7	38	ELL	Against	-0.19	No Flag	No Flag
7	38	Spanish	Against	-0.15	No Flag	No Flag
8	7	Female	Against	-0.15	7128.90	-2.25
8	11	ELL	Against	-0.11	969.84	-1.68
8	11	Spanish	Against	-0.11	239.24	-1.52
8	32	Female	In Favor	0.10	No Flag	No Flag
8	32	High needs	Against	-0.14	No Flag	No Flag
8	37	Spanish	Against	-0.14	No Flag	No Flag
8	38	Spanish	In Favor	0.13	No Flag	No Flag
8	40	Hispanic	Against	-0.11	No Flag	No Flag
8	40	Asian	Against	-0.10	No Flag	No Flag
8	40	ELL	Against	-0.25	No Flag	No Flag
8	40	Spanish	Against	-0.18	No Flag	No Flag
8	42	Black	Against	-0.18	No Flag	No Flag
8	42	Asian	Against	-0.12	No Flag	No Flag
8	43	Black	Against	-0.13	No Flag	No Flag
8	43	ELL	Against	-0.14	No Flag	No Flag
8	43	Spanish	Against	-0.14	No Flag	No Flag
8	44	Black	In Favor	0.12	No Flag	No Flag
8	45	High needs	Against	-0.12	No Flag	No Flag
8	45	ELL	In Favor	0.10	No Flag	No Flag
8	45	Spanish	In Favor	0.10	No Flag	No Flag

Table D2. Items Flagged for DIF by the Linn-Harnisch Method

Grade	Item	Focal Group	Direction	Magnitude
4	12	Spanish	Against	-0.114
4	18	Spanish	Against	-0.104
4	34	ELL	Against	-0.123
5	5	Spanish	Against	-0.112
5	29	Spanish	In Favor	0.112
5	31	Asian	Against	-0.101
5	33	Asian	Against	-0.115
6	30	ELL	Against	-0.169
6	30	Spanish	Against	-0.105
6	34	Spanish	Against	-0.108
7	21	Spanish	Against	-0.107
7	29	ELL	In Favor	0.157
7	29	Spanish	In Favor	0.589
7	31	ELL	Against	-0.200
7	31	Spanish	Against	-0.152
7	32	Spanish	In Favor	0.125
7	35	ELL	Against	-0.155
7	35	Spanish	Against	-0.215
7	37	ELL	In Favor	0.103
7	38	ELL	Against	-0.175
7	38	Spanish	Against	-0.135
8	37	Spanish	Against	-0.134
8	38	Spanish	In Favor	0.108
8	40	ELL	Against	-0.244
8	40	Spanish	Against	-0.183
8	42	Black	Against	-0.113
8	43	ELL	Against	-0.143
8	43	Spanish	Against	-0.148

Appendix E—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, N-count, obtained-*Z* fit statistic, and critical-*Z* fit statistic are presented for each item. Fit for most items in the Grades 3–8 Mathematics Tests was acceptable (critical *Z* >obtained *Z*).

