New York State Testing Program 2010: Mathematics, Grades 3–8

Technical Report

Submitted
2010

CTB/McGraw-Hill
Monterey, California 93940

Copyright

Table of Contents

List of Tables

Section I: Introduction and Overview

Introduction

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

Test Purpose

The NYSTP is an assessment system designed to measure concepts, processes, and skills taught in schools in New York 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 students' proficiency into one of four levels based on their test performance.

Target Population

Students in New York State public schools in 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 2010, nonpublic schools participated in all grade tests but were not well represented in the testing program. The New York State Education Department (NYSED) made a decision to exclude these schools from the data analyses in 2010. 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.p12.nysed.gov/osa/math/home.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 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 (Below Standards), Level II (Meets Basic Standards), Level III (Meets Proficiency Standards), and Level IV (Exceeds Proficiency Standards). The original proficiency cut scores used to distinguish among Levels I, II, III, and IV were established during the process of Standard Setting in 2006. In 2010, change in the test administration window between the 2008–2009 and 2009–2010 school years and a decision to align the proficiency standards with Grade 8 student performance on the NYS Regents Math A exams led to changes in the proficiency cut scores. The process of cut score adjustment after the 2010 OP test administration is described in detail in Section VII of this report.

Detailed information on a process of establishing original performance cut scores and their association with test content is provided in the *Bookmark Standard Setting Technical Report 2006 for Grades 3, 4, 5, 6, 7, and 8 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 uses 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 visually impaired students, large type and braille editions of the test books are provided. The students dictate and/or record their responses; the teachers transcribe student responses to multiple-choice (MC) questions onto scannable answer sheets; and the teachers transcribe the responses to constructed-response (CR) 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 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 programs.

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 Grades 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.p12.nysed.gov/osa/math/samplers/chinese/ (Chinese)
http://www.p12.nysed.gov/osa/math/samplers/haitian/ (Haitian-Creole)
http://www.p12.nysed.gov/osa/math/samplers/korean/ (Korean)
http://www.p12.nysed.gov/osa/math/samplers/russian/ (Russian)
http://www.p12.nysed.gov/osa/math/samplers/spanish/ (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 in May 2010 over a two-day period for Grades 3, 5, 6, 7, and 8 and over a three-day period for Grade 4. The tests were printed in black and white and incorporated the concepts of universal design. Copies of the OP tests are available online at http://www.nysedregents.org/elementary.html and http://www.nysedregents.org/intermediate.html. Details on the administration and scoring of these tests can be found in Section IV, "Test Administration and Scoring."

Test Configuration

The OP test books were administered, in order, on two to three consecutive days, depending on the grade. Table 1 provides information on the number and type of items in each book, as well as testing times. Book 1 contained only MC items. Book 2 and Book 3 contained only CR items. The 2010 *Teacher's Directions* (http://www.p12.nysed.gov/osa/ei/ directions/m3-5-td-10.pdf and http://www.p12.nysed.gov/osa/ei/directions/m6-8-td-10.pdf) as well as the 2010 *School Administrator's Manual* (http://www.p12.nysed.gov/osa/sam /math/mathei-sam-10.pdf) provide details on security, scheduling, classroom organization and preparation, test materials, and administration.

Table 1. NYSTP Mathematics 2010 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 2010 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	6	0	6	40	10
	2	3	0	6	6	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 2010 OP tests.

Table 2. NYSTP Mathematics 2010 Test Blueprint

Grade	Total Points	Content Strand	Target Points	Selected Points	Target % of Test	Selected % of Test
3	39	Number Sense and Operations	19	19	48.0	49.0
		Algebra	5	4	13.0	10.0
		Geometry	5	5	13.0	13.0
		Measurement	5	4	13.0	10.0
		Statistics and Probability	5	7	13.0	18.0

(Continued on next page)

Table 2. NYSTP Mathematics 2010 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	35	45.0	50.0
		Algebra	10	11	14.0	16.0
		Geometry	8	8	12.0	11.0
		Measurement	12	10	17.0	14.0
		Statistics and Probability	8	6	12.0	9.0
5	46	Number Sense and Operations	18	15	39.0	33.0
		Algebra	5	8	11.0	17.0
		Geometry	12	12	25.0	26.0
		Measurement	6	6	14.0	13.0
		Statistics and Probability	5	5	11.0	11.0
6	49	Number Sense and Operations	18	18	37.0	37.0
		Algebra	9	12	19.0	25.0
		Geometry	8	7	16.5	14.0
		Measurement	6	5	11.0	10.0
		Statistics and Probability	8	7	16.5	14.0
7	50	Number Sense and Operations	15	16	30.0	32.0
		Algebra	6	7	12.0	14.0
		Geometry	7	8	14.0	16.0
		Measurement	7	5	14.0	10.0
		Statistics and Probability	15	14	30.0	28.0
8	69	Number Sense and Operations	8	9	11.0	13.0
		Algebra	30	26	44.0	38.0
		Geometry	24	24	35.0	35.0
		Measurement	7	10	10.0	14.0

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 2010 Operational Test Map, Grade 3

Question	Type	Points	Strand	Content Performance Indicator	Answer Key
Book 1					
1	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)	D
2	Multiple Choice	1	Number Sense and Operations	3.N02 Read and write whole numbers to 1,000	C
3	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
4	Multiple Choice	1	Number Sense and Operations	3.N16 Identify odd and even numbers	C
5	Multiple Choice	1	Geometry	3.G03 Name, describe, compare, and sort three-dimensional shapes: cube, cylinder, sphere, prism, and cone	B
6	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	D
7	Multiple Choice	1	Measurement	3.M07 Count and represent combined coins and dollars, using currency symbols ($0.00)	C
8	Multiple Choice	1	Geometry	3.G02 Identify congruent and similar figures	A
9	Multiple Choice	1	Number Sense and Operations	3.N19 Develop fluency with single-digit multiplication facts	C
10	Multiple Choice	1	Number Sense and Operations	3.N21 Use the area model, tables, patterns, arrays, and doubling to provide meaning for multiplication	A
11	Multiple Choice	1	Number Sense and Operations	3.N08 Use the zero property of multiplication	A
12	Multiple Choice	1	Number Sense and Operations	3.N27 Check reasonableness of an answer by using estimation	C
13	Multiple Choice	1	Number Sense and Operations	3.N24 Develop strategies for selecting the appropriate computational and operational method in problem solving situations	D
14	Multiple Choice	1	Number Sense and Operations	3.N06 Use and explain the commutative property of addition and multiplication	B

(Continued on next page)

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

Question	Type	Points	Strand	Content Performance Indicator	Answer Key
Book 1 (continued)					
15	Multiple Choice	1	Algebra	3.A02 Describe and extend numeric (+, –) and geometric patterns	A
16	Multiple Choice	1	Number Sense and Operations	3.N11 Use manipulatives, visual models, and illustrations to name and represent unit fractions (1/2, 1/3, 1/4, 1/5, 1/6, and 1/10) as part of a whole or a set of objects	D
17	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)	B
18	Multiple Choice	1	Measurement	3.M09 Tell time to the minute, using digital and analog clocks	A
19	Multiple Choice	1	Statistics and Probability	3.S07 Read and interpret data in bar graphs and pictographs	D
20	Multiple Choice	1	Measurement	3.M01 Select tools and units (customary) appropriate for the length measured	B
21	Multiple Choice	1	Geometry	3.G05 Identify and construct lines of symmetry	D
22	Multiple Choice	1	Number Sense and Operations	3.N12 Understand and recognize the meaning of numerator and denominator in the symbolic form of a fraction	A
23	Multiple Choice	1	Number Sense and Operations	3.N07 Use 1 as the identity element for multiplication	B
24	Multiple Choice	1	Number Sense and Operations	3.N22 Demonstrate fluency and apply single-digit division facts	D
25	Multiple Choice	1	Statistics and Probability	3.S08 Formulate conclusions and make predictions from graphs	D
Book 2					
26	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
27	Short Response	2	Statistics and Probability	3.S05 Display data in pictographs and bar graphs	n/a
28	Short Response	2	Algebra	3.A02 Describe and extend numeric (+, –) and geometric patterns	n/a
29	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
30	Extended Response	3	Statistics and Probability	3.S07 Read and interpret data in bar graphs and pictographs	n/a
31	Extended Response	3	Number Sense and Operations	3.N18 Use a variety of strategies to add and subtract 3-digit numbers (with and without regrouping)	n/a

Table 3b. NYSTP Mathematics 2010 Operational Test Map, Grade 4

Question	Type	Points	Strand	Content Performance Indicator	Answer Key
Book 1					
1	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
2	Multiple Choice	1	Number Sense and Operations	4.N02 Read and write whole numbers to 10,000	D
3	Multiple Choice	1	Geometry	4.G02 Identify points and line segments when drawing a plane figure	C
4	Multiple Choice	1	Number Sense and Operations	4.N03 Compare and order numbers to 10,000	D
5	Multiple Choice	1	Number Sense and Operations	4.N26 Round numbers less than 1,000 to the nearest tens and hundreds	B
6	Multiple Choice	1	Number Sense and Operations	4.N12 Use concrete materials and visual models to compare and order decimals (less than 1) to the hundredths place in the context of money	D
7	Multiple Choice	1	Algebra	3.A01 Use the symbols <, >, = (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)	A
8	Multiple Choice	1	Number Sense and Operations	4.N15 Select appropriate computational and operational methods to solve problems	B
9	Multiple Choice	1	Geometry	4.G04 Find the area of a rectangle by counting the number of squares needed to cover the rectangle	C
10	Multiple Choice	1	Number Sense and Operations	4.N11 Read and write decimals to hundredths, using money as a context	B
11	Multiple Choice	1	Number Sense and Operations	3.N25 Estimate numbers up to 500	B
12	Multiple Choice	1	Number Sense and Operations	3.N14 Explore equivalent fractions (1/2, 1/3, and 1/4)	D
13	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	C
14	Multiple Choice	1	Measurement	4.M09 Calculate elapsed time in hours and half hours, not crossing A.M./P.M.	C
15	Multiple Choice	1	Geometry	3.G02 Identify congruent and similar figures	A
16	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
17	Multiple Choice	1	Number Sense and Operations	4.N13 Develop an understanding of the properties of odd/even numbers as a result of multiplication	A

(Continued on next page)

Table 3b. NYSTP Mathematics 2010 Operational Test Map, Grade 4 (cont.)

Question	Type	Points	Strand	Content Performance Indicator	Answer Key
Book 1 (continued)					
18	Multiple Choice	1	Number Sense and Operations	4.N27 Check reasonableness of an answer by using estimation	C
19	Multiple Choice	1	Measurement	4.M06 Select tools and units appropriate to the capacity being measured (milliliters and liters)	A
20	Multiple Choice	1	Measurement	4.M10 Calculate elapsed time in days and weeks, using a calendar	B
21	Multiple Choice	1	Number Sense and Operations	4.N08 Recognize and generate equivalent fractions (halves, fourths, thirds, fifths, sixths, and tenths) using manipulatives, visual models, and illustrations	D
22	Multiple Choice	1	Measurement	4.M01 Select tools and units (customary and metric) appropriate for the length measured	A
23	Multiple Choice	1	Statistics and Probability	4.S05 Develop and make predictions that are based on data	C
24	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)	C
25	Multiple Choice	1	Number Sense and Operations	4.N24 Express decimals as an equivalent form of fractions to tenths and hundredths	D
26	Multiple Choice	1	Algebra	4.A03 Find the value or values that will make an open sentence true, if it contains < or >	A
27	Multiple Choice	1	Statistics and Probability	4.S06 Formulate conclusions and make predictions from graphs	C
28	Multiple Choice	1	Number Sense and Operations	3.N26 Recognize real world situations in which an estimate (rounding) is more appropriate	B
29	Multiple Choice	1	Measurement	4.M04 Select tools and units appropriate to the mass of the object being measured (grams and kilograms)	C
30	Multiple Choice	1	Statistics and Probability	4.S04 Read and interpret line graphs	D
Book 2					
31	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
32	Short Response	2	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)	n/a

(Continued on next page)

Table 3b. NYSTP Mathematics 2010 Operational Test Map, Grade 4 (cont.)

Question	Type	Points	Strand	Content Performance Indicator	Answer Key
Book 2 (continued)					
33	Short Response	2	Measurement	4.M03 Know and understand equivalent standard units of length: 12 inches = 1 foot 3 feet = 1 yard	n/a
34	Short Response	2	Number Sense and Operations	4.N16 Understand various meanings of multiplication and division	n/a
35	Short Response	2	Number Sense and Operations	4.N17 Use multiplication and division as inverse operations to solve problems	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	Number Sense and Operations	4.N07 Develop an understanding of fractions as locations on number lines and as divisions of whole numbers	n/a
38	Extended Response	3	Statistics and Probability	4.S03 Represent data using tables, bar graphs, and pictographs	n/a
39	Extended Response	3	Algebra	4.A04 Describe, extend, and make generalizations about numeric (+, –, ×, ÷) and geometric patterns	n/a
Book 3					
40	Short Response	2	Number Sense and Operations	3.N20 Use a variety of strategies to solve multiplication problems with factors up to 12 x 12 4.N16 Understand various meanings of multiplication and division	n/a
41	Short Response	2	Algebra	4.A01 Evaluate and express relationships using open sentences with one operation	n/a
42	Short Response	2	Measurement	4.M08 Make change, using combined coins and dollar amounts	n/a
43	Short Response	2	Number Sense and Operations	4.N21 Use a variety of strategies to divide two-digit dividends by one-digit divisors (with and without remainders)	n/a
44	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
45	Short Response	2	Number Sense and Operations	4.N06 Understand, use, and explain the associative property of multiplication	n/a
46	Short Response	2	Number Sense and Operations	4.N22 Interpret the meaning of remainders	n/a

(Continued on next page)

Table 3b. NYSTP Mathematics 2010 Operational Test Map, Grade 4 (cont.)

Question	Type	Points	Strand	Content Performance Indicator	Answer Key
Book 3 (continued)					
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	4.N18 Use a variety of strategies to multiply two-digit numbers by one-digit numbers (with and without regrouping)	n/a

Table 3c. NYSTP Mathematics 2010 Operational Test Map, Grade 5

Question	Type	Points	Strand	Content Performance Indicator	Answer Key
Book 1					
1	Multiple Choice	1	Measurement	5.M02 Identify customary equivalent units of length	C
2	Multiple Choice	1	Geometry	5.G07 Know that the sum of the interior angles of a triangle is 180 degrees	A
3	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)	B
4	Multiple Choice	1	Number Sense and Operations	5.N02 Compare and order numbers to millions	D
5	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 *Note: Division by anything greater than a two-digit divisor should be done using technology.*	A
6	Multiple Choice	1	Geometry	5.G09 Identify pairs of congruent triangles	C
7	Multiple Choice	1	Number Sense and Operations	5.N16 Use a variety of strategies to multiply three-digit by three-digit numbers *Note: Multiplication by anything greater than a three-digit multiplier/multiplicand should be done using technology.*	C
8	Multiple Choice	1	Number Sense and Operations	5.N20 Convert improper fractions to mixed numbers, and mixed numbers to improper fractions	B

(Continued on next page)

Table 3c. NYSTP Mathematics 2010 Operational Test Map, Grade 5 (cont.)