Table E1. Mathematics Grade 3 Item Fit Statistics

Item	Model	Chi-Square	DF	N-count	Obtained *Z*	Critical *Z*	Fit OK?
1	3PL	333.14	7	170139	87.16	453.70	Y
2	3PL	229.69	7	170139	59.52	453.70	Y
3	3PL	113.78	7	170139	28.54	453.70	Y
4	3PL	781.23	7	170139	206.92	453.70	Y
5	3PL	230.09	7	170139	59.62	453.70	Y
6	3PL	971.37	7	170139	257.74	453.70	Y
7	3PL	86.55	7	170139	21.26	453.70	Y
8	3PL	152.90	7	170139	38.99	453.70	Y
9	3PL	662.65	7	170139	175.23	453.70	Y
10	3PL	153.27	7	170139	39.09	453.70	Y
11	3PL	203.25	7	170139	52.45	453.70	Y
12	3PL	157.79	7	170139	40.30	453.70	Y
13	3PL	255.56	7	170139	66.43	453.70	Y
14	3PL	302.84	7	170139	79.07	453.70	Y
15	3PL	122.75	7	170139	30.94	453.70	Y
16	3PL	1427.33	7	170139	379.60	453.70	Y
17	3PL	771.86	7	170139	204.42	453.70	Y
18	3PL	1386.86	7	170139	368.78	453.70	Y
19	3PL	294.03	7	170139	76.71	453.70	Y
20	3PL	275.73	7	170139	71.82	453.70	Y
21	3PL	100.63	7	170139	25.02	453.70	Y
22	3PL	227.54	7	170139	58.94	453.70	Y
23	3PL	217.18	7	170139	56.17	453.70	Y
24	3PL	1281.90	7	170139	340.73	453.70	Y
25	3PL	659.53	7	170139	174.40	453.70	Y
26	2PPC	944.62	17	170139	159.09	453.70	Y
27	2PPC	1410.46	17	170139	238.98	453.70	Y
28	2PPC	2204.15	17	170139	375.09	453.70	Y
29	2PPC	1499.07	17	170139	254.17	453.70	Y
30	2PPC	2637.00	26	170139	362.08	453.70	Y
31	2PPC	1986.36	26	170139	271.85	453.70	Y

Table E2. Mathematics Grade 4 Item Fit Statistics

Item	Model	Chi-Square	DF	N-count	Obtained *Z*	Critical *Z*	Fit OK?
1	3PL	158.91	7	192988	40.60	514.63	Y
2	3PL	39.31	7	192988	8.64	514.63	Y
3	3PL	25.55	7	192988	4.96	514.63	Y
4	3PL	53.07	7	192988	12.31	514.63	Y
5	3PL	131.53	7	192988	33.28	514.63	Y
6	3PL	681.82	7	192988	180.35	514.63	Y
7	3PL	20.49	7	192988	3.61	514.63	Y
8	3PL	248.18	7	192988	64.46	514.63	Y
9	3PL	33.41	7	192988	7.06	514.63	Y
10	3PL	44.56	7	192988	10.04	514.63	Y
11	3PL	47.91	7	192988	10.93	514.63	Y
12	3PL	341.95	7	192988	89.52	514.63	Y
13	3PL	48.56	7	192988	11.11	514.63	Y
14	3PL	39.03	7	192988	8.56	514.63	Y
15	3PL	60.53	7	192988	14.31	514.63	Y
16	3PL	102.82	7	192988	25.61	514.63	Y
17	3PL	50.15	7	192988	11.53	514.63	Y
18	3PL	93.25	7	192988	23.05	514.63	Y
19	3PL	90.40	7	192988	22.29	514.63	Y
20	3PL	198.71	7	192988	51.24	514.63	Y
21	3PL	105.14	7	192988	26.23	514.63	Y
22	3PL	300.95	7	192988	78.56	514.63	Y
23	3PL	100.12	7	192988	24.89	514.63	Y
24	3PL	73.66	7	192988	17.82	514.63	Y
25	3PL	338.43	7	192988	88.58	514.63	Y
26	3PL	113.26	7	192988	28.40	514.63	Y
27	3PL	1558.27	7	192988	414.59	514.63	Y
28	3PL	84.08	7	192988	20.60	514.63	Y
29	3PL	155.70	7	192988	39.74	514.63	Y
30	3PL	95.74	7	192988	23.72	514.63	Y
31	2PPC	700.65	17	192988	117.24	514.63	Y
32	2PPC	210.63	17	192988	33.21	514.63	Y
33	2PPC	417.39	17	192988	68.67	514.63	Y
34	2PPC	443.48	17	192988	73.14	514.63	Y
35	2PPC	2220.80	17	192988	377.95	514.63	Y
36	2PPC	697.16	17	192988	116.65	514.63	Y
37	2PPC	370.68	17	192988	60.65	514.63	Y
38	2PPC	1839.29	26	192988	251.46	514.63	Y
39	2PPC	9325.98	26	192988	1289.67	514.63	N
40	2PPC	218.26	17	192988	34.52	514.63	Y
41	2PPC	500.56	17	192988	82.93	514.63	Y
42	2PPC	728.06	17	192988	121.95	514.63	Y

(Continued on next page)

Table E2. Mathematics Grade 4 Item Fit Statistics (cont.)

Item	Model	Chi-Square	DF	N-count	Obtained *Z*	Critical *Z*	Fit OK?
43	2PPC	116.57	17	192988	17.08	514.63	Y
44	2PPC	320.74	17	192988	52.09	514.63	Y
45	2PPC	425.32	17	192988	70.03	514.63	Y
46	2PPC	485.55	17	192988	80.36	514.63	Y
47	2PPC	1072.95	26	192988	145.19	514.63	Y
48	2PPC	4090.09	26	192988	563.59	514.63	N

Table E3. Mathematics Grade 5 Item Fit Statistics

Item	Model	Chi-Square	DF	N-count	Obtained *Z*	Critical *Z*	Fit OK?
1	3PL	88.78	7	194048	21.86	517.46	Y
2	3PL	84.17	7	194048	20.63	517.46	Y
3	3PL	93.44	7	194048	23.10	517.46	Y
4	3PL	581.75	7	194048	153.61	517.46	Y
5	3PL	61.92	7	194048	14.68	517.46	Y
6	3PL	62.60	7	194048	14.86	517.46	Y
7	3PL	369.42	7	194048	96.86	517.46	Y
8	3PL	181.93	7	194048	46.75	517.46	Y
9	3PL	32.31	7	194048	6.76	517.46	Y
10	3PL	272.28	7	194048	70.90	517.46	Y
11	3PL	173.52	7	194048	44.50	517.46	Y
12	3PL	161.37	7	194048	41.26	517.46	Y
13	3PL	62.78	7	194048	14.91	517.46	Y
14	3PL	112.36	7	194048	28.16	517.46	Y
15	3PL	366.04	7	194048	95.96	517.46	Y
16	3PL	44.90	7	194048	10.13	517.46	Y
17	3PL	61.18	7	194048	14.48	517.46	Y
18	3PL	110.93	7	194048	27.78	517.46	Y
19	3PL	446.64	7	194048	117.50	517.46	Y
20	3PL	452.92	7	194048	119.18	517.46	Y
21	3PL	164.73	7	194048	42.16	517.46	Y
22	3PL	435.04	7	194048	114.40	517.46	Y
23	3PL	119.21	7	194048	29.99	517.46	Y
24	3PL	83.06	7	194048	20.33	517.46	Y
25	3PL	80.41	7	194048	19.62	517.46	Y
26	3PL	118.25	7	194048	29.73	517.46	Y
27	2PPC	402.94	17	194048	66.19	517.46	Y
28	2PPC	564.96	17	194048	93.97	517.46	Y
29	2PPC	2646.84	17	194048	451.01	517.46	Y
30	2PPC	337.56	17	194048	54.98	517.46	Y
31	2PPC	5864.19	26	194048	809.61	517.46	N
32	2PPC	2223.64	26	194048	304.76	517.46	Y
33	2PPC	5185.37	26	194048	715.48	517.46	N
34	2PPC	1052.38	26	194048	142.33	517.46	Y