Question	Type	Points	Strand	Content Performance Indicator	Answer Key
Book 1 (continued)					
9	Multiple Choice	1	Geometry	5.G05 Know that the sum of the interior angles of a quadrilateral is 360 degrees	C
10	Multiple Choice	1	Number Sense and Operations	5.N15 Find the common factors and the greatest common factor of two numbers	C
11	Multiple Choice	1	Statistics and Probability	4.S04 Read and interpret line graphs	B
12	Multiple Choice	1	Geometry	4.G08 Classify angles as acute, obtuse, right, and straight	D
13	Multiple Choice	1	Algebra	5.A04 Solve simple one-step equations using basic whole-number facts	C
14	Multiple Choice	1	Geometry	5.G04 Classify quadrilaterals by properties of their angles and sides	B
15	Multiple Choice	1	Algebra	5.A02 Translate simple verbal expressions into algebraic expressions	D
16	Multiple Choice	1	Number Sense and Operations	5.N01 Read and write whole numbers to millions	C
17	Multiple Choice	1	Number Sense and Operations	5.N22 Add and subtract mixed numbers with like denominators	A
18	Multiple Choice	1	Geometry	5.G02 Identify pairs of similar triangles	A
19	Multiple Choice	1	Number Sense and Operations	4.N25 Add and subtract decimals to tenths and hundredths using a hundreds chart	C
20	Multiple Choice	1	Algebra	5.A03 Substitute assigned values into variable expressions and evaluate using order of operations	A
21	Multiple Choice	1	Number Sense and Operations	4.N23 Add and subtract proper fractions with common denominators	D
22	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	C
23	Multiple Choice	1	Number Sense and Operations	5.N10 Compare decimals using <, >, or =	D
24	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 a ruler, measuring cup, etc.*	C
25	Multiple Choice	1	Geometry	5.G10 Identify corresponding parts of congruent triangles	D
26	Multiple Choice	1	Number Sense and Operations	5.N08 Read, write, and order decimals to thousandths	D

(Continued on next page)

Table 3c. NYSTP Mathematics 2010 Operational Test Map, Grade 5 (cont.)

Question	Type	Points	Strand	Content Performance Indicator	Answer Key
Book 2					
27	Short Response	2	Algebra	5.A06 Evaluate the perimeter formula for given input values	n/a
28	Short Response	2	Measurement	5.M08 Measure and draw angles using a protractor	n/a
29	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
30	Short Response	2	Geometry	5.G08 Find a missing angle when given two angles of a triangle	n/a
31	Extended Response	3	Number Sense and Operations	5.N11 Understand that percent means part of 100, and write percents as fractions and decimals	n/a
32	Extended Response	3	Geometry	5.G01 Calculate the perimeter of regular and irregular polygons	n/a
33	Extended Response	3	Statistics and Probability	5.S02 Display data in a line graph to show an increase or decrease over time	n/a
34	Extended Response	3	Measurement	5.M07 Calculate elapsed time in hours and minutes	n/a

Table 3d. NYSTP Mathematics 2010 Operational Test Map, Grade 6

Question	Type	Points	Strand	Content Performance Indicator	Answer Key
Book 1					
1	Multiple Choice	1	Number Sense and Operations	6.N07 Express equivalent ratios as a proportion	C
2	Multiple Choice	1	Number Sense and Operations	6.N23 Represent repeated multiplication in exponential form	C
3	Multiple Choice	1	Algebra	5.A04 Solve simple one-step equations using basic whole-number facts	B
4	Multiple Choice	1	Number Sense and Operations	6.N13 Define absolute value and determine the absolute value of rational numbers (including positive and negative)	D
5	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)	C
6	Multiple Choice	1	Number Sense and Operations	6.N05 Define and identify the zero property of multiplication	A
7	Multiple Choice	1	Geometry	6.G05 Identify radius, diameter, chords, and central angles of a circle	C

(Continued on next page)

Table 3d. NYSTP Mathematics 2010 Operational Test Map, Grade 6 (cont.)

Question	Type	Points	Strand	Content Performance Indicator	Answer Key
8	Multiple Choice	1	Algebra	6.A05 Solve simple proportions within context	B
9	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
10	Multiple Choice	1	Geometry	6.G10 Identify and plot points in all four quadrants	D
11	Multiple Choice	1	Statistics and Probability	6.S05 Determine the mean, mode, and median for a given set of data	C
12	Multiple Choice	1	Algebra	6.A03 Translate two-step verbal sentences into algebraic equations	D
13	Multiple Choice	1	Measurement	6.M05 Identify equivalent metric units of capacity (milliliter to liter and liter to milliliter)	C
14	Multiple Choice	1	Algebra	6.A06 Evaluate formulas for given input values (circumference, area, volume, distance, temperature, interest, etc.)	A
15	Multiple Choice	1	Number Sense and Operations	6.N22 Evaluate numerical expressions using order of operations (may include exponents of two and three)	C
16	Multiple Choice	1	Number Sense and Operations	6.N16 Add and subtract fractions with unlike denominators	A
17	Multiple Choice	1	Statistics and Probability	6.S07 Read and interpret graphs	C
18	Multiple Choice	1	Measurement	6.M03 Identify equivalent customary units of capacity (cups to pints, pints to quarts, and quarts to gallons)	B
19	Multiple Choice	1	Number Sense and Operations	6.N25 Evaluate expressions having exponents where the power is an exponent of one, two, or three	C
20	Multiple Choice	1	Number Sense and Operations	6.N12 Solve percent problems involving percent, rate, and base	A
21	Multiple Choice	1	Statistics and Probability	5.S06 Record experiment results using fractions/ratios	B
22	Multiple Choice	1	Geometry	6.G07 Determine the area and circumference of a circle, using the appropriate formula	D
23	Multiple Choice	1	Geometry	6.G06 Understand the relationship between the diameter and radius of a circle	C
24	Multiple Choice	1	Algebra	6.A01 Translate two-step verbal expressions into algebraic expressions	D

(Continued on next page)

Table 3d. NYSTP Mathematics 2010 Operational Test Map, Grade 6 (cont.)

Question	Type	Points	Strand	Content Performance Indicator	Answer Key
Book 1 (continued)					
25	Multiple Choice	1	Measurement	6.M03 Identify equivalent customary units of capacity (cups to pints, pints to quarts, and quarts to gallons)	B
Book 2					
26	Short Response	2	Algebra	5.A05 Solve and explain simple one-step equations using inverse operations involving whole numbers	n/a
27	Short Response	2	Measurement	6.M01 Measure capacity and calculate volume of a rectangular prism	n/a
28	Short Response	2	Number Sense and Operations	6.N10 Verify the proportionality using the product of the means equals the product of the extremes	n/a
29	Short Response	2	Geometry	5.G12 Identify and plot points in the first quadrant	n/a
30	Short Response	2	Number Sense and Operations	6.N26 Estimate a percent of quantity (0% to 100%)	n/a
31	Short Response	2	Algebra	5.A05 Solve and explain simple one-step equations using inverse operations involving whole numbers	n/a
32	Extended Response	3	Algebra	5.A05 Solve and explain simple one-step equations using inverse operations involving whole numbers	n/a
33	Extended Response	3	Number Sense and Operations	6.N09 Solve proportions using equivalent fractions	n/a
34	Extended Response	3	Number Sense and Operations	6.N03 Define and identify the distributive property of multiplication over addition	n/a
35	Extended Response	3	Statistics and Probability	6.S07 Read and interpret graphs	n/a

Table 3e. NYSTP Mathematics 2010 Operational Test Map, Grade 7

Question	Type	Points	Strand	Content Performance Indicator	Answer Key
Book 1					
1	Multiple Choice	1	Measurement	7.M04 Convert mass within a given system	C
2	Multiple Choice	1	Number Sense and Operations	7.N13 Add and subtract two integers (with and without the use of a number line)	A
3	Multiple Choice	1	Statistics and Probability	7.S04 Calculate the range for a given set of data	B
4	Multiple Choice	1	Algebra	7.A03 Identify a polynomial as an algebraic expression containing one or more terms	C
5	Multiple Choice	1	Statistics and Probability	6.S10 Determine the probability of dependent events	A
6	Multiple Choice	1	Algebra	7.A02 Add and subtract monomials with exponents of one	C
7	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.*	C
8	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)	A
9	Multiple Choice	1	Number Sense and Operations	7.N12 Add, subtract, multiply, and divide integers	A
10	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	D
11	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)	A
12	Multiple Choice	1	Algebra	6.A05 Solve simple proportions within context	B
13	Multiple Choice	1	Measurement	7.M03 Identify customary and metric units of mass	A
14	Multiple Choice	1	Statistics and Probability	7.S12 Compare actual results to predicted results	C
15	Multiple Choice	1	Geometry	7.G10 Graph the solution set of an inequality (positive coefficients only) on a number line	A

Continued on next page)

Table 3e. NYSTP Mathematics 2010 Operational Test Map, Grade 7 (cont.)

Question	Type	Points	Strand	Content Performance Indicator	Answer Key
Book 1 (continued)					
16	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	D
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)	A
18	Multiple Choice	1	Number Sense and Operations	7.N09 Determine multiples and least common multiple of two or more numbers	C
19	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	A
20	Multiple Choice	1	Algebra	6.A04 Solve and explain two-step equations involving whole numbers using inverse operations	A
21	Multiple Choice	1	Statistics and Probability	7.S10 Predict the outcome of an experiment	C
22	Multiple Choice	1	Statistics and Probability	7.S08 Interpret data to provide the basis for predictions and to establish experimental probabilities	B
23	Multiple Choice	1	Number Sense and Operations	7.N06 Translate numbers from scientific notation into standard form	C
24	Multiple Choice	1	Number Sense and Operations	7.N08 Find the common factors and greatest common factor of two or more numbers	B
25	Multiple Choice	1	Number Sense and Operations	7.N03 Place rational and irrational numbers (approximations) on a number line and justify the placement of the numbers	A
26	Multiple Choice	1	Measurement	7.M09 Determine the tool and technique to measure with an appropriate level of precision: mass	B
27	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

Continued on next page)

Table 3e. NYSTP Mathematics 2010 Operational Test Map, Grade 7 (cont.)

Question	Type	Points	Strand	Content Performance Indicator	Answer Key
Book 1 (continued)					
28	Multiple Choice	1	Number Sense and Operations	7.N18 Identify the two consecutive whole numbers between which the square root of a non-perfect square whole number less than 225 lies (with and without the use of a number line)	C
29	Multiple Choice	1	Number Sense and Operations	7.N02 Recognize the difference between rational and irrational numbers (e.g., explore different approximations of π)	A
30	Multiple Choice	1	Statistics and Probability	7.S09 Determine the validity of sampling methods to predict outcomes	D
Book 2					
31	Short Response	2	Algebra	7.A06 Evaluate formulas for given input values (surface area, rate, and density problems)	n/a
32	Short Response	2	Geometry	7.G09 Determine whether a given triangle is a right triangle by applying the Pythagorean Theorem and using a calculator	n/a
33	Short Response	2	Statistics and Probability	6.S09 List possible outcomes for compound events	n/a
34	Short Response	2	Measurement	7.M08 Draw central angles in a given circle using a protractor (circle graphs)	n/a
35	Extended Response	3	Statistics and Probability	6.S02 Record data in a frequency table	n/a
36	Extended Response	3	Number Sense and Operations	7.N10 Determine the prime factorization of a given number and write in exponential form	n/a
37	Extended Response	3	Number Sense and Operations	7.N19 Justify the reasonableness of answers using estimation	n/a
38	Extended Response	3	Geometry	7.G04 Determine the surface area of prisms and cylinders, using a calculator and a variety of methods	n/a

Table 3f. NYSTP Mathematics 2010 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	B
2	Multiple Choice	1	Geometry	8.G05 Calculate the missing angle measurements when given two parallel lines cut by a transversal	D
3	Multiple Choice	1	Algebra	8.A02 Write verbal expressions that match given mathematical expressions	B
4	Multiple Choice	1	Geometry	8.G03 Calculate the missing angle in a supplementary or complementary pair	B
5	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
6	Multiple Choice	1	Geometry	7.G05 Identify the right angle, hypotenuse, and legs of a right triangle	D
7	Multiple Choice	1	Geometry	8.G03 Calculate the missing angle in a supplementary or complementary pair	A
8	Multiple Choice	1	Geometry	7.G08 Use the Pythagorean Theorem to determine the unknown length of a side of a right triangle	B
9	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	A
10	Multiple Choice	1	Algebra	8.A07 Add and subtract polynomials (integer coefficients)	B
11	Multiple Choice	1	Geometry	8.G05 Calculate the missing angle measurements when given two parallel lines cut by a transversal	D
12	Multiple Choice	1	Geometry	8.G07 Describe and identify transformations in the plane, using proper function notation (rotations, reflections, translations, and dilations)	D
13	Multiple Choice	1	Number Sense and Operations	8.N05 Estimate a percent of quantity, given an application	B
14	Multiple Choice	1	Geometry	8.G04 Determine angle pair relationships when given two parallel lines cut by a transversal	B
15	Multiple Choice	1	Geometry	8.G12 Identify the properties preserved and not preserved under a reflection, rotation, translation, and dilation	C

Continued on next page)

Table 3f. NYSTP Mathematics 2010 Operational Test Map, Grade 8 (cont.)

Question	Type	Points	Strand	Content Performance Indicator	Answer Key
Book 1 (continued)					
16	Multiple Choice	1	Algebra	8.A12 Apply algebra to determine the measure of angles formed by or contained in parallel lines cut by a transversal and by intersecting lines	D
17	Multiple Choice	1	Algebra	8.A06 Multiply and divide monomials	A
18	Multiple Choice	1	Measurement	7.M01 Calculate distance using a map scale	C
19	Multiple Choice	1	Algebra	7.A03 Identify a polynomial as an algebraic expression containing one or more terms	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	7.A03 Identify a polynomial as an algebraic expression containing one or more terms	D
22	Multiple Choice	1	Algebra	8.A10 Factor algebraic expressions using the GCF	B
23	Multiple Choice	1	Measurement	7.M01 Calculate distance using a map scale	C
24	Multiple Choice	1	Algebra	7.A10 Write an equation to represent a function from a table of values	D
25	Multiple Choice	1	Algebra	8.A01 Translate verbal sentences into algebraic inequalities	D
26	Multiple Choice	1	Algebra	8.A09 Divide a polynomial by a monomial (integer coefficients) *Note: The degree of the denominator is less than or equal to the degree of the numerator for all variables.*	A
27	Multiple Choice	1	Number Sense and Operations	8.N04 Apply percents to: Tax; Percent increase/decrease; Simple interest; Sale price; Commission; Interest rates; Gratuities	B
Book 2					
28	Short Response	2	Number Sense and Operations	8.N04 Apply percents to: Tax; Percent increase/decrease; Simple interest; Sale price; Commission; Interest rates; Gratuities	n/a
29	Short Response	2	Algebra	7.A10 Write an equation to represent a function from a table of values	n/a
30	Short Response	2	Geometry	8.G03 Calculate the missing angle in a supplementary or complementary pair	n/a

Continued on next page)

Table 3f. NYSTP Mathematics 2010 Operational Test Map, Grade 8 (cont.)

Question	Type	Points	Strand	Content Performance Indicator	Answer Key
Book 2 (continued)					
31	Short Response	2	Measurement	8.M01 Solve equations/proportions to convert to equivalent measurements within metric and customary measurement systems *Note: Also allow Fahrenheit to Celsius and vice versa.*	n/a
32	Short Response	2	Algebra	7.A08 Create algebraic patterns using charts/tables, graphs, equations, and expressions	n/a
33	Short Response	2	Geometry	8.G09 Draw the image of a figure under a reflection over a given line	n/a
Book 3					
34	Short Response	2	Geometry	8.G02 Identify pairs of supplementary and complementary angles	n/a
35	Short Response	2	Number Sense and Operations	8.N01 Develop and apply the laws of exponents for multiplication and division	n/a
36	Short Response	2	Algebra	8.A07 Add and subtract polynomials (integer coefficients)	n/a
37	Extended Response	3	Geometry	8.G01 Identify pairs of vertical angles as congruent	n/a
38	Extended Response	3	Measurement	7.M06 Compare unit prices	n/a
39	Short Response	2	Geometry	7.G09 Determine whether a given triangle is a right triangle by applying the Pythagorean Theorem and using a calculator	n/a
40	Short Response	2	Algebra	8.A06 Multiply and divide monomials	n/a
41	Extended Response	3	Measurement	7.M07 Convert money between different currencies with the use of an exchange rate table and a calculator	n/a
42	Extended Response	3	Geometry	8.G08 Draw the image of a figure under rotations of 90 and 180 degrees	n/a
43	Extended Response	3	Algebra	8.A16 Find a set of ordered pairs to satisfy a given linear numerical pattern (expressed algebraically); then plot the ordered pairs and draw the line	n/a
44	Extended Response	3	Number Sense and Operations	8.N06 Justify the reasonableness of answers using estimation	n/a
45	Short Response	2	Geometry	8.G02 Identify pairs of supplementary and complementary angles	n/a

Table 4. NYSTP Mathematics 2010 Strand Coverage

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

(Continued on next page)

Table 4. NYSTP Mathematics 2010 Strand Coverage (cont.)