Table E4. Mathematics Grade 6 Item Fit Statistics

Item	Model	Chi-Square	DF	N-count	Obtained Z	Critical Z	Fit OK?
1	3PL	368.45	7	194194	96.60	517.85	Y
2	3PL	58.47	7	194194	13.76	517.85	Y
3	3PL	25.68	7	194194	4.99	517.85	Y
4	3PL	166.41	7	194194	42.60	517.85	Y
5	3PL	422.33	7	194194	111.00	517.85	Y
6	3PL	161.83	7	194194	41.38	517.85	Y
7	3PL	288.45	7	194194	75.22	517.85	Y
8	3PL	211.77	7	194194	54.73	517.85	Y
9	3PL	106.55	7	194194	26.60	517.85	Y
10	3PL	516.13	7	194194	136.07	517.85	Y
11	3PL	108.32	7	194194	27.08	517.85	Y
12	3PL	36.16	7	194194	7.79	517.85	Y
13	3PL	274.36	7	194194	71.46	517.85	Y
14	3PL	704.52	7	194194	186.42	517.85	Y
15	3PL	372.37	7	194194	97.65	517.85	Y
16	3PL	141.87	7	194194	36.05	517.85	Y
17	3PL	57.14	7	194194	13.40	517.85	Y
18	3PL	827.79	7	194194	219.37	517.85	Y
19	3PL	52.03	7	194194	12.03	517.85	Y
20	3PL	227.01	7	194194	58.80	517.85	Y
21	3PL	327.62	7	194194	85.69	517.85	Y
22	3PL	106.96	7	194194	26.71	517.85	Y
23	3PL	168.52	7	194194	43.17	517.85	Y
24	3PL	109.10	7	194194	27.29	517.85	Y
25	3PL	212.71	7	194194	54.98	517.85	Y
26	2PPC	801.14	17	194194	134.48	517.85	Y
27	2PPC	542.81	17	194194	90.18	517.85	Y
28	2PPC	922.51	17	194194	155.29	517.85	Y
29	2PPC	1142.88	17	194194	193.09	517.85	Y
30	2PPC	1013.73	17	194194	170.94	517.85	Y
31	2PPC	621.59	17	194194	103.69	517.85	Y
32	2PPC	1246.57	26	194194	169.26	517.85	Y
33	2PPC	2864.30	26	194194	393.60	517.85	Y
34	2PPC	771.47	26	194194	103.38	517.85	Y
35	2PPC	3928.66	26	194194	541.20	517.85	N

Table E5. Mathematics Grade 7 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	Obtained Z	Critical Z	Fit OK?
1	3PL	125.70	7	196666	31.72	524.44	Y
2	3PL	136.76	7	196666	34.68	524.44	Y
3	3PL	305.63	7	196666	79.81	524.44	Y
4	3PL	236.70	7	196666	61.39	524.44	Y
5	3PL	1021.83	7	196666	271.22	524.44	Y
6	3PL	190.19	7	196666	48.96	524.44	Y
7	3PL	259.15	7	196666	67.39	524.44	Y
8	3PL	48.55	7	196666	11.10	524.44	Y
9	3PL	116.16	7	196666	29.18	524.44	Y
10	3PL	295.47	7	196666	77.10	524.44	Y
11	3PL	883.26	7	196666	234.19	524.44	Y
12	3PL	251.41	7	196666	65.32	524.44	Y
13	3PL	1085.91	7	196666	288.35	524.44	Y
14	3PL	1784.08	7	196666	474.94	524.44	Y
15	3PL	234.53	7	196666	60.81	524.44	Y
16	3PL	41.88	7	196666	9.32	524.44	Y
17	3PL	420.83	7	196666	110.60	524.44	Y
18	3PL	82.93	7	196666	20.29	524.44	Y
19	3PL	149.45	7	196666	38.07	524.44	Y
20	3PL	705.86	7	196666	186.78	524.44	Y
21	3PL	174.77	7	196666	44.84	524.44	Y
22	3PL	332.41	7	196666	86.97	524.44	Y
23	3PL	51.28	7	196666	11.83	524.44	Y
24	3PL	130.75	7	196666	33.07	524.44	Y
25	3PL	1276.28	7	196666	339.23	524.44	Y
26	3PL	44.73	7	196666	10.08	524.44	Y
27	3PL	151.53	7	196666	38.63	524.44	Y
28	3PL	86.67	7	196666	21.29	524.44	Y
29	3PL	402.63	7	196666	105.74	524.44	Y
30	3PL	122.29	7	196666	30.81	524.44	Y
31	2PPC	236.82	17	196666	37.70	524.44	Y
32	2PPC	1731.61	17	196666	294.05	524.44	Y
33	2PPC	1215.84	17	196666	205.60	524.44	Y
34	2PPC	1187.35	17	196666	200.71	524.44	Y
35	2PPC	424.13	26	196666	55.21	524.44	Y
36	2PPC	2312.67	26	196666	317.10	524.44	Y
37	2PPC	734.48	26	196666	98.25	524.44	Y
38	2PPC	798.84	26	196666	107.17	524.44	Y

Table E6. Mathematics Grade 8 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	Obtained Z	Critical Z	Fit OK?
1	3PL	190.34	7	202667	49.00	540.45	Y
2	3PL	44.05	7	202667	9.90	540.45	Y
3	3PL	197.45	7	202667	50.90	540.45	Y
4	3PL	43.41	7	202667	9.73	540.45	Y
5	3PL	172.81	7	202667	44.32	540.45	Y
6	3PL	85.78	7	202667	21.06	540.45	Y
7	3PL	483.15	7	202667	127.26	540.45	Y
8	3PL	555.02	7	202667	146.46	540.45	Y
9	3PL	221.57	7	202667	57.35	540.45	Y
10	3PL	67.89	7	202667	16.27	540.45	Y
11	3PL	194.57	7	202667	50.13	540.45	Y
12	3PL	112.96	7	202667	28.32	540.45	Y
13	3PL	333.51	7	202667	87.26	540.45	Y
14	3PL	169.91	7	202667	43.54	540.45	Y
15	3PL	398.20	7	202667	104.55	540.45	Y
16	3PL	171.12	7	202667	43.86	540.45	Y
17	3PL	471.15	7	202667	124.05	540.45	Y
18	3PL	243.65	7	202667	63.25	540.45	Y
19	3PL	51.47	7	202667	11.88	540.45	Y
20	3PL	237.53	7	202667	61.61	540.45	Y
21	3PL	145.85	7	202667	37.11	540.45	Y
22	3PL	48.04	7	202667	10.97	540.45	Y
23	3PL	4583.73	7	202667	1223.18	540.45	N
24	3PL	91.94	7	202667	22.70	540.45	Y
25	3PL	46.23	7	202667	10.49	540.45	Y
26	3PL	347.61	7	202667	91.03	540.45	Y
27	3PL	641.20	7	202667	169.50	540.45	Y
28	2PPC	232.28	17	202667	36.92	540.45	Y
29	2PPC	2948.16	17	202667	502.69	540.45	Y
30	2PPC	3601.33	17	202667	614.71	540.45	N
31	2PPC	241.24	17	202667	38.46	540.45	Y
32	2PPC	1941.66	26	202667	265.65	540.45	Y
33	2PPC	2053.61	26	202667	281.18	540.45	Y
34	2PPC	4363.49	17	202667	745.42	540.45	N
35	2PPC	5022.06	17	202667	858.36	540.45	N
36	2PPC	1170.62	17	202667	197.84	540.45	Y
37	2PPC	2142.30	17	202667	364.49	540.45	Y
38	2PPC	850.45	17	202667	142.93	540.45	Y
39	2PPC	1180.82	17	202667	199.59	540.45	Y
40	2PPC	249.90	17	202667	39.94	540.45	Y
41	2PPC	2958.56	17	202667	504.47	540.45	Y
42	2PPC	430.44	26	202667	56.09	540.45	Y
43	2PPC	2862.47	26	202667	393.35	540.45	Y
44	2PPC	5926.04	26	202667	818.19	540.45	N
45	2PPC	1532.70	26	202667	208.94	540.45	Y

Appendix F—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 is 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

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

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 $1_i - 1$:

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

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

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

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:

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

and

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

The OPI is

$$\tilde{T}_j = \frac{p_j}{p_j + q_j} \quad (20)$$

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

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

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

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

$$w_j = \frac{n_j^*}{n_j^* + n_j} . \quad (23)$$

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 the 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. While 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 G—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:

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

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

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 H—Scale Score Frequency Distributions

Tables H1–H6 depict the scale score (SS) distributions, by frequency (N-count), percent, cumulative frequency, and cumulative percent for each grade (total population of students from public and charter schools).