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

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, and special education and bilingual instructors also 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, were 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 ten participants per grade band. Upon completion of the committee meetings, NYSED reviewed the committees' determinations and approved them, with minor adjustments when necessary, to maintain consistency across the grades. 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 ten participants per grade band. The committee members were provided with the items, the New York State Learning Standards, and the test specifications, and they considered the following elements as they reviewed the test items:

- the accuracy and grade-level appropriateness of the items
- the mapping of the items to the assigned performance indicators
- the accompanying exemplary responses (CR items)
- the appropriateness of the correct responses and distractors (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 16, 2009. 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 2009 in New York State to examine a sampling of student responses to the short- and extended-response items. Committees of New York State educators with content and grade-level expertise were again assembled. Each committee was composed of approximately eight to ten participants per grade level. CTB/McGraw-Hill staff facilitated the meetings, and NYSED staff reviewed the decisions made by the committees and verified that the decisions made were consistent across grades. The committees' charge was to select student responses that exemplified each score point of each CR item. These responses, in conjunction with the 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 fifth year of Grades 3–8 Mathematics OP Tests were administered in May 2010. 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 and reviewed by psychometricians from CTB/McGraw-Hill and 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 (Green, Yen, & 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 2009 to finalize item selection and test creation. There, they discussed the content and data of the proposed selections, explored alternate selections for consideration, determined the final item selections, and ordered those items (assigned positions) in the OP test books. The final test forms were approved by the final-eyes committee that consisted of approximately 20 participants across all grade levels. After approval by NYSED, the tests were produced and administered in May 2010.

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 site:

- http://www.p12.nysed.gov/osa/math/home.html#ei

Proficiency and Performance Standards

A change in the test administration window between the 2008–2009 and 2009–2010 school years and a decision to align the proficiency standards with Grade 8 student performance on the NYS Regents Math A exams led to changes in the proficiency cut scores after the 2010 test administration. The results were reviewed by the NYS Technical Advisory Group and were approved by the Board of Regents in July 2010. For each grade level, there are four proficiency levels. Three cut points demarcate the performance standards needed to demonstrate each ascending level of proficiency.

Section III: Validity

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

The American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) *Standards for Educational and Psychological Testing* (1999) address 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 of 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 & Lehmann, 1991). In addition, tests are now also used for the purposes 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 AERA/APA/NCME (1999) 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 VIII, "Reliability and Standard Error of Measurement." For the total populations, the reliability coefficients (Cronbach's alpha) ranged from 0.88–0.94, and for all subgroups, the reliability coefficients are greater than 0.80. 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 5 Mathematics Tests displayed good item-model fit. Two items in Grade 4, one item in Grade 6, one item in Grade 7, and three 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 the 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 principal components accounting for more than 20% 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**	**7.50**	**24.21**	**24.21**
	2	1.47	4.75	28.96
	3	1.01	3.26	32.22
4	**1**	**12.64**	**26.34**	**26.34**
	2	1.51	3.14	29.48
	3	1.14	2.38	31.86
	4	1.00	2.09	33.95
5	**1**	**8.44**	**24.83**	**24.83**
	2	1.17	3.43	28.26
	3	1.02	2.99	31.25
6	**1**	**8.65**	**24.71**	**24.71**
	2	1.35	3.86	28.57
	3	1.13	3.23	31.80
7	**1**	**8.95**	**23.56**	**23.56**
	2	1.59	4.20	27.76
	3	1.16	3.04	30.80
	4	1.03	2.71	33.51
8	**1**	**14.09**	**31.32**	**31.32**
	2	1.43	3.18	34.50
	3	1.12	2.49	36.99

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. The ELL/SUA subgroup is defined as examinees whose ELL status are true and use one or more ELL-related accommodation. The SWD/SUA subgroup includes examinees who are classified with disabilities and use one or more disability-related accommodations.

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 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 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, 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 & 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 (2010) located at http://www.p12.nysed.gov/osa/ei/ssom-10.pdf.

Test Administration

NYSTP Grades 3–8 Mathematics Tests were administered at the classroom level during May 2010. The testing window for Grades 3–8 (including the makeup test administration) was May 5–14, 2010. The makeup test administration 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 tests 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 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 the scoring committee members 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 member" included scorers at every site.

Scoring Models

For the 2009–10 school year, schools and school districts used local decision-making processes to select the model that best met their needs for the scoring of the Grades 3–8 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 schools' 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 schools' 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 schools' test papers included staff from all schools administering this test in a district, provided at least three schools were represented.

4. Two schools within a district—The first readers for the schools' test papers included staff from all schools administering this test in a district, provided that two schools were represented.

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

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

For further information, refer to the following link for a brief comparison between regional, district, and local scoring: http://www.p12.nysed.gov/osa/ei/ssom-10.pdf (see 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 were selected on the basis of their handscoring experience along with their educational and professional backgrounds.

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

Scoring was conducted with pen-and-pencil scoring as opposed to electronic scoring, and each scoring committee member evaluated actual student papers instead of electronically scanned papers. All scoring committee members were trained by previously trained and approved trainers along with guidance from scoring guides, the Mathematics Frequently Asked Questions (FAQs) document, and a CD containing audio files that highlighted important elements of the scoring guides. Each test book was scored by three separate scoring committee members, who scored three distinct sections of the test book. After test books were completed, the table facilitator or 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 and audio files to train scoring committee members on the criteria for scoring CR items. Part of the training process was the administration of a consistency assurance set (CAS) that provided the State's scoring sites with information regarding strengths and weaknesses of their scorers. This tool 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 book after reviewing the scoring guides and audio files, 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 email. When a member of the staff was unable to resolve an issue, they deferred to NYSED for a scoring decision. A quality check was also performed on each completed box of scored tests to certify that all questions were scored and that the scoring committee members darkened each score on the answer document appropriately. The log of calls received by the scoring Helpline was delivered to NYSED after the scoring window. To affirm that all schools across the state adhered to scoring guidelines and policies, approximately 5% of the schools' 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 June 2010. These data were used for all data analyses. Phase 2 involved submitting “straggler files” to CTB/McGraw-Hill in late June 2010. The straggler files contained less than 2% of the total population cases and were excluded from research data analyses due to late submission. Nonpublic school data were also 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 the results of the checking. CTB/McGraw-Hill Research performed data cleaning on 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 6 students who were administered an incorrect version of the Grade 6 test were rescored in Grade 6 data files (refer to the “Item Rescoring” subsection 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		197658
Out of grade	92	197566
No grade	0	197566
Duplicate record	0	197566
Non-public and out-of-district schools	3295	194271
Missing values for ALL items on OP form	3	194268
Out-of-range CR scores	0	194268

Table 6b. NYSTP Mathematics Data Cleaning, Grade 4

Exclusion Rule	# Deleted	# Cases Remain
Initial N		210695
Out of grade	87	210608
No grade	0	210608
Duplicate record	0	210608
Non-public and out-of-district schools	12948	197660
Missing values for ALL items on OP form	1	197659
Out-of-range CR scores	0	197659

Table 6c. NYSTP Mathematics Data Cleaning, Grade 5

Exclusion Rule	# Deleted	# Cases Remain
Initial N		199477
Out of grade	35	199442
No grade	0	199442
Duplicate record	0	199442
Non-public and out-of-district schools	3199	196243
Missing values for ALL items on OP form	0	196243
Out-of-range CR scores	0	196243

Table 6d. NYSTP Mathematics Data Cleaning, Grade 6

Exclusion Rule	# Deleted	# Cases Remain
Initial N		207815
Out of grade	158	207657
No grade	0	207657
Duplicate record	0	207657
Non-public and out-of-district schools	10633	197024
Missing values for ALL items on OP form	1	197023
Out-of-range CR scores	0	197023

Table 6e. NYSTP Mathematics Data Cleaning, Grade 7

Exclusion Rule	# Deleted	# Cases Remain
Initial N		201832
Out of grade	199	201633
No grade	0	201633
Duplicate record	0	201633
Non-public and out-of-district schools	3113	198520
Missing values for ALL items on OP form	4	198516
Out-of-range CR scores	0	198516

Table 6f. NYSTP Mathematics Data Cleaning, Grade 8

Exclusion Rule	# Deleted	# Cases Remain
Initial N		215226
Out of grade	177	215049
No grade	0	215049
Duplicate record	0	215049
Non-public and out-of-district schools	11940	203109
Missing values for ALL items on OP form	3	203106
Out-of-range CR scores	0	203106

Classical Analysis and Calibration Sample Characteristics

The demographic characteristics of students in the classical analysis and calibration sample data sets are presented in the following tables. The needs resource code (NRC) is assigned at the 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 = 194268)

Demographic Category		N-count	% of Total N-count
NRC	NYC	71308	36.82
	Big cities	8481	4.38
	Urban/Suburban	15664	8.09
	Rural	11185	5.78
	Average needs	56573	29.22
	Low needs	26325	13.59
	Charter	4106	2.12
Ethnicity	Asian	15407	7.93
	Black	36854	18.97
	Hispanic	43849	22.57
	American Indian	943	0.49
	Multi-Racial	1096	0.56
	Unknown	123	0.06
	White	95996	49.41
ELL	No	177853	91.55
	Yes	16415	8.45
SWD	No	166583	85.75
	Yes	27685	14.25
SUA	No	145986	75.15
	Yes	48282	24.85

Table 7b. Grade 4 Sample Characteristics (N = 197659)

Demographic Category		N-count	% of Total N-count
NRC	NYC	72002	36.54
	Big cities	8070	4.10
	Urban/Suburban	15939	8.09
	Rural	11333	5.75
	Average needs	58313	29.59
	Low needs	27941	14.18
	Charter	3444	1.75
Ethnicity	Asian	16530	8.36
	Black	37517	18.98
	Hispanic	43174	21.84
	American Indian	922	0.47
	Multi-Racial	983	0.50
	Unknown	111	0.06
	White	98422	49.79
ELL	No	183042	92.60
	Yes	14617	7.40

(Continued on next page)

Table 7b. Grade 4 Sample Characteristics (N = 197659) (cont.)

Demographic Category		N-count	% of Total N-count
SWD	No	168506	85.25
	Yes	29153	14.75
SUA	No	148242	75.00
	Yes	49417	25.00

Table 7c. Grade 5 Sample Characteristics (N = 196243)

Demographic Category		N-count	% of Total N-count
NRC	NYC	69360	35.46
	Big cities	7964	4.07
	Urban/Suburban	15256	7.80
	Rural	11280	5.77
	Average needs	58576	29.94
	Low needs	28675	14.66
	Charter	4501	2.30
Ethnicity	Asian	15543	7.92
	Black	37519	19.12
	Hispanic	42539	21.68
	American Indian	916	0.47
	Multi-Racial	857	0.44
	Unknown	97	0.05
	White	98772	50.33
ELL	No	184600	94.07
	Yes	11643	5.93
SWD	No	166383	84.78
	Yes	29860	15.22
SUA	No	148283	75.56
	Yes	47960	24.44

Table 7d. Grade 6 Sample Characteristics (N = 197023)

Demographic Category		N-count	% of Total N-count
NRC	NYC	68874	35.09
	Big cities	7655	3.90
	Urban/Suburban	15208	7.75
	Rural	11247	5.73
	Average needs	60308	30.73
	Low needs	29210	14.88
	Charter	3780	1.93

(Continued on next page)

Table 7d. Grade 6 Sample Characteristics (N = 197023) (cont.)

Demographic Category		N-count	% of Total N-count
Ethnicity	Asian	15444	7.84
	Black	37832	19.20
	Hispanic	41963	21.30
	American Indian	959	0.49
	Multi-Racial	764	0.39
	White	101	0.05
	Unknown	99960	50.74
ELL	No	187263	95.05
	Yes	9760	4.95
SWD	No	166865	84.69
	Yes	30158	15.31
SUA	No	153045	77.68
	Yes	43978	22.32

Table 7e. Grade 7 Sample Characteristics (N = 198516)

Demographic Category		N-count	% of Total N-count
NRC	NYC	69861	35.34
	Big cities	7704	3.90
	Urban/Suburban	15142	7.66
	Rural	11439	5.79
	Average needs	61699	31.21
	Low needs	28959	14.65
	Charter	2879	1.46
Ethnicity	Asian	15640	7.88
	Black	38066	19.18
	Hispanic	41663	20.99
	American Indian	955	0.48
	Multi-Racial	719	0.36
	Unknown	76	0.04
	White	101397	51.08
ELL	No	189756	95.59
	Yes	8760	4.41
SWD	No	168735	85.00
	Yes	29781	15.00
SUA	No	155982	78.57
	Yes	42534	21.43

Table 7f. Grade 8 Sample Characteristics (N = 203106)

Demographic Category		N-count	% of Total N-count
NRC	NYC	72646	35.95
	Big cities	7646	3.78
	Urban/Suburban	15110	7.48
	Rural	11724	5.80
	Average needs	62379	30.87
	Low needs	30200	14.95
	Charter	2347	1.16
Ethnicity	Asian	16061	7.91
	Black	38191	18.80
	Hispanic	42569	20.96
	American Indian	922	0.45
	Multi-Racial	604	0.30
	Unknown	94	0.05
	White	104665	51.53
ELL	No	194666	95.84
	Yes	8440	4.16
SWD	No	173071	85.21
	Yes	30035	14.79
SUA	No	160244	78.90
	Yes	42862	21.10

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 (e.g., 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 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 Section III, "Validity," and Section VIII, "Reliability and Standard Error of Measurement").

Item Rescoring

One item in the Grade 6 Spanish language version was rescored during the data analysis. In item 11 of the Spanish language version of the Grade 6 Mathematics Test, the phrase *median* was translated as *media* and should have been translated as *mediana*. To adjust for this, any student who used the Spanish language version was given credit for either choice A or choice B for this item.

Item Difficulty and Response Distribution

Item difficulty and response distribution tables (Tables 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 a valid score on the item, from zero to the maximum score.