Table H1. Grade 3 Mathematics 2009 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
470	98	0.05	98	0.05
560	105	0.05	203	0.10
587	183	0.09	386	0.19
600	247	0.12	633	0.32
608	365	0.18	998	0.50
615	428	0.21	1426	0.71
620	542	0.27	1968	0.98
624	668	0.33	2636	1.32
628	794	0.40	3430	1.71
631	807	0.40	4237	2.12
634	966	0.48	5203	2.60
637	1047	0.52	6250	3.12
639	1226	0.61	7476	3.74
642	1346	0.67	8822	4.41
644	1562	0.78	10384	5.19
646	1696	0.85	12080	6.04
648	1846	0.92	13926	6.96
650	2142	1.07	16068	8.03
652	2369	1.18	18437	9.22
654	2564	1.28	21001	10.50
656	2872	1.44	23873	11.93
658	3236	1.62	27109	13.55
660	3756	1.88	30865	15.43
662	4120	2.06	34985	17.49
664	4792	2.40	39777	19.88
667	5444	2.72	45221	22.60
669	6251	3.12	51472	25.73

(Continued on next page)

Table H1. Grade 3 Mathematics 2009 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
671	7484	3.74	58956	29.47
674	8601	4.30	67557	33.77
677	10267	5.13	77824	38.90
681	12386	6.19	90210	45.09
685	15078	7.54	105288	52.63
690	18299	9.15	123587	61.78
697	22500	11.25	146087	73.02
710	26765	13.38	172852	86.40
770	27206	13.60	200058	100.00

Table H2. Grade 4 Mathematics 2009 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
485	211	0.11	211	0.11
525	182	0.09	393	0.20
550	227	0.12	620	0.31
564	281	0.14	901	0.46
574	345	0.17	1246	0.63
582	399	0.20	1645	0.83
588	425	0.22	2070	1.05
593	468	0.24	2538	1.29
598	503	0.25	3041	1.54
602	515	0.26	3556	1.80
606	549	0.28	4105	2.08
609	578	0.29	4683	2.37
612	603	0.31	5286	2.68
615	615	0.31	5901	2.99
618	690	0.35	6591	3.34
621	691	0.35	7282	3.69
623	731	0.37	8013	4.06
625	831	0.42	8844	4.48
627	861	0.44	9705	4.92

(Continued on next page)

Table H2. Grade 4 Mathematics 2009 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
630	868	0.44	10573	5.36
632	961	0.49	11534	5.84
634	978	0.50	12512	6.34
636	1098	0.56	13610	6.90
637	1166	0.59	14776	7.49
639	1219	0.62	15995	8.10
641	1269	0.64	17264	8.75
643	1358	0.69	18622	9.43
644	1440	0.73	20062	10.16
646	1500	0.76	21562	10.92
648	1690	0.86	23252	11.78
649	1799	0.91	25051	12.69
651	1890	0.96	26941	13.65
653	1959	0.99	28900	14.64
654	2070	1.05	30970	15.69
656	2228	1.13	33198	16.82
658	2300	1.17	35498	17.98
659	2548	1.29	38046	19.28
661	2556	1.29	40602	20.57
663	2767	1.40	43369	21.97
664	2892	1.47	46261	23.44
666	3032	1.54	49293	24.97
668	3211	1.63	52504	26.60
669	3502	1.77	56006	28.37
671	3597	1.82	59603	30.20
673	3846	1.95	63449	32.15
675	4102	2.08	67551	34.22
677	4326	2.19	71877	36.42
679	4625	2.34	76502	38.76
681	5073	2.57	81575	41.33
683	5273	2.67	86848	44.00
686	5727	2.90	92575	46.90

(Continued on next page)

Table H2. Grade 4 Mathematics 2009 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
688	6028	3.05	98603	49.96
691	6418	3.25	105021	53.21
694	6960	3.53	111981	56.73
697	7448	3.77	119429	60.51
700	7896	4.00	127325	64.51
704	8620	4.37	135945	68.88
708	9026	4.57	144971	73.45
713	9665	4.90	154636	78.34
719	9761	4.95	164397	83.29
727	9657	4.89	174054	88.18
737	9044	4.58	183098	92.76
751	7607	3.85	190705	96.62
777	4896	2.48	195601	99.10
800	1778	0.90	197379	100.00

Table H3. Grade 5 Mathematics 2009 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
495	458	0.23	458	0.23
562	314	0.16	772	0.39
583	435	0.22	1207	0.61
595	566	0.28	1773	0.89
603	654	0.33	2427	1.22
610	864	0.43	3291	1.65
615	1006	0.51	4297	2.16
620	1125	0.56	5422	2.72
625	1269	0.64	6691	3.36
629	1440	0.72	8131	4.08
632	1665	0.84	9796	4.92
635	1798	0.90	11594	5.82
638	2007	1.01	13601	6.83
641	2137	1.07	15738	7.90

(Continued on next page)

Table H3. Grade 5 Mathematics 2009 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
644	2385	1.20	18123	9.10
647	2666	1.34	20789	10.44
649	2760	1.39	23549	11.82
652	2941	1.48	26490	13.30
654	3220	1.62	29710	14.92
656	3437	1.73	33147	16.64
659	3748	1.88	36895	18.52
661	4111	2.06	41006	20.59
663	4470	2.24	45476	22.83
666	4595	2.31	50071	25.14
668	5103	2.56	55174	27.70
670	5618	2.82	60792	30.52
673	6048	3.04	66840	33.56
675	6540	3.28	73380	36.84
678	7190	3.61	80570	40.45
681	7665	3.85	88235	44.30
684	8420	4.23	96655	48.53
687	9459	4.75	106114	53.28
690	10413	5.23	116527	58.50
694	11177	5.61	127704	64.11
699	12123	6.09	139827	70.20
704	13193	6.62	153020	76.82
711	13566	6.81	166586	83.64
720	12728	6.39	179314	90.03
734	10553	5.30	189867	95.32
758	6732	3.38	196599	98.70
780	2581	1.30	199180	100.00

Table H4. Grade 6 Mathematics 2009 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
500	555	0.28	555	0.28
556	415	0.21	970	0.49
576	607	0.30	1577	0.79
588	815	0.41	2392	1.20
597	931	0.47	3323	1.66
604	1088	0.55	4411	2.21
610	1243	0.62	5654	2.83
615	1459	0.73	7113	3.56
620	1643	0.82	8756	4.39
624	1798	0.90	10554	5.29
627	1949	0.98	12503	6.26
630	2050	1.03	14553	7.29
634	2216	1.11	16769	8.40
637	2382	1.19	19151	9.59
639	2504	1.25	21655	10.85
642	2727	1.37	24382	12.22
645	2898	1.45	27280	13.67
647	3107	1.56	30387	15.22
649	3265	1.64	33652	16.86
652	3500	1.75	37152	18.61
654	3581	1.79	40733	20.41
656	3746	1.88	44479	22.28
658	3990	2.00	48469	24.28
660	4202	2.11	52671	26.39
663	4553	2.28	57224	28.67
665	4788	2.40	62012	31.07
667	5050	2.53	67062	33.60
669	5192	2.60	72254	36.20
671	5583	2.80	77837	39.00
673	5836	2.92	83673	41.92
675	6261	3.14	89934	45.06
678	6469	3.24	96403	48.30
680	6939	3.48	103342	51.77
683	7177	3.60	110519	55.37

(Continued on next page)

Table H4. Grade 6 Mathematics 2009 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
685	7604	3.81	118123	59.18
688	8158	4.09	126281	63.27
691	8346	4.18	134627	67.45
695	8847	4.43	143474	71.88
699	9535	4.78	153009	76.66
704	9900	4.96	162909	81.62
710	9886	4.95	172795	86.57
719	9694	4.86	182489	91.43
731	8473	4.24	190962	95.67
756	6052	3.03	197014	98.70
780	2591	1.30	199605	100.00

Table H5. Grade 7 Mathematics 2009 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
500	217	0.11	217	0.11
513	188	0.09	405	0.20
570	348	0.17	753	0.37
589	507	0.25	1260	0.62
601	743	0.36	2003	0.98
609	906	0.44	2909	1.42
616	1134	0.56	4043	1.98
621	1355	0.66	5398	2.64
626	1555	0.76	6953	3.40
630	1723	0.84	8676	4.25
633	1964	0.96	10640	5.21
636	2125	1.04	12765	6.25
639	2289	1.12	15054	7.37
642	2407	1.18	17461	8.55
644	2569	1.26	20030	9.80
646	2764	1.35	22794	11.16
649	2921	1.43	25715	12.59

(Continued on next page)