Item difficulty is classically measured by the p-value statistic. It assesses the proportion of students who responded correctly for each MC item or the average proportion of the maximum score that students earned on each CR item. It is important to have a good range of p-values to increase test information and to avoid floor or ceiling effects. Generally, p-values should range between 0.30 and 0.90. P-values represent the overall degree of difficulty, but do not account for demonstrated student performance on other test items. Usually, p-value information is coupled with point biserial (pbis) statistics to verify that items are functioning as intended. (Point biserials are discussed in the next subsection.) Item difficulties (p-values) on the tests ranged from 0.23 to 0.97. For Grade 3, the item p-values were between 0.68 and 0.97 with a mean of 0.85. For Grade 4, the item p-values were between 0.43 and 0.96 with a mean of 0.76. For Grade 5, the item p-values were between 0.56 and 0.94 with a mean of 0.76. For Grade 6, the item p-values were between 0.37 and 0.92 with a mean of 0.74. For Grade 7, the item p-values were between 0.23 and 0.91 with a mean of 0.69. For Grade 8, the item p-values were between 0.51 and 0.96 with a mean of 0.72. These statistics are provided in Tables 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	194133	0.93	0.03	0.00	1.82	3.62	1.54	92.95	-0.16	-0.22	-0.21	0.35*	0.35
2	194123	0.94	0.04	0.00	0.49	4.34	93.91	1.19	-0.16	-0.22	0.34*	-0.23	0.34
3	193943	0.69	0.10	0.00	20.77	5.45	5.22	68.40	-0.43	-0.10	-0.09	0.47*	0.47
4	194002	0.80	0.08	0.00	8.19	3.92	80.31	7.44	-0.25	-0.30	0.48*	-0.23	0.48
5	194022	0.92	0.05	0.00	1.28	92.37	2.20	4.02	-0.19	0.35*	-0.20	-0.20	0.35
6	193997	0.87	0.07	0.00	6.75	2.23	4.15	86.74	-0.31	-0.29	-0.17	0.46*	0.46
7	194050	0.87	0.08	0.00	3.67	6.12	86.70	3.40	-0.24	-0.24	0.46*	-0.27	0.46
8	194054	0.71	0.06	0.00	70.56	4.63	14.21	10.50	0.35*	-0.20	-0.16	-0.20	0.35

(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	194034	0.76	0.09	0.00	13.29	6.02	76.37	4.20	-0.42	-0.19	0.53*	-0.17	0.53
10	194080	0.93	0.07	0.00	92.67	2.12	1.13	3.97	0.42*	-0.21	-0.21	-0.28	0.42
11	194079	0.97	0.05	0.00	96.45	1.37	0.72	1.36	0.33*	-0.17	-0.13	-0.26	0.33
12	194033	0.75	0.09	0.00	18.54	3.34	75.01	2.99	-0.33	-0.27	0.50*	-0.23	0.50
13	193873	0.75	0.11	0.00	8.65	12.73	3.83	74.58	-0.23	-0.12	-0.31	0.38*	0.38
14	194018	0.84	0.08	0.00	4.11	84.19	8.68	2.88	-0.29	0.28*	-0.09	-0.12	0.28
15	193936	0.95	0.09	0.00	94.55	1.93	1.40	1.95	0.35*	-0.21	-0.18	-0.19	0.35
16	193900	0.93	0.10	0.00	1.49	1.81	3.62	92.89	-0.25	-0.14	-0.24	0.37*	0.37
17	194042	0.87	0.08	0.00	7.47	87.11	2.65	2.66	-0.27	0.42*	-0.21	-0.22	0.42
18	194058	0.90	0.06	0.00	90.04	1.06	7.34	1.46	0.37*	-0.17	-0.26	-0.20	0.37
19	193947	0.92	0.11	0.00	1.52	4.31	2.28	91.72	-0.19	-0.40	-0.19	0.49*	0.49
20	193939	0.94	0.11	0.00	1.96	93.50	1.01	3.37	-0.16	0.31*	-0.16	-0.19	0.31
21	193694	0.95	0.13	0.00	1.48	1.61	1.41	95.21	-0.20	-0.16	-0.16	0.31*	0.31
22	193982	0.95	0.12	0.00	94.97	2.50	1.36	1.03	0.36*	-0.19	-0.23	-0.19	0.36
23	193900	0.94	0.15	0.00	3.27	93.84	1.19	1.52	-0.22	0.43*	-0.27	-0.26	0.43
24	193692	0.81	0.26	0.00	6.30	5.75	7.38	80.27	-0.26	-0.17	-0.17	0.38*	0.38
25	193049	0.80	0.59	0.00	7.36	4.87	7.51	79.63	-0.37	-0.20	-0.23	0.51*	0.51
26	194160	0.90	0.06	7.01	6.45	86.49							
27	194156	0.87	0.06	7.48	10.16	82.31							
28	194112	0.86	0.08	3.32	20.78	75.82							
29	194154	0.70	0.06	16.27	27.78	55.89							
30	194168	0.80	0.05	1.55	21.31	12.17	64.91						
31	194024	0.69	0.13	5.64	28.94	17.28	48.02						

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	197576	0.74	0.02	0.00	2.82	73.77	22.01	1.35	-0.16	0.35*	-0.26	-0.15	0.35
2	197555	0.96	0.03	0.00	1.07	0.59	2.17	96.11	-0.12	-0.14	-0.17	0.25*	0.25
3	197525	0.93	0.03	0.00	0.47	4.13	93.24	2.10	-0.09	-0.19	0.26*	-0.14	0.26
4	197480	0.93	0.04	0.00	1.78	3.01	1.93	93.19	-0.24	-0.16	-0.20	0.35*	0.35
5	197472	0.83	0.04	0.00	2.41	82.45	9.49	5.56	-0.21	0.39*	-0.24	-0.20	0.39
6	197427	0.88	0.05	0.00	3.75	1.09	7.35	87.70	-0.24	-0.18	-0.25	0.40*	0.40
7	197519	0.93	0.04	0.00	92.62	0.95	3.67	2.68	0.36*	-0.14	-0.24	-0.21	0.36
8	197518	0.74	0.05	0.00	4.30	74.17	19.20	2.26	-0.33	0.44*	-0.24	-0.21	0.44
9	197590	0.94	0.02	0.00	0.96	2.42	93.50	3.09	-0.13	-0.11	0.21*	-0.13	0.21
10	197502	0.79	0.06	0.00	6.76	78.79	7.29	7.09	-0.30	0.52*	-0.27	-0.27	0.52
11	197463	0.80	0.07	0.00	7.41	79.79	6.32	6.39	-0.23	0.44*	-0.19	-0.28	0.44
12	197472	0.69	0.07	0.00	4.20	4.75	21.91	69.05	-0.22	-0.23	-0.38	0.54*	0.54
13	197497	0.85	0.06	0.00	8.51	4.62	84.93	1.86	-0.29	-0.28	0.47*	-0.21	0.47
14	197462	0.90	0.07	0.00	2.83	5.03	89.70	2.34	-0.28	-0.20	0.39*	-0.19	0.39
15	197447	0.81	0.06	0.00	80.84	0.76	1.53	16.76	0.34*	-0.09	-0.13	-0.29	0.34

(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	197418	0.67	0.07	0.00	3.15	2.43	67.30	27.00	-0.22	-0.15	0.47*	-0.35	0.47
17	197434	0.83	0.08	0.00	82.87	5.18	6.10	5.73	0.43*	-0.21	-0.26	-0.21	0.43
18	197426	0.77	0.09	0.00	7.65	8.24	76.45	7.55	-0.24	-0.27	0.54*	-0.33	0.54
19	197454	0.69	0.08	0.00	69.23	11.11	12.08	7.48	0.43*	-0.23	-0.16	-0.26	0.43
20	197513	0.89	0.05	0.00	2.91	88.84	5.84	2.33	-0.25	0.38*	-0.20	-0.19	0.38
21	197379	0.59	0.09	0.00	3.64	29.61	7.62	59.00	-0.11	-0.38	-0.07	0.44*	0.44
22	197391	0.92	0.08	0.00	91.87	5.37	1.78	0.85	0.27*	-0.18	-0.14	-0.14	0.27
23	197431	0.90	0.08	0.00	2.11	5.53	89.83	2.41	-0.26	-0.16	0.38*	-0.25	0.38
24	197263	0.51	0.16	0.00	24.02	11.11	50.51	14.16	-0.17	-0.17	0.40*	-0.20	0.40
25	197349	0.56	0.11	0.00	21.71	7.01	14.79	56.33	-0.25	-0.20	-0.07	0.36*	0.36
26	197336	0.80	0.12	0.00	79.69	10.00	4.74	5.41	0.44*	-0.21	-0.23	-0.28	0.44
27	197360	0.70	0.11	0.00	8.38	4.96	69.99	16.52	-0.19	-0.22	0.40*	-0.22	0.40
28	197333	0.72	0.12	0.00	7.85	72.03	8.29	11.67	-0.24	0.44*	-0.19	-0.24	0.44
29	197263	0.52	0.17	0.00	11.50	20.82	51.99	15.49	-0.12	-0.13	0.29*	-0.14	0.29
30	197053	0.77	0.28	0.00	6.50	8.62	8.04	76.53	-0.28	-0.32	-0.28	0.56*	0.56
31	197566	0.86	0.05	3.00	22.56	74.39							
32	197477	0.72	0.09	6.91	42.24	50.76							
33	197272	0.81	0.20	15.62	7.51	76.68							
34	197357	0.73	0.15	21.66	9.87	68.31							
35	197367	0.84	0.15	8.19	16.41	75.25							
36	197383	0.84	0.14	3.31	25.92	70.64							
37	196894	0.43	0.39	52.67	7.62	39.32							
38	197419	0.81	0.12	5.19	10.06	20.94	63.69						
39	197324	0.75	0.17	7.33	17.92	18.41	56.16						
40	197477	0.66	0.09	24.34	18.55	57.01							
41	197390	0.77	0.14	11.08	24.49	64.29							
42	197394	0.80	0.13	9.89	20.12	69.86							
43	197386	0.74	0.14	22.14	7.91	69.81							
44	197427	0.78	0.12	16.05	12.40	71.43							
45	197201	0.66	0.23	12.53	43.07	44.17							
46	197267	0.55	0.20	26.56	36.16	37.08							
47	197391	0.70	0.14	5.90	24.10	24.79	45.08						
48	197225	0.61	0.22	20.66	18.55	16.31	44.26						

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	196083	0.76	0.05	0.00	5.34	14.84	76.04	3.71	-0.22	-0.33	0.45*	-0.13	0.45
2	196058	0.76	0.06	0.00	75.81	11.22	3.75	9.14	0.51*	-0.28	-0.16	-0.35	0.51
3	196077	0.92	0.04	0.00	2.04	91.50	4.91	1.47	-0.15	0.31*	-0.22	-0.14	0.31
4	196066	0.91	0.04	0.00	0.75	2.72	5.74	90.71	-0.14	-0.21	-0.25	0.36*	0.36
5	195863	0.75	0.15	0.00	75.26	8.47	10.76	5.32	0.45*	-0.26	-0.20	-0.25	0.45
6	196125	0.93	0.04	0.00	0.69	5.63	92.88	0.75	-0.11	-0.18	0.23*	-0.10	0.23

(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	195944	0.81	0.13	0.00	3.63	8.07	81.05	7.10	-0.25	-0.30	0.46*	-0.19	0.46
8	195981	0.76	0.10	0.00	7.05	75.95	8.22	8.65	-0.14	0.48*	-0.30	-0.31	0.48
9	195991	0.73	0.11	0.00	6.20	9.22	72.63	11.83	-0.15	-0.23	0.54*	-0.43	0.54
10	196019	0.70	0.08	0.00	12.12	12.03	69.96	5.78	-0.19	-0.14	0.40*	-0.32	0.40
11	196105	0.83	0.05	0.00	2.80	82.48	8.10	6.56	-0.21	0.48*	-0.31	-0.25	0.48
12	196034	0.68	0.07	0.00	3.79	20.64	7.38	68.08	-0.18	-0.15	-0.24	0.34*	0.34
13	196015	0.87	0.10	0.00	3.86	2.48	86.85	6.69	-0.29	-0.24	0.39*	-0.15	0.39
14	196034	0.66	0.07	0.00	6.29	65.68	22.08	5.85	-0.26	0.36*	-0.11	-0.27	0.36
15	195901	0.61	0.13	0.00	24.05	4.72	10.05	61.01	-0.32	-0.13	-0.13	0.42	0.42
16	196040	0.84	0.08	0.00	3.12	11.99	83.47	1.31	-0.20	-0.24	0.35*	-0.17	0.35
17	195901	0.67	0.15	0.00	67.13	4.13	16.21	12.36	0.53	-0.22	-0.22	-0.38*	0.53
18	196021	0.84	0.08	0.00	84.01	2.52	9.11	4.24	0.36*	-0.17	-0.29	-0.11	0.36
19	196049	0.94	0.08	0.00	4.02	1.25	93.85	0.78	-0.14	-0.19	0.26*	-0.14	0.26
20	195724	0.56	0.20	0.00	56.19	21.16	5.92	16.46	0.41*	-0.23	-0.20	-0.17	0.41
21	195840	0.75	0.15	0.00	9.84	6.94	8.52	74.49	-0.37	-0.35	-0.22	0.60*	0.60
22	195631	0.75	0.27	0.00	6.07	8.08	75.01	10.53	-0.28	-0.28	0.52*	-0.25	0.52
23	195711	0.84	0.22	0.00	5.26	4.19	6.49	83.79	-0.20	-0.21	-0.20	0.38*	0.38
24	195640	0.67	0.26	0.00	14.45	5.78	67.23	12.24	-0.42	-0.04	0.40*	-0.09	0.40
25	195582	0.87	0.32	0.00	4.58	3.96	4.27	86.86	-0.24	-0.22	-0.21	0.41*	0.41
26	195428	0.88	0.40	0.00	6.29	2.52	2.83	87.94	-0.12	-0.20	-0.18	0.29*	0.29
27	196108	0.84	0.07	7.13	17.88	74.93							
28	196059	0.65	0.09	17.86	34.78	47.27							
29	196107	0.74	0.07	2.98	45.69	51.26							
30	195518	0.65	0.37	24.71	19.34	55.58							
31	195790	0.69	0.23	11.21	20.89	17.09	50.58						
32	196005	0.78	0.12	3.45	14.73	25.36	56.34						
33	196050	0.68	0.10	5.06	21.69	38.01	35.14						
34	195992	0.58	0.13	22.68	18.98	19.19	39.03						

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	196826	0.91	0.09	0.00	3.07	3.42	90.78	2.62	-0.24	-0.25	0.44*	-0.24	0.44
2	196878	0.90	0.04	0.00	2.91	2.71	90.14	4.17	-0.26	-0.19	0.44*	-0.28	0.44
3	196908	0.92	0.03	0.00	5.56	92.20	1.12	1.06	-0.19	0.30*	-0.18	-0.17	0.30
4	196845	0.80	0.06	0.00	3.85	12.75	3.19	80.12	-0.21	-0.20	-0.19	0.36*	0.36
5	196813	0.75	0.09	0.00	12.62	8.03	74.46	4.79	-0.18	-0.21	0.30*	-0.05	0.30
6	196851	0.85	0.05	0.00	85.39	3.35	7.87	3.29	0.17*	-0.12	-0.04	-0.14	0.17
7	196882	0.82	0.05	0.00	7.92	4.87	81.53	5.61	-0.27	-0.28	0.43*	-0.14	0.43
8	196786	0.67	0.10	0.00	19.69	66.62	9.73	3.85	-0.28	0.48*	-0.24	-0.21	0.48
9	196845	0.56	0.07	0.00	19.78	22.21	56.35	1.57	-0.28	-0.25	0.45*	-0.06	0.45
10	196856	0.70	0.07	0.00	7.80	10.53	11.88	69.70	-0.24	-0.24	-0.10	0.38*	0.38
11	196864	0.80	0.07	0.00	5.10	3.37	79.68	11.77	-0.11	-0.16	0.41*	-0.34	0.41
12	196791	0.66	0.09	0.00	10.68	8.71	14.13	66.36	-0.17	-0.13	-0.07	0.24*	0.24

(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	196855	0.71	0.07	0.00	11.37	7.93	70.57	10.05	-0.33	-0.24	0.46*	-0.13	0.46
14	196724	0.79	0.13	0.00	79.15	4.66	8.94	7.10	0.46*	-0.29	-0.24	-0.22	0.46
15	196804	0.64	0.09	0.00	6.77	17.62	63.51	11.98	-0.29	-0.10	0.36*	-0.18	0.36
16	196809	0.73	0.09	0.00	73.20	13.48	3.47	9.74	0.45*	-0.38	-0.17	-0.13	0.45
17	196865	0.87	0.07	0.00	1.87	5.06	86.98	6.01	-0.15	-0.19	0.37*	-0.26	0.37
18	196831	0.83	0.08	0.00	7.12	82.84	2.30	7.65	-0.39	0.47*	-0.14	-0.21	0.47
19	196798	0.84	0.09	0.00	2.75	7.38	83.71	6.05	-0.22	-0.29	0.45*	-0.22	0.45
20	196662	0.65	0.16	0.00	64.92	5.78	10.55	18.57	0.50*	-0.17	-0.21	-0.34	0.50
21	196797	0.87	0.10	0.00	8.92	86.40	2.18	2.39	-0.30	0.44*	-0.20	-0.23	0.44
22	196634	0.54	0.16	0.00	6.63	34.57	4.79	53.81	-0.30	-0.29	-0.15	0.50*	0.50
23	196628	0.82	0.18	0.00	8.46	3.40	81.72	6.22	-0.29	-0.23	0.47*	-0.23	0.47
24	196579	0.57	0.20	0.00	4.53	4.33	34.11	56.81	-0.31	-0.26	-0.21	0.44*	0.44
25	196457	0.91	0.28	0.00	5.24	90.40	3.08	0.99	-0.18	0.29*	-0.17	-0.14	0.29
26	196747	0.89	0.14	7.09	6.83	85.94							
27	196801	0.86	0.11	10.14	8.49	81.26							
28	194030	0.56	1.52	28.89	27.97	41.62							
29	196617	0.74	0.21	11.55	28.46	59.79							
30	196341	0.37	0.35	49.40	26.96	23.29							
31	196540	0.72	0.25	24.14	7.52	68.10							
32	196488	0.78	0.27	10.86	5.25	21.73	61.89						
33	195680	0.66	0.68	16.57	13.14	24.24	45.36						
34	196727	0.45	0.15	16.53	43.29	29.26	10.77						
35	196754	0.80	0.14	6.07	13.15	15.03	65.62						