Table H5. Grade 7 Mathematics 2009 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
651	3083	1.51	28798	14.10
653	3212	1.57	32010	15.67
654	3522	1.72	35532	17.39
656	3707	1.81	39239	19.21
658	3888	1.90	43127	21.11
660	4217	2.06	47344	23.17
662	4405	2.16	51749	25.33
664	4624	2.26	56373	27.59
665	4969	2.43	61342	30.03
667	5212	2.55	66554	32.58
669	5542	2.71	72096	35.29
671	5874	2.88	77970	38.17
673	6151	3.01	84121	41.18
675	6491	3.18	90612	44.35
677	6709	3.28	97321	47.64
679	7088	3.47	104409	51.11
681	7204	3.53	111613	54.63
684	7533	3.69	119146	58.32
686	7820	3.83	126966	62.15
689	8200	4.01	135166	66.16
692	8326	4.08	143492	70.24
695	8508	4.16	152000	74.40
699	8703	4.26	160703	78.66
703	8941	4.38	169644	83.04
708	8800	4.31	178444	87.35
715	8495	4.16	186939	91.51
725	7652	3.75	194591	95.25
743	6155	3.01	200746	98.26
800	3546	1.74	204292	100.00

Table H6. Grade 8 Mathematics 2009 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
480	398	0.19	398	0.19
531	324	0.16	722	0.35
567	419	0.20	1141	0.55
581	606	0.29	1747	0.84
590	737	0.35	2484	1.19
597	827	0.40	3311	1.59
602	919	0.44	4230	2.03
606	993	0.48	5223	2.50
610	1018	0.49	6241	2.99
613	1002	0.48	7243	3.47
616	1072	0.51	8315	3.98
619	1120	0.54	9435	4.52
621	1194	0.57	10629	5.09
624	1231	0.59	11860	5.68
626	1349	0.65	13209	6.33
628	1386	0.66	14595	6.99
630	1414	0.68	16009	7.67
631	1440	0.69	17449	8.36
633	1552	0.74	19001	9.10
635	1609	0.77	20610	9.87
636	1714	0.82	22324	10.69
638	1799	0.86	24123	11.55
640	1881	0.90	26004	12.45
641	1865	0.89	27869	13.34
642	2045	0.98	29914	14.32
644	2122	1.02	32036	15.34
645	2114	1.01	34150	16.35
647	2148	1.03	36298	17.38
648	2300	1.10	38598	18.48
649	2427	1.16	41025	19.64
651	2445	1.17	43470	20.82
652	2559	1.23	46029	22.04
653	2601	1.25	48630	23.29
655	2834	1.36	51464	24.64

(Continued on next page)

Table H6. Grade 8 Mathematics 2009 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
656	2868	1.37	54332	26.02
657	2838	1.36	57170	27.38
658	2990	1.43	60160	28.81
660	3144	1.51	63304	30.31
661	3238	1.55	66542	31.86
662	3430	1.64	69972	33.51
663	3526	1.69	73498	35.19
665	3647	1.75	77145	36.94
666	3689	1.77	80834	38.71
667	3829	1.83	84663	40.54
669	3844	1.84	88507	42.38
670	4155	1.99	92662	44.37
671	4195	2.01	96857	46.38
673	4458	2.13	101315	48.51
674	4660	2.23	105975	50.75
676	4897	2.34	110872	53.09
678	4961	2.38	115833	55.47
680	5335	2.55	121168	58.02
682	5666	2.71	126834	60.73
684	5977	2.86	132811	63.60
686	6335	3.03	139146	66.63
689	6710	3.21	145856	69.84
692	7090	3.40	152946	73.24
695	7631	3.65	160577	76.89
699	8025	3.84	168602	80.73
704	8446	4.04	177048	84.78
709	8538	4.09	185586	88.87
717	8260	3.96	193846	92.82
727	7300	3.50	201146	96.32
745	5260	2.52	206406	98.84
775	2429	1.16	208835	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.

Hambleton, R.K., B.E. Clauser, K.M. Mazor, and R.W. Jones. 1993. Advances in the detection of differentially functioning test items. *European Journal of Psychological Assessment* 9 (1):1–18.

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, NY: 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 Measurement* 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: 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: 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: 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.