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	198399	0.86	0.04	0.00	1.80	2.16	85.63	10.35	-0.17	-0.17	0.30*	-0.19	0.30
2	198382	0.81	0.04	0.00	81.21	3.96	7.60	7.16	0.48*	-0.18	-0.30	-0.28	0.48
3	198230	0.79	0.12	0.00	6.20	79.21	9.64	4.81	-0.14	0.46*	-0.34	-0.22	0.46
4	198008	0.42	0.21	0.00	17.36	12.07	41.56	28.75	-0.13	-0.10	0.25*	-0.10	0.25
5	198003	0.62	0.23	0.00	62.32	19.96	12.05	5.41	0.24*	-0.17	-0.09	-0.09	0.24
6	198082	0.52	0.18	0.00	9.91	14.17	51.94	23.76	-0.17	-0.20	0.27*	-0.03	0.27
7	198316	0.75	0.08	0.00	17.42	6.18	74.55	1.74	-0.22	-0.31	0.41*	-0.16	0.41
8	198388	0.91	0.05	0.00	91.32	1.05	6.10	1.46	0.37*	-0.12	-0.31	-0.13	0.37
9	198391	0.86	0.05	0.00	85.93	4.71	2.16	7.13	0.38*	-0.23	-0.19	-0.22	0.38
10	198039	0.42	0.20	0.00	12.71	20.66	24.90	41.49	-0.24	-0.21	-0.06	0.39*	0.39
11	198400	0.87	0.04	0.00	87.39	3.22	6.26	3.07	0.42*	-0.26	-0.25	-0.19	0.42
12	198321	0.86	0.08	0.00	6.13	86.28	4.00	3.49	-0.23	0.40*	-0.22	-0.20	0.40
13	198330	0.86	0.07	0.00	85.50	6.61	4.36	3.44	0.36*	-0.25	-0.16	-0.16	0.36
14	198346	0.84	0.06	0.00	5.77	3.26	83.51	7.38	-0.31	-0.18	0.42*	-0.20	0.42
15	198244	0.57	0.11	0.00	57.13	16.29	18.47	7.96	0.36*	-0.18	-0.15	-0.18	0.36
16	198238	0.65	0.11	0.00	22.70	4.82	7.37	64.97	-0.42	-0.23	-0.15	0.56*	0.56
17	198363	0.90	0.05	0.00	90.04	3.62	3.46	2.81	0.22*	-0.13	-0.12	-0.10	0.22
18	198209	0.42	0.13	0.00	6.54	31.49	41.98	19.84	-0.13	-0.43	0.44*	0.04	0.44

(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	198230	0.52	0.12	0.00	52.38	28.59	11.82	7.07	0.46*	-0.17	-0.30	-0.23	0.46
20	198336	0.85	0.07	0.00	85.31	2.52	4.55	7.53	0.51*	-0.17	-0.26	-0.38	0.51
21	198333	0.88	0.07	0.00	1.55	4.11	87.42	6.83	-0.14	-0.21	0.43*	-0.33	0.43
22	198277	0.82	0.10	0.00	5.03	82.22	6.90	5.73	-0.26	0.35*	-0.22	-0.09	0.35
23	198297	0.79	0.08	0.00	15.80	2.40	79.39	2.30	-0.18	-0.21	0.31*	-0.17	0.31
24	198200	0.63	0.12	0.00	15.48	63.12	8.15	13.09	-0.17	0.45*	-0.19	-0.29	0.45
25	198109	0.58	0.17	0.00	58.17	9.54	14.25	17.84	0.49*	-0.29	-0.22	-0.20	0.49
26	198152	0.65	0.15	0.00	5.18	64.50	16.53	13.62	-0.17	0.32*	-0.31	-0.01	0.32
27	198085	0.75	0.18	0.00	4.94	75.23	11.97	7.64	-0.25	0.50*	-0.27	-0.27	0.50
28	198107	0.72	0.17	0.00	8.61	11.35	72.33	7.50	-0.33	-0.27	0.53*	-0.22	0.53
29	198002	0.64	0.23	0.00	63.75	14.19	10.15	11.65	0.45*	-0.17	-0.26	-0.22	0.45
30	198013	0.84	0.23	0.00	2.82	10.63	2.53	83.77	-0.17	-0.09	-0.17	0.23*	0.23
31	197902	0.84	0.31	7.42	16.92	75.36							
32	196984	0.23	0.77	70.82	10.54	17.87							
33	198093	0.85	0.21	7.13	15.98	76.67							
34	197878	0.67	0.32	24.28	16.31	59.09							
35	197633	0.75	0.44	6.98	12.16	30.54	49.87						
36	194975	0.51	1.78	27.29	11.61	37.99	21.33						
37	196999	0.35	0.76	29.03	49.78	6.67	13.75						
38	196756	0.42	0.89	31.95	26.91	23.59	16.67						

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	202992	0.92	0.05	0.00	1.00	91.62	5.78	1.55	-0.10	0.26*	-0.16	-0.18	0.26
2	202959	0.84	0.04	0.00	5.05	4.96	5.82	84.10	-0.26	-0.29	-0.30	0.53*	0.53
3	202962	0.79	0.06	0.00	13.03	78.50	4.68	3.72	-0.42	0.55*	-0.26	-0.16	0.55
4	202894	0.86	0.08	0.00	4.64	86.13	4.55	4.58	-0.35	0.48*	-0.23	-0.20	0.48
5	202961	0.81	0.05	0.00	4.15	4.85	10.38	80.55	-0.26	-0.27	-0.22	0.45*	0.45
6	202944	0.83	0.07	0.00	7.30	3.54	5.99	83.09	-0.22	-0.24	-0.30	0.46*	0.46
7	202966	0.78	0.05	0.00	78.29	7.75	12.59	1.31	0.52*	-0.35	-0.30	-0.19	0.52
8	202723	0.69	0.17	0.00	8.86	68.90	14.03	8.02	-0.32	0.57*	-0.27	-0.28	0.57
9	202770	0.67	0.14	0.00	66.50	11.83	13.29	8.21	0.52*	-0.26	-0.28	-0.22	0.52
10	202842	0.62	0.11	0.00	7.28	61.42	12.66	18.51	-0.18	0.49*	-0.26	-0.27	0.49
11	202942	0.86	0.06	0.00	5.21	3.94	4.71	86.06	-0.24	-0.25	-0.31	0.49*	0.49
12	202960	0.77	0.05	0.00	11.50	3.55	7.60	77.28	-0.24	-0.19	-0.33	0.48*	0.48
13	202818	0.71	0.12	0.00	5.34	71.13	19.93	3.47	-0.14	0.34*	-0.21	-0.20	0.34
14	202915	0.69	0.08	0.00	8.52	68.98	13.25	9.15	-0.28	0.46*	-0.20	-0.23	0.46
15	202877	0.83	0.09	0.00	6.08	6.73	82.71	4.36	-0.23	-0.29	0.47*	-0.23	0.47
16	202875	0.64	0.09	0.00	5.30	25.97	4.30	64.31	-0.27	-0.21	-0.26	0.43*	0.43

(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	202881	0.76	0.09	0.00	75.49	9.51	4.04	10.84	0.33*	-0.21	-0.21	-0.12	0.33
18	202949	0.92	0.06	0.00	1.77	2.70	91.52	3.92	-0.22	-0.20	0.39*	-0.23	0.39
19	202835	0.83	0.11	0.00	5.99	3.85	6.69	83.33	-0.23	-0.21	-0.27	0.44*	0.44
20	202953	0.81	0.06	0.00	13.32	3.88	80.83	1.90	-0.23	-0.28	0.41*	-0.21	0.41
21	202845	0.62	0.10	0.00	8.20	22.57	6.98	62.12	-0.09	-0.32	-0.18	0.42*	0.42
22	202746	0.70	0.15	0.00	10.59	69.60	4.05	15.58	-0.19	0.37*	-0.21	-0.19	0.37
23	202910	0.96	0.08	0.00	1.03	1.74	96.05	1.08	-0.19	-0.21	0.33*	-0.17	0.33
24	202803	0.58	0.12	0.00	8.45	13.18	20.00	58.22	-0.27	-0.22	-0.28	0.53*	0.53
25	202880	0.83	0.09	0.00	11.27	2.20	3.94	82.48	-0.37	-0.23	-0.24	0.52*	0.52
26	202790	0.82	0.14	0.00	81.76	6.48	5.90	5.70	0.56*	-0.26	-0.30	-0.34	0.56
27	202646	0.78	0.21	0.00	8.85	78.06	8.50	4.36	-0.24	0.50*	-0.31	-0.23	0.50
28	202260	0.70	0.42	19.62	21.29	58.67							
29	202368	0.72	0.36	6.57	42.04	51.02							
30	202136	0.76	0.48	14.14	20.02	65.36							
31	202300	0.69	0.40	11.65	38.48	49.48							
32	201484	0.70	0.80	13.23	18.31	12.78	54.88						
33	202353	0.68	0.37	13.72	15.39	23.98	46.55						
34	201884	0.68	0.60	23.95	14.79	60.65							
35	202228	0.61	0.43	34.95	8.76	55.86							
36	202360	0.57	0.37	21.64	42.91	35.09							
37	200838	0.56	1.12	32.85	21.16	44.87							
38	201821	0.58	0.63	19.22	45.08	35.06							
39	202351	0.63	0.37	25.62	22.31	51.70							
40	198782	0.53	2.13	38.27	15.19	44.41							
41	199711	0.59	1.67	32.43	16.71	49.18							
42	201993	0.75	0.55	9.81	12.36	20.99	56.29						
43	201867	0.52	0.61	33.59	12.39	18.56	34.85						
44	201629	0.75	0.73	10.70	14.69	14.32	59.57						
45	201731	0.66	0.68	8.52	27.55	21.50	41.76						

Point-Biserial Correlation Coefficients

Point biserial statistics are used to examine item-test correlations, or item discrimination. As shown in Tables 8a–8f, point biserial correlation coefficients were computed for each answer option. Point biserials for the correct answer options are denoted with asterisks (*) 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. Point biserials for correct answer options (pbis*) on the tests ranged from 0.17–0.60. For Grade 3, the pbis* were between 0.28 and 0.53. For Grade 4, the pbis* were between 0.21 and 0.56. For Grade 5, the pbis* were between 0.23 and 0.60. For Grade 6, pbis* were between 0.17 and 0.50. For Grade 7, the pbis* were between 0.21 and 0.56. For Grade 8, the pbis* were between 0.26 and 0.57.

Distractor Analysis

Item distractors provide additional information about student performance on test questions. Two types of information on item distractors are available from New York State test data: information on the 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 the proportions of students selecting any other answer choice. This was true for all New York State mathematics items.

As mentioned in the "Point Biserial Correlation 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 18, with a point biserial of 0.04).

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 from 0.88–0.94. Feldt-Raju reliability coefficients ranged from 0.89–0.95. 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 Grades 4 and 8 tests. All reliabilities exceeded 0.85 across statistics, which is a good indication that the NYSTP Grades 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 2010 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	32.79	6.31	0.85	0.68	0.97	0.88	0.89
4	70	52.44	13.94	0.76	0.43	0.96	0.94	0.95
5	46	34.20	8.88	0.76	0.56	0.94	0.90	0.91
6	49	35.40	9.56	0.74	0.37	0.92	0.90	0.91
7	50	32.80	9.75	0.69	0.23	0.91	0.90	0.91
8	69	48.07	15.94	0.72	0.51	0.96	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 DIF was evaluated using two methods. First, the standardized mean difference (SMD) was computed for all items. The SMD statistic (Dorans, Schmitt & 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, & 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 & Jones, 1993).

Classical DIF analyses were conducted on subgroups of NRC (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 ELLs (focal group: ELLs; reference group: Non-ELLs). 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 2010 Classical DIF Sample N-Counts

Grade	Ethnicity				Gender		Needs Resource Category		Test Language	
	Black	Hispanic	Asian	White	Female	Male	High	Low	Spanish	English
3	36854	43849	15407	95996	94822	99446	105709	83950	3504	190764
4	37517	43174	16530	98422	96513	101146	106334	87350	3289	194370
5	37519	42539	15543	98772	95613	100630	102979	88204	3195	193048
6	37832	41963	15444	99960	96513	100510	102108	90492	2873	194150
7	38066	41663	15640	101397	96859	101657	103092	91655	3175	195341
8	38191	42569	16061	104665	99011	104095	105770	93823	3192	199913

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	5
4	4
5	6
6	6
7	11
8	11

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, 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

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

where

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

and

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

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

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

and

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

where

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

and

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

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

Calibration Sample

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

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

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 detail in the final OP test technical reports. As shown in Tables 12 through 14, the 2010 samples were comparable to 2009 populations in terms of NRC, student ethnicity, proportions of ELL, proportions of SWD, and proportions of SUA.

Table 12. Grades 3 and 4 Demographic Statistics

Demographics	2009 Grade 3 Population	2010 Grade 3 Sample	2009 Grade 4 Population	2010 Grade 4 Sample
	%	%	%	%
NRC SUBGROUPS				
NYC	35.07	36.82	34.73	36.54
Big cities	4.17	4.38	4.10	4.10
Urban/Suburban	8.24	8.09	8.16	8.09
Rural	5.83	5.78	5.87	5.75
Average needs	29.60	29.22	30.24	29.59
Low needs	15.04	13.59	15.13	14.18
Charter	1.74	2.12	1.47	1.75
ETHNICITY				
Asian	8.19	7.93	7.67	8.36
Black	18.61	18.97	18.83	18.98
Hispanics	21.36	22.57	21.27	21.84
American Indian	0.47	0.49	0.46	0.47
Multi-Racial	0.33	0.56	0.25	0.50
White	50.98	49.41	51.47	49.79
Unknown	0.05	0.06	0.04	0.06
ELL STATUS				
No	91.91	91.55	93.2	92.60
Yes	8.09	8.45	6.80	7.40
DISABILITY				
No	86.68	85.75	85.72	85.25
Yes	13.32	14.25	14.28	14.75
ACCOMMODATIONS				
No	76.77	75.15	76.50	75.00
Yes	23.23	24.85	23.50	25.00

Table 13. Grades 5 and 6 Demographic Statistics

Demographics	2009 Grade 5 Population	2010 Grade 5 Sample	2009 Grade 6 Population	2010 Grade 6 Sample
	%	%	%	%
NRC SUBGROUPS				
NYC	34.39	35.46	34.30	35.09
Big cities	3.86	4.07	3.86	3.90
Urban/Suburban	7.92	7.80	7.69	7.75
Rural	5.81	5.77	5.77	5.73
Average needs	30.48	29.94	30.99	30.73
Low needs	15.39	14.66	15.39	14.88
Charter	1.79	2.30	1.61	1.93
ETHNICITY				
Asian	7.59	7.92	7.66	7.84
Black	18.92	19.12	19.02	19.20
Hispanics	21.03	21.68	20.58	21.30
American Indian	0.48	0.47	0.45	0.49
Multi-Racial	0.26	0.44	0.23	0.39
White	51.68	50.33	52.02	50.74
Unknown	0.05	0.05	0.04	0.05
ELL STATUS				
No	94.35	94.07	95.41	95.05
Yes	5.65	5.93	4.59	4.95
DISABILITY				
No	84.97	84.78	84.87	84.69
Yes	15.03	15.22	15.13	15.31
ACCOMMODATIONS				
No	76.38	75.56	78.28	77.68
Yes	23.62	24.44	21.72	22.32

Table 14. Grades 7 and 8 Demographic Statistics

Demographics	2009 Grade 7 Population	2010 Grade 7 Sample	2009 Grade 8 Population	2010 Grade 8 Sample
	%	%	%	%
NRC SUBGROUPS				
NYC	34.36	35.34	34.98	35.95
Big cities	3.89	3.90	3.83	3.78
Urban/Suburban	7.70	7.66	7.74	7.48
Rural	5.99	5.79	6.06	5.80
Average needs	31.07	31.21	31.83	30.87
Low needs	15.29	14.65	15.46	14.95
Charter	1.22	1.46	1.06	1.16
ETHNICITY				
Asian	7.55	7.88	7.47	7.91
Black	18.75	19.18	18.86	18.80
Hispanics	20.42	20.99	20.32	20.96
American Indian	0.46	0.48	0.49	0.45
Multi-Racial	0.20	0.36	0.16	0.30
White	52.58	51.08	52.68	51.53
Unknown	0.04	0.04	0.03	0.05
ELL STATUS				
No	96.06	95.59	96.10	95.84
Yes	3.94	4.41	3.90	4.16
DISABILITY				
No	84.94	85.00	85.62	85.21
Yes	15.06	15.00	14.38	14.79
ACCOMMODATIONS				
No	79.23	78.57	79.78	78.90
Yes	20.77	21.43	20.22	21.10

The NRC and ethnicity distributions of the FT anchor form samples are compared with those of the OP samples in Tables 15 through 17. It is apparent that the FT anchor samples represent the OP student population well.

Table 15. Grades 3 and 4 Demographic Statistics for Field Test Anchor Forms

Demographics	2010 Grade 3 FT Anchor Form 1	2010 Grade 3 FT Anchor Form 2	2010 Grade 3 OP Sample	2010 Grade 4 FT Anchor Form 1	2010 Grade 4 FT Anchor Form 2	2010 Grade 4 OP Sample
	%	%	%	%	%	%
NRC SUBGROUPS						
NYC	39.46	35.38	36.82	39.85	34.62	36.54
Big cities	4.52	4.14	4.38	3.86	3.85	4.10
Urban/Suburban	9.40	7.86	8.09	9.35	7.72	8.09
Rural	4.76	5.59	5.78	4.91	5.78	5.75
Average needs	26.92	30.22	29.22	26.92	31.19	29.59
Low needs	12.97	14.47	13.59	13.48	14.97	14.18
Charter	1.76	2.07	2.12	1.45	1.63	1.75
ETHNICITY						
Asian	9.21	7.76	7.93	10.94	8.10	8.36
Black	15.37	17.86	18.97	13.74	17.98	18.98
Hispanics	30.17	21.79	22.57	29.51	20.73	21.84
American Indian	0.35	0.61	0.49	0.62	0.40	0.47
Multi-Racial	0.42	0.61	0.56	0.33	0.60	0.50
White	44.45	51.32	49.41	44.82	52.12	49.79
Unknown	0.03	0.04	0.06	0.05	0.08	0.06

Table 16. Grades 5 and 6 Demographic Statistics for Field Test Anchor Forms

Demographics	2010 Grade 5 FT Anchor Form 1	2010 Grade 5 FT Anchor Form 2	2010 Grade 5 OP Sample	2010 Grade 6 FT Anchor Form 1	2010 Grade 6 FT Anchor Form 2	2010 Grade 6 OP Sample
	%	%	%	%	%	%
NRC SUBGROUPS						
NYC	39.19	33.89	35.46	37.49	32.89	35.09
Big cities	4.04	3.80	4.07	3.56	3.67	3.90
Urban/Suburban	8.00	7.46	7.80	8.32	7.55	7.75
Rural	4.72	5.93	5.77	5.19	5.68	5.73
Average needs	27.77	31.10	29.94	28.76	32.32	30.73
Low needs	14.23	15.47	14.66	15.06	15.92	14.88
Charter	1.80	2.12	2.30	1.29	1.70	1.93
ETHNICITY						
Asian	10.02	7.97	7.92	9.08	8.02	7.84
Black	14.79	18.30	19.12	15.11	18.68	19.20
Hispanics	28.79	20.95	21.68	27.37	19.54	21.30
American Indian	0.48	0.45	0.47	0.40	0.40	0.49
Multi-Racial	0.29	0.36	0.44	0.48	0.34	0.39
White	45.61	51.94	50.33	0.1	52.95	50.74
Unknown	0.02	0.04	0.05	47.47	0.05	0.05

Table 17. Grades 7 and 8 Demographic Statistics for Field Test Anchor Forms

Demographics	2010 Grade 7 OP Anchor Form 1	2010 Grade 7 OP Anchor Form 2	2010 Grade 7 OP Sample	2010 Grade 8 FT Anchor Form 1	2010 Grade 8 FT Anchor Form 2	2010 Grade 8 OP Sample
	%	%	%	%	%	%
NRC SUBGROUPS						
NYC	38.83	33.28	35.34	40.03	34.04	35.95
Big cities	3.29	3.63	3.90	3.54	3.45	3.78
Urban/Suburban	8.02	8.02	7.66	8.19	7.86	7.48
Rural	4.79	5.88	5.79	5.10	6.14	5.80
Average needs	29.67	32.52	31.21	28.33	31.95	30.87
Low needs	14.04	14.82	14.65	13.55	15.11	14.95
Charter	1.13	1.48	1.46	0.93	1.06	1.16
ETHNICITY						
Asian	10.38	7.60	7.88	9.46	7.93	7.91
Black	14.32	18.92	19.18	15.49	19.00	18.80
Hispanics	27.50	20.08	20.99	28.58	19.75	20.96
American Indian	0.49	0.62	0.48	0.34	0.29	0.45
Multi-Racial	0.26	0.24	0.36	0.27	0.42	0.30
White	46.98	52.51	51.08	45.81	52.55	51.53
Unknown	0.07	0.03	0.04	0.05	0.06	0.05

Calibration Process

The IRT model parameters were estimated using CTB/McGraw-Hill's PARDUX software (Burket, 2002). PARDUX estimates parameters simultaneously for MC and CR items using marginal maximum likelihood procedures implemented via the expectation-maximization (EM) algorithm (Bock & 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 & Julian, 1996).

The NYSTP Mathematics Tests item calibrations did not incur any problems. The number of estimation cycles was set to 50 with a 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. 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. Table 18 presents a summary of calibration results. For the Grades 3–8 Mathematics Tests, all of the calibration estimation results are reasonable.

Table 18. NYSTP Mathematics 2010 Calibration Results

Grade	Largest *a*-parameter	Lowest and highest *b*-parameters		# Items with Default *c*-parameters	Theta Mean	Theta Standard Deviation	# Students
3	2.191	-2.847	-0.163	10	0.21	1.675	194268
4	2.472	-4.713	1.057	15	0.06	1.189	197659
5	2.611	-3.776	0.770	8	0.02	1.236	196243
6	2.683	-3.196	1.148	5	0.01	1.194	197023
7	2.319	-2.877	1.229	9	-0.06	1.155	198516
8	2.654	-2.598	0.976	7	0.04	1.206	203106

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 3PL model. Students are rank-ordered on the basis of $\hat{\theta}$ values and sorted into ten cells with 10% of the sample in each cell. For each item, the number of students in cell *k* who answered item *i*, N_{ik}, and the number of students in that cell who answered item *i* correctly, R_{ik}, were determined. The observed proportion in cell *k* passing item *i*, O_{ik}, is R_{ik}/N_{ik}. The fit index for item *i* is

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

with

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

A modification of this procedure was used to measure fit to the 2PPC model. For the 2PPC 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 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 2010 Grades 3–8 Mathematics Tests demonstrated a good model fit, further supporting use of the chosen models. No items in Grades 3 and 5 exhibited poor item-model fit statistics. The following items exhibited misfit: Grade 4 items 40 and 46, Grade 6 item 28, Grade 7 item 35, and Grade 8 items 29, 35, and 36. The fact that so few items were flagged for poor fit across all mathematics tests further supports the use of the chosen models. Fit statistics and status for all items in Grades 3–8 Mathematics Tests are presented in Tables 19–24.

Table 19. Mathematics Grade 3 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	Z_{Q1}	Z_{Q1} critical	Fit OK?
1	3PL	84.34	7	172362	20.67	459.632	Y
2	3PL	108.47	7	172362	27.12	459.632	Y
3	3PL	660.65	7	172362	174.69	459.632	Y
4	3PL	250.83	7	172362	65.17	459.632	Y
5	3PL	53.75	7	172362	12.50	459.632	Y
6	3PL	145.51	7	172362	37.02	459.632	Y
7	3PL	166.43	7	172362	42.61	459.632	Y
8	3PL	963.03	7	172362	255.51	459.632	Y
9	3PL	399.55	7	172362	104.91	459.632	Y
10	3PL	86.42	7	172362	21.23	459.632	Y
11	3PL	57.52	7	172362	13.50	459.632	Y
12	3PL	397.58	7	172362	104.39	459.632	Y
13	3PL	480.87	7	172362	126.65	459.632	Y
14	3PL	164.55	7	172362	42.11	459.632	Y
15	3PL	88.63	7	172362	21.82	459.632	Y
16	3PL	182.22	7	172362	46.83	459.632	Y
17	3PL	187.35	7	172362	48.20	459.632	Y
18	3PL	172.77	7	172362	44.30	459.632	Y
19	3PL	926.20	7	172362	245.67	459.632	Y
20	3PL	51.61	7	172362	11.92	459.632	Y
21	3PL	67.81	7	172362	16.25	459.632	Y
22	3PL	87.68	7	172362	21.56	459.632	Y
23	3PL	310.04	7	172362	80.99	459.632	Y

(Continued on next page)

Table 19. Mathematics Grade 3 Item Fit Statistics(cont.)

Item	Model	Chi Square	DF	Total N	Z_{Q1}	Z_{Q1} critical	Fit OK?
24	3PL	570.38	7	172362	150.57	459.632	Y
25	3PL	322.81	7	172362	84.40	459.632	Y
26	2PPC	416.11	17	172362	68.45	459.632	Y
27	2PPC	1437.38	17	172362	243.59	459.632	Y
28	2PPC	1051.83	17	172362	177.47	459.632	Y
29	2PPC	2251.01	17	172362	383.13	459.632	Y
30	2PPC	889.10	26	172362	119.69	459.632	Y
31	2PPC	2286.07	26	172362	313.42	459.632	Y

Table 20. Mathematics Grade 4 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	Z_{Q1}	Z_{Q1} critical	Fit OK?
1	3PL	46.78	7	193793	10.63	516.781	Y
2	3PL	62.08	7	193793	14.72	516.781	Y
3	3PL	95.16	7	193793	23.56	516.781	Y
4	3PL	339.33	7	193793	88.82	516.781	Y
5	3PL	12.87	7	193793	1.57	516.781	Y
6	3PL	36.59	7	193793	7.91	516.781	Y
7	3PL	82.81	7	193793	20.26	516.781	Y
8	3PL	42.19	7	193793	9.40	516.781	Y
9	3PL	234.65	7	193793	60.84	516.781	Y
10	3PL	30.66	7	193793	6.32	516.781	Y
11	3PL	17.22	7	193793	2.73	516.781	Y
12	3PL	28.35	7	193793	5.71	516.781	Y
13	3PL	54.97	7	193793	12.82	516.781	Y
14	3PL	341.58	7	193793	89.42	516.781	Y
15	3PL	15.80	7	193793	2.35	516.781	Y
16	3PL	18.63	7	193793	3.11	516.781	Y
17	3PL	52.16	7	193793	12.07	516.781	Y
18	3PL	90.52	7	193793	22.32	516.781	Y
19	3PL	76.29	7	193793	18.52	516.781	Y
20	3PL	24.28	7	193793	4.62	516.781	Y
21	3PL	246.69	7	193793	64.06	516.781	Y
22	3PL	128.47	7	193793	32.46	516.781	Y
23	3PL	1108.83	7	193793	294.48	516.781	Y
24	3PL	139.18	7	193793	35.33	516.781	Y
25	3PL	68.48	7	193793	16.43	516.781	Y
26	3PL	286.94	7	193793	74.82	516.781	Y
27	3PL	44.16	7	193793	9.93	516.781	Y
28	3PL	20.90	7	193793	3.71	516.781	Y
29	3PL	101.21	7	193793	25.18	516.781	Y
30	3PL	91.84	7	193793	22.67	516.781	Y
31	2PPC	419.08	17	193793	68.96	516.781	Y

(Continued on next page)

Table 20. Mathematics Grade 4 Item Fit Statistics (cont.)

32	2PPC	1909.87	17	193793	324.62	516.781	Y
33	2PPC	173.70	17	193793	26.87	516.781	Y
34	2PPC	179.33	17	193793	27.84	516.781	Y
35	2PPC	843.25	17	193793	141.70	516.781	Y
36	2PPC	1191.87	17	193793	201.49	516.781	Y
37	2PPC	633.59	17	193793	105.74	516.781	Y
38	2PPC	1061.66	26	193793	143.62	516.781	Y
39	2PPC	1024.97	26	193793	138.53	516.781	Y
40	2PPC	5403.40	17	193793	923.76	516.781	N
41	2PPC	720.56	17	193793	120.66	516.781	Y
42	2PPC	943.08	17	193793	158.82	516.781	Y
43	2PPC	279.93	17	193793	45.09	516.781	Y
44	2PPC	1744.48	17	193793	296.26	516.781	Y
45	2PPC	855.27	17	193793	143.76	516.781	Y
46	2PPC	3133.55	17	193793	534.48	516.781	N
47	2PPC	661.06	26	193793	88.07	516.781	Y
48	2PPC	1363.49	26	193793	185.48	516.781	Y

Table 21. Mathematics Grade 5 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	Z_{Q1}	Z_{Q1} critical	Fit OK?
1	3PL	33.56	7	192496	7.10	513.323	Y
2	3PL	124.16	7	192496	31.31	513.323	Y
3	3PL	286.68	7	192496	74.75	513.323	Y
4	3PL	51.94	7	192496	12.01	513.323	Y
5	3PL	40.63	7	192496	8.99	513.323	Y
6	3PL	218.55	7	192496	56.54	513.323	Y
7	3PL	39.28	7	192496	8.63	513.323	Y
8	3PL	33.14	7	192496	6.99	513.323	Y
9	3PL	289.23	7	192496	75.43	513.323	Y
10	3PL	35.11	7	192496	7.51	513.323	Y
11	3PL	107.50	7	192496	26.86	513.323	Y
12	3PL	139.14	7	192496	35.32	513.323	Y
13	3PL	521.18	7	192496	137.42	513.323	Y
14	3PL	245.08	7	192496	63.63	513.323	Y
15	3PL	36.66	7	192496	7.93	513.323	Y
16	3PL	31.91	7	192496	6.66	513.323	Y
17	3PL	208.85	7	192496	53.95	513.323	Y
18	3PL	41.31	7	192496	9.17	513.323	Y
19	3PL	138.37	7	192496	35.11	513.323	Y
20	3PL	148.88	7	192496	37.92	513.323	Y
21	3PL	239.89	7	192496	62.24	513.323	Y

(Continued on next page)

Table 21. Mathematics Grade 5 Item Fit Statistics (cont.)

Item	Model	Chi Square	DF	Total N	Z_{Q1}	Z_{Q1} critical	Fit OK?
22	3PL	94.50	7	192496	23.39	513.323	Y
23	3PL	167.97	7	192496	43.02	513.323	Y
24	3PL	80.84	7	192496	19.74	513.323	Y
25	3PL	104.48	7	192496	26.05	513.323	Y
26	3PL	230.19	7	192496	59.65	513.323	Y
27	2PPC	494.60	17	192496	81.91	513.323	Y
28	2PPC	593.50	17	192496	98.87	513.323	Y
29	2PPC	512.69	17	192496	85.01	513.323	Y
30	2PPC	2036.35	17	192496	346.32	513.323	Y
31	2PPC	633.92	26	192496	84.30	513.323	Y
32	2PPC	1292.39	26	192496	175.62	513.323	Y
33	2PPC	3312.12	26	192496	455.70	513.323	Y
34	2PPC	1448.44	26	192496	197.26	513.323	Y

Table 22. Mathematics Grade 6 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	Z_{Q1}	Z_{Q1} critical	Fit OK?
1	3PL	75.49	7	194773	18.30	519.395	Y
2	3PL	77.31	7	194773	18.79	519.395	Y
3	3PL	329.19	7	194773	86.11	519.395	Y
4	3PL	41.73	7	194773	9.28	519.395	Y
5	3PL	298.58	7	194773	77.93	519.395	Y
6	3PL	52.85	7	194773	12.25	519.395	Y
7	3PL	57.80	7	194773	13.58	519.395	Y
8	3PL	369.17	7	194773	96.79	519.395	Y
9	3PL	173.33	7	194773	44.45	519.395	Y
10	3PL	74.94	7	194773	18.16	519.395	Y
11	3PL	22.10	7	194773	4.04	519.395	Y
12	3PL	513.44	7	194773	135.35	519.395	Y
13	3PL	109.99	7	194773	27.53	519.395	Y
14	3PL	48.76	7	194773	11.16	519.395	Y
15	3PL	38.82	7	194773	8.50	519.395	Y
16	3PL	200.76	7	194773	51.79	519.395	Y
17	3PL	90.64	7	194773	22.35	519.395	Y
18	3PL	43.51	7	194773	9.76	519.395	Y
19	3PL	138.84	7	194773	35.23	519.395	Y
20	3PL	294.11	7	194773	76.73	519.395	Y
21	3PL	48.68	7	194773	11.14	519.395	Y
22	3PL	190.61	7	194773	49.07	519.395	Y
23	3PL	154.49	7	194773	39.42	519.395	Y
24	3PL	89.86	7	194773	22.14	519.395	Y
25	3PL	45.66	7	194773	10.33	519.395	Y
26	2PPC	1527.60	17	194773	259.07	519.395	Y

(Continued on next page)

Table 22. Mathematics Grade 6 Item Fit Statistics (cont.)

27	2PPC	1400.51	17	194773	237.27	519.395	Y
28	2PPC	3612.36	17	194773	616.60	519.395	N
29	2PPC	746.19	17	194773	125.06	519.395	Y
30	2PPC	753.31	17	194773	126.28	519.395	Y
31	2PPC	711.82	17	194773	119.16	519.395	Y
32	2PPC	2966.97	26	194773	407.84	519.395	Y
33	2PPC	1217.46	26	194773	165.23	519.395	Y
34	2PPC	1237.67	26	194773	168.03	519.395	Y
35	2PPC	521.12	26	194773	68.66	519.395	Y

Table 23. Mathematics Grade 7 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	Z_{Q1}	Z_{Q1} critical	Fit OK?
1	3PL	261.94	7	197405	68.14	526.413	Y
2	3PL	48.28	7	197405	11.03	526.413	Y
3	3PL	28.80	7	197405	5.83	526.413	Y
4	3PL	55.06	7	197405	12.84	526.413	Y
5	3PL	269.45	7	197405	70.14	526.413	Y
6	3PL	145.79	7	197405	37.09	526.413	Y
7	3PL	64.15	7	197405	15.27	526.413	Y
8	3PL	214.02	7	197405	55.33	526.413	Y
9	3PL	148.41	7	197405	37.79	526.413	Y
10	3PL	293.51	7	197405	76.57	526.413	Y
11	3PL	58.72	7	197405	13.82	526.413	Y
12	3PL	92.53	7	197405	22.86	526.413	Y
13	3PL	21.53	7	197405	3.88	526.413	Y
14	3PL	63.74	7	197405	15.16	526.413	Y
15	3PL	19.17	7	197405	3.25	526.413	Y
16	3PL	154.96	7	197405	39.54	526.413	Y
17	3PL	293.97	7	197405	76.70	526.413	Y
18	3PL	100.60	7	197405	25.01	526.413	Y
19	3PL	264.68	7	197405	68.87	526.413	Y
20	3PL	242.62	7	197405	62.97	526.413	Y
21	3PL	94.06	7	197405	23.27	526.413	Y
22	3PL	669.06	7	197405	176.94	526.413	Y
23	3PL	205.65	7	197405	53.09	526.413	Y
24	3PL	82.28	7	197405	20.12	526.413	Y
25	3PL	210.76	7	197405	54.46	526.413	Y
26	3PL	188.69	7	197405	48.56	526.413	Y
27	3PL	127.55	7	197405	32.22	526.413	Y
28	3PL	134.94	7	197405	34.19	526.413	Y
29	3PL	89.35	7	197405	22.01	526.413	Y
30	3PL	1277.16	7	197405	339.46	526.413	Y
31	2PPC	1524.11	17	197405	258.47	526.413	Y

(Continued on next page)

Table 23. Mathematics Grade 7 Item Fit Statistics (cont.)

Item	Model	Chi Square	DF	Total N	Z_{Q1}	Z_{Q1} critical	Fit OK?
32	2PPC	235.64	17	197405	37.50	526.413	Y
33	2PPC	339.78	17	197405	55.36	526.413	Y
34	2PPC	421.49	17	197405	69.37	526.413	Y
35	2PPC	5455.65	26	197405	752.96	526.413	N
36	2PPC	1305.86	26	197405	177.48	526.413	Y
37	2PPC	2844.88	26	197405	390.91	526.413	Y
38	2PPC	1558.41	26	197405	212.51	526.413	Y

Table 24. Mathematics Grade 8 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	Z_{Q1}	Z_{Q1} critical	Fit OK?
1	3PL	68.56	7	199307	16.45	531.485	Y
2	3PL	55.65	7	199307	13.00	531.485	Y
3	3PL	55.19	7	199307	12.88	531.485	Y
4	3PL	227.82	7	199307	59.02	531.485	Y
5	3PL	20.49	7	199307	3.60	531.485	Y
6	3PL	24.93	7	199307	4.79	531.485	Y
7	3PL	59.26	7	199307	13.97	531.485	Y
8	3PL	131.26	7	199307	33.21	531.485	Y
9	3PL	66.78	7	199307	15.98	531.485	Y
10	3PL	74.07	7	199307	17.93	531.485	Y
11	3PL	35.34	7	199307	7.57	531.485	Y
12	3PL	77.18	7	199307	18.76	531.485	Y
13	3PL	160.54	7	199307	41.03	531.485	Y
14	3PL	107.18	7	199307	26.77	531.485	Y
15	3PL	30.08	7	199307	6.17	531.485	Y
16	3PL	17.05	7	199307	2.69	531.485	Y
17	3PL	49.03	7	199307	11.23	531.485	Y
18	3PL	60.84	7	199307	14.39	531.485	Y
19	3PL	21.57	7	199307	3.89	531.485	Y
20	3PL	55.30	7	199307	12.91	531.485	Y
21	3PL	30.58	7	199307	6.30	531.485	Y
22	3PL	144.80	7	199307	36.83	531.485	Y
23	3PL	113.92	7	199307	28.57	531.485	Y
24	3PL	129.01	7	199307	32.61	531.485	Y
25	3PL	32.72	7	199307	6.87	531.485	Y
26	3PL	90.26	7	199307	22.25	531.485	Y
27	3PL	22.90	7	199307	4.25	531.485	Y
28	2PPC	401.30	17	199307	65.91	531.485	Y
29	2PPC	6258.79	17	199307	1070.46	531.485	N
30	2PPC	1213.99	17	199307	205.28	531.485	Y
31	2PPC	1219.23	17	199307	206.18	531.485	Y
32	2PPC	333.67	26	199307	42.67	531.485	Y

(Continued on next page)

Table 24. Mathematics Grade 8 Item Fit Statistics (cont.)

Item	Model	Chi Square	DF	Total N	Z_{Q1}	Z_{Q1} critical	Fit OK?
33	2PPC	317.93	26	199307	40.48	531.485	Y
34	2PPC	318.23	17	199307	51.66	531.485	Y
35	2PPC	4631.25	17	199307	791.34	531.485	N
36	2PPC	3457.77	17	199307	590.09	531.485	N
37	2PPC	198.44	17	199307	31.12	531.485	Y
38	2PPC	1059.12	17	199307	178.72	531.485	Y
39	2PPC	468.62	17	199307	77.45	531.485	Y
40	2PPC	617.91	17	199307	103.05	531.485	Y
41	2PPC	194.91	17	199307	30.51	531.485	Y
42	2PPC	1172.34	26	199307	158.97	531.485	Y
43	2PPC	1928.09	26	199307	263.77	531.485	Y
44	2PPC	377.46	26	199307	48.74	531.485	Y
45	2PPC	274.90	26	199307	34.52	531.485	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 is not dependent upon their response on 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 the Grades 3–8 Mathematics Tests and no items were found to be locally dependent in Grades 3–7. In Grade 8, two pairs of items were found to be locally dependent: items 2 and 11 (Q_3 = 0.377) and items 32 and 44 (Q_3 = 0.404). The magnitudes of these statistics were not sufficient to warrant any concern. Anchor items were excluded from Q_3 computation.

Scaling and Equating

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

In the first equating procedure, the new 2010 OP forms were pre-equated to the corresponding 2009 assessments. Prior to pre-equating, the FT items administered in 2009 were placed onto the OP scales in each grade. The equating of 2009 FT items to the 2009 OP scales was conducted via common examinees. FT items that were eligible for future OP administrations were then included in the NYS item pool. Other items in the NYS item pool were items field tested in 2008, 2007, 2006, and 2005. All items field tested between 2005 and 2008 were also equated to the NYS OP scales. For more details on equating of FT items to the NYS OP scales, refer to page 44 of *New York State Testing Program 2006: Grades 3 through 8 Mathematics Field Test Technical Report*.

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

- Content coverage of each form matched the test blueprint
- Psychometric properties of the items:
  - item fit
  - differential item functioning
  - item difficulty
  - item discrimination
  - omit rates
- Test Characteristic Curve (TCC) and Standard Error (SE) curve alignment of the 2010 forms with the target 2009 OP forms (note that the 2009 OP TCC and SE curves were based on OP parameters, and the 2010 TCC and SE curves were based on FT parameters transformed to the NYS OP scale).

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

a) at least 10 characters of last name match (including blank spaces)
b) at least 5 characters of first name match (including blank spaces)
c) gender must be the same or one must be blank
d) school Bedscode must be the same or one must be blank

e) 2 of 3 parts of date of birth (MM or DD or YY) must be the same or one must be blank

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

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

c-parameters of anchor items were fixed to their 2008 values.

The relationships between the new and old linear transformation constants that are applied to the original ability metric parameters to place them onto the NYS scale via the Stocking and Lord (1983) method are presented below:

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

where

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

The A and B values are derived from the input (2008 FT anchor parameter estimates) and estimate (2010 FT anchor parameter estimates) values of anchor items. Anchor input values are known item parameter estimates entered into equating. Anchor estimate 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 $M1$ and $M2$ transformation parameters obtained in the Stocking and Lord (1983) equating process were used to transform item parameters obtained in the calibration process into the final scale score metric. Table 25 presents the 2010 OP transformation parameters for New York State Grades 3–8 Mathematics Tests.

Table 25. NYSTP Mathematics 2010 Final Transformation Constants

Grade	*M*1	*M*2
3	17.2059	687.6159
4	28.9457	685.0496
5	26.3727	684.5696
6	27.4497	680.5049
7	26.0047	678.5270
8	26.5541	676.5651

Anchor Item Security

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

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 procedure 3 was 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 2008 and 2010 FT anchor form administrations.

2. Correlations of anchor input and estimate of *a*- and *b*-parameters. Correlations of anchor input and estimate of *a*- and *b*-parameters and p-values were evaluated for magnitude. Ideally, the correlations between anchor input and estimate for *a*-parameter should be at least 0.80 and at least 0.90 for *b*-parameter.

3. Iterative linking using Stocking and Lord's (1983) TCC method. This procedure, called the TCC method, minimizes the mean squared difference between the two TCCs: one based on 2008 FT anchor estimates and the other on transformed estimates from the 2010 equating of OP test forms. Differential item performance was evaluated by examining previous (input) and transformed (estimated) item parameters. Items with an absolute difference of parameters greater than two times the root mean square deviation were flagged.

In all cases, the overall TCC alignment for anchor set input and estimate parameters was very good. Correlations for *b*-parameter input and estimates ranged from 0.94 for Grade 8 to 0.98 for Grades 4 and 5. Correlations for *a*-parameter input and estimate ranged from 0.64 for Grade 3 to 0.93 for Grade 6. Only correlation between *a*-parameter input and estimates for Grade 3 was below the NYS criterion.

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

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.

Item Parameters

The OP test item parameters were estimated by the software PARDUX (Burket, 2002) and are presented in Tables 26–31. The parameter estimates are expressed in scale score metrics and are defined below:

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

As described in 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 *mj* – 1 independent gammas and one alpha for a total of m_j independent parameters estimated for each item, while there is one *a*-parameter and one *b*-parameter per item in the 3PL model.

Table 26. Grade 3 2010 Operational Item Parameter Estimates

Item	Max Pts	*a*-par/ alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3
1	1	0.0429	643.5340	0.1368	
2	1	0.0453	643.5304	0.1935	
3	1	0.0582	677.4951	0.1289	
4	1	0.0620	669.8259	0.2143	
5	1	0.0422	644.1674	0.1193	
6	1	0.0547	659.4844	0.1386	
7	1	0.0591	662.1629	0.2167	
8	1	0.0334	671.0156	0.1093	
9	1	0.0701	672.4298	0.1406	
10	1	0.0574	651.0522	0.1718	
11	1	0.0528	638.8137	0.1935	
12	1	0.0680	674.2241	0.1759	
13	1	0.0362	666.5793	0.0776	
14	1	0.0284	652.3484	0.1935	
15	1	0.0501	645.1137	0.2405	
16	1	0.0463	644.7352	0.0931	
17	1	0.0486	657.0290	0.1291	
18	1	0.0463	654.7067	0.2717	
19	1	0.0628	652.1354	0.0435	
20	1	0.0408	643.1226	0.2424	
21	1	0.0442	638.7993	0.1935	
22	1	0.0497	640.8296	0.0921	
23	1	0.0597	649.2918	0.1935	
24	1	0.0504	671.7261	0.3312	
25	1	0.0689	670.3083	0.1763	
26	2	0.0552	36.9134	35.0364	
27	2	0.0782	51.6662	51.1628	
28	2	0.0720	45.9968	47.8061	
29	2	0.0412	27.3187	27.6046	
30	3	0.0690	43.0003	47.1217	45.7566
31	3	0.0651	41.7692	44.7221	43.9237

Table 27. Grade 4 2010 Operational Item Parameter Estimates

Item	Max Pts	*a*-par/ alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3
1	1	0.0199	657.1581	0.1585	
2	1	0.0252	599.5527	0.2000	
3	1	0.0210	606.2830	0.2000	
4	1	0.0281	615.5534	0.0554	
5	1	0.0259	647.9416	0.2053	
6	1	0.0272	631.3243	0.0496	
7	1	0.0288	618.6633	0.0547	
8	1	0.0276	660.7518	0.1324	
9	1	0.0175	593.1804	0.2000	
10	1	0.0370	656.7765	0.1077	
11	1	0.0278	650.8040	0.1168	
12	1	0.0377	669.0377	0.0867	
13	1	0.0318	641.5935	0.0439	
14	1	0.0280	627.1210	0.0554	
15	1	0.0210	646.3292	0.2057	
16	1	0.0330	674.2675	0.1720	
17	1	0.0325	654.8614	0.2783	
18	1	0.0444	665.1143	0.1917	
19	1	0.0285	671.4348	0.1886	
20	1	0.0267	629.5389	0.1068	
21	1	0.0455	690.4255	0.2589	
22	1	0.0255	636.8994	0.4985	
23	1	0.0265	624.6312	0.0554	
24	1	0.0435	698.1311	0.2181	
25	1	0.0311	694.6043	0.2520	
26	1	0.0261	646.6499	0.0419	
27	1	0.0276	673.2459	0.2423	
28	1	0.0292	667.5664	0.1901	
29	1	0.0223	702.2383	0.2324	
30	1	0.0502	666.1062	0.1975	
31	2	0.0418	25.2275	27.2015	
32	2	0.0556	34.7447	37.9828	
33	2	0.0503	33.8866	31.9781	
34	2	0.0506	34.3620	32.7455	
35	2	0.0458	29.3037	29.6210	
36	2	0.0274	16.0673	17.6658	
37	2	0.0442	31.9251	29.2009	
38	3	0.0432	27.2325	27.9818	28.4919
39	3	0.0402	25.3841	26.9791	26.6404
40	2	0.0403	27.1854	26.6246	

(Continued on next page)

Table 27. Grade 4 2010 Operational Item Parameter Estimates (cont.)

Item	Max Pts	*a*-par/ alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3
41	2	0.0510	32.8165	33.9278	
42	2	0.0450	28.8906	29.4519	
43	2	0.0483	33.0436	30.8712	
44	2	0.0451	30.0465	29.0482	
45	2	0.0309	19.4551	21.2902	
46	2	0.0463	30.8082	32.1486	
47	3	0.0251	15.1661	16.9385	16.7609
48	3	0.0379	25.2729	25.8987	25.3810

Table 28. Grade 5 2010 Operational Item Parameter Estimates

Item	Max Pts	*a*-par/ alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3
1	1	0.0344	662.1100	0.1766	
2	1	0.0470	666.2870	0.2011	
3	1	0.0285	632.7664	0.3403	
4	1	0.0317	634.2208	0.2332	
5	1	0.0374	666.6138	0.2352	
6	1	0.0203	606.1445	0.2000	
7	1	0.0378	656.6349	0.1982	
8	1	0.0367	661.5861	0.1417	
9	1	0.0575	671.3974	0.2072	
10	1	0.0313	672.7298	0.2367	
11	1	0.0455	659.7298	0.2821	
12	1	0.0248	674.6860	0.2471	
13	1	0.0293	640.2851	0.1915	
14	1	0.0310	681.5958	0.2815	
15	1	0.0291	677.3114	0.1125	
16	1	0.0235	636.9893	0.0704	
17	1	0.0481	674.2779	0.1490	
18	1	0.0312	656.3167	0.3737	
19	1	0.0236	609.7018	0.2000	
20	1	0.0316	684.8466	0.1392	
21	1	0.0582	665.8516	0.1025	
22	1	0.0460	666.1953	0.1762	
23	1	0.0328	656.0915	0.3464	
24	1	0.0245	665.2591	0.0638	
25	1	0.0324	641.4930	0.1485	
26	1	0.0215	623.6997	0.1073	

(Continued on next page)

Table 28. Grade 5 2010 Operational Item Parameter Estimates (cont.)

Item	Max Pts	*a*-par/ alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3
27	2	0.0323	20.4109	20.5288	
28	2	0.0381	24.8005	25.8971	
29	2	0.0314	18.1400	21.3755	
30	2	0.0492	33.1148	32.7733	
31	3	0.0416	26.7895	28.2833	27.6546
32	3	0.0314	19.0650	20.4943	20.7537
33	3	0.0300	18.3262	19.7180	20.8269
34	3	0.0295	19.8569	20.0683	19.7502

Table 29. Grade 6 2010 Operational Item Parameter Estimates

Item	Max Pts	*a*-par/ alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3
1	1	0.0497	643.0248	0.3678	
2	1	0.0403	635.6012	0.2001	
3	1	0.0242	613.4509	0.2000	
4	1	0.0233	643.7159	0.1676	
5	1	0.0249	671.2623	0.4038	
6	1	0.0107	592.7794	0.2000	
7	1	0.0336	650.6897	0.2222	
8	1	0.0575	677.6850	0.2880	
9	1	0.0423	683.1460	0.1813	
10	1	0.0276	668.7889	0.2468	
11	1	0.0273	647.2859	0.1598	
12	1	0.0258	692.1872	0.4418	
13	1	0.0420	670.4437	0.2560	
14	1	0.0334	651.0574	0.1349	
15	1	0.0232	671.9728	0.1662	
16	1	0.0468	671.3289	0.3307	
17	1	0.0264	628.5331	0.0782	
18	1	0.0386	649.6210	0.1997	
19	1	0.0417	653.7484	0.3179	
20	1	0.0558	676.7208	0.2335	
21	1	0.0360	640.9793	0.1763	
22	1	0.0389	680.2298	0.0787	
23	1	0.0539	661.7233	0.3726	
24	1	0.0353	680.8878	0.1502	
25	1	0.0229	621.3586	0.2810	
26	2	0.0418	27.1392	25.5558	

(Continued on next page)

Table 29. Grade 6 2010 Operational Item Parameter Estimates (cont.)

Item	Max Pts	*a*-par/ alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3
27	2	0.0392	25.7723	24.1893	
28	2	0.0474	31.6841	32.1798	
29	2	0.0330	20.8881	21.6872	
30	2	0.0513	35.2639	35.8689	
31	2	0.0364	25.2890	22.5780	
32	3	0.0437	29.0004	27.5589	28.7132
33	3	0.0231	15.5270	14.9519	15.2313
34	3	0.0319	20.2776	22.1957	23.3250
35	3	0.0268	16.6617	17.7049	16.7717

Table 30. Grade 7 2010 Operational Item Parameter Estimates

Item	Max Pts	*a*-par/ alpha	*b*-par/ gamma1	c-par/ gamma2	gamma3
1	1	0.0230	630.0510	0.2000	
2	1	0.0406	647.8977	0.1219	
3	1	0.0354	649.3287	0.1329	
4	1	0.0223	707.4928	0.1928	
5	1	0.0298	695.3773	0.4354	
6	1	0.0498	700.9687	0.3701	
7	1	0.0313	658.7432	0.2289	
8	1	0.0348	623.1909	0.0528	
9	1	0.0367	646.9689	0.3531	
10	1	0.0490	696.0953	0.1811	
11	1	0.0454	645.8022	0.3251	
12	1	0.0378	644.7668	0.3078	
13	1	0.0295	639.9033	0.2632	
14	1	0.0342	642.2297	0.1426	
15	1	0.0276	682.3614	0.2181	
16	1	0.0468	667.2548	0.0948	
17	1	0.0176	603.7209	0.2000	
18	1	0.0321	688.7643	0.0545	
19	1	0.0523	685.1267	0.2014	
20	1	0.0525	645.1108	0.1313	
21	1	0.0522	649.3943	0.3802	
22	1	0.0246	638.3118	0.1719	
23	1	0.0314	666.8854	0.4887	
24	1	0.0368	673.3110	0.1996	
25	1	0.0498	678.8263	0.1960	
26	1	0.0198	666.7363	0.1719	

(Continued on next page)

Table 30. Grade 7 2010 Operational Item Parameter Estimates (cont.)

Item	Max Pts	*a*-par/ alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3
27	1	0.0454	659.7665	0.1997	
28	1	0.0480	662.0001	0.1579	
29	1	0.0429	676.2098	0.2603	
30	1	0.0150	617.8142	0.2000	
31	2	0.0364	22.9547	22.9821	
32	2	0.0264	19.7927	17.7297	
33	2	0.0399	25.2238	25.2876	
34	2	0.0424	28.4476	27.5373	
35	3	0.0471	29.8681	30.2752	31.5725
36	3	0.0447	30.4222	29.0290	31.5011
37	3	0.0406	26.6248	29.8723	27.7607
38	3	0.0557	37.2339	38.1539	39.2778

Table 31. Grade 8 2010 Operational Item Parameter Estimates

Item	Max Pts	*a*-par/ alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3
1	1	0.0216	607.5853	0.2000	
2	1	0.0440	641.5895	0.0534	
3	1	0.0538	656.1551	0.2078	
4	1	0.0390	636.8172	0.0761	
5	1	0.0351	649.2824	0.2149	
6	1	0.0361	643.9223	0.1743	
7	1	0.0408	651.5114	0.1346	
8	1	0.0542	666.1794	0.1779	
9	1	0.0511	670.8489	0.2283	
10	1	0.0518	676.8550	0.2360	
11	1	0.0410	638.0939	0.0855	
12	1	0.0404	657.2252	0.2491	
13	1	0.0275	669.0886	0.3379	
14	1	0.0324	662.4160	0.1557	
15	1	0.0367	644.8004	0.1738	
16	1	0.0334	671.4647	0.2098	
17	1	0.0213	649.1750	0.2002	
18	1	0.0350	622.7054	0.0951	
19	1	0.0329	640.5349	0.1356	
20	1	0.0312	648.7996	0.2549	
21	1	0.0296	671.1216	0.1616	
22	1	0.0262	663.8521	0.2288	

(Continued on next page)

Table 31. Grade 8 2010 Operational Item Parameter Estimates (cont.)

Item	Max Pts	*a*-par/ alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3
23	1	0.0429	612.2216	0.1028	
24	1	0.0588	677.5839	0.1872	
25	1	0.0459	647.6072	0.1582	
26	1	0.0505	648.4337	0.1174	
27	1	0.0410	654.5309	0.2073	
28	2	0.0492	32.3007	32.3782	
29	2	0.0671	41.7314	45.2367	
30	2	0.0552	35.7454	36.1195	
31	2	0.0539	34.1025	36.3185	
32	3	0.0442	28.4860	29.8020	28.6940
33	3	0.0408	26.4452	26.7397	27.1866
34	2	0.0619	41.1900	40.5681	
35	2	0.0316	22.4186	19.6787	
36	2	0.0401	25.9169	27.6088	
37	2	0.0531	35.7190	35.5647	
38	2	0.0439	28.2356	30.2621	
39	2	0.0425	28.2376	28.0674	
40	2	0.0432	29.7576	28.5003	
41	2	0.0455	30.9013	29.9711	
42	3	0.0336	21.6121	21.7756	21.8479
43	3	0.0255	17.9597	16.8604	16.9462
44	3	0.0443	28.4541	29.4327	28.7211
45	3	0.0316	19.5159	21.3604	20.9311

Test Characteristic Curves

Test Characteristic Curves (TCCs) provide an overview of the test in the IRT scale score metric. The 2009 and 2010 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 2009 and 2010 TCCs and SE curves are presented in Figures 1–6. Following the adoption of the chain-equating method by New York State, the TCCs for new OP test forms are compared to the previous year's TCCs rather than to the baseline 2006 test form TCCs. Therefore, the 2009 OP curves are considered to be target curves for the 2010 OP test TCCs. This equating process enables the comparisons of impact results (i.e., percentages of examinees at and above each proficiency level) between adjacent test administrations. Note that in all figures the blue TCCs and SE curves represent the 2010 OP test and pink TCCs and SE curves represent the 2009 OP test. The *x*-axis is the ability scale expressed in a scale score metric, with the lower and upper bounds established in Year 1 of test administration and presented in the lower corners of the graphs. The *y*-axis is the proportion of the test that the students can answer correctly.

Figure 1. Grade 3 Mathematics 2009 and 2010 OP TCCs and SE

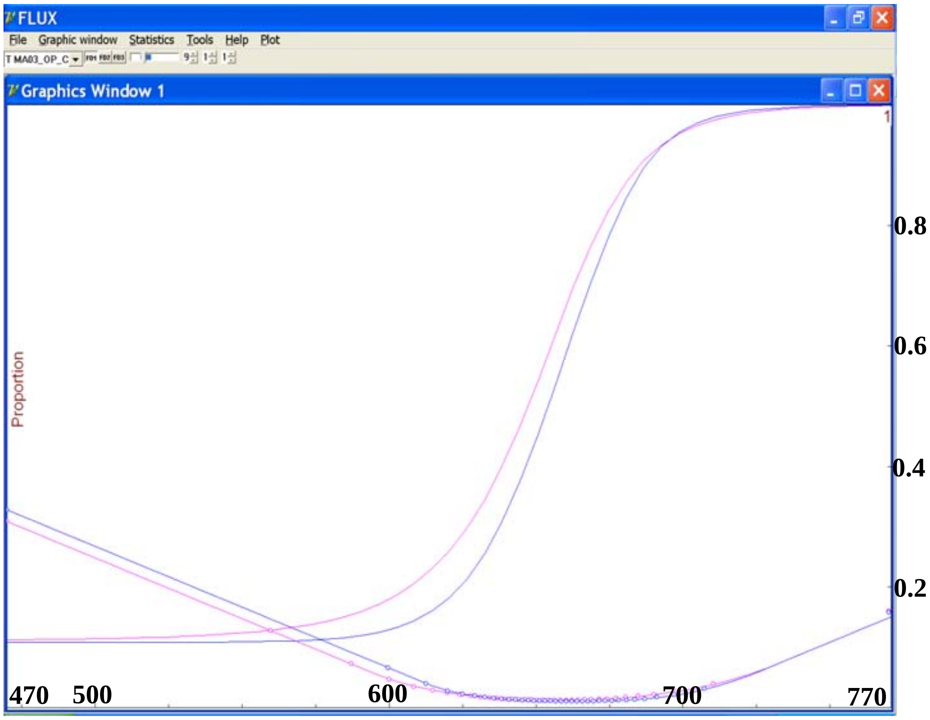


Figure 2. Grade 4 Mathematics 2009 and 2010 OP TCCs and SE

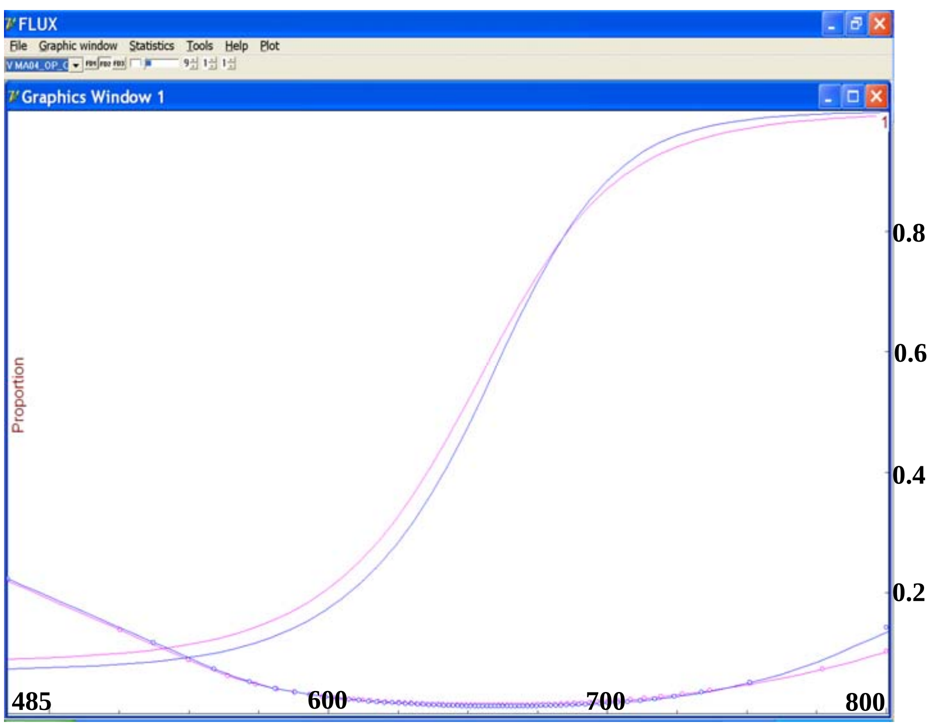


Figure 3. Grade 5 Mathematics 2009 and 2010 OP TCCs and SE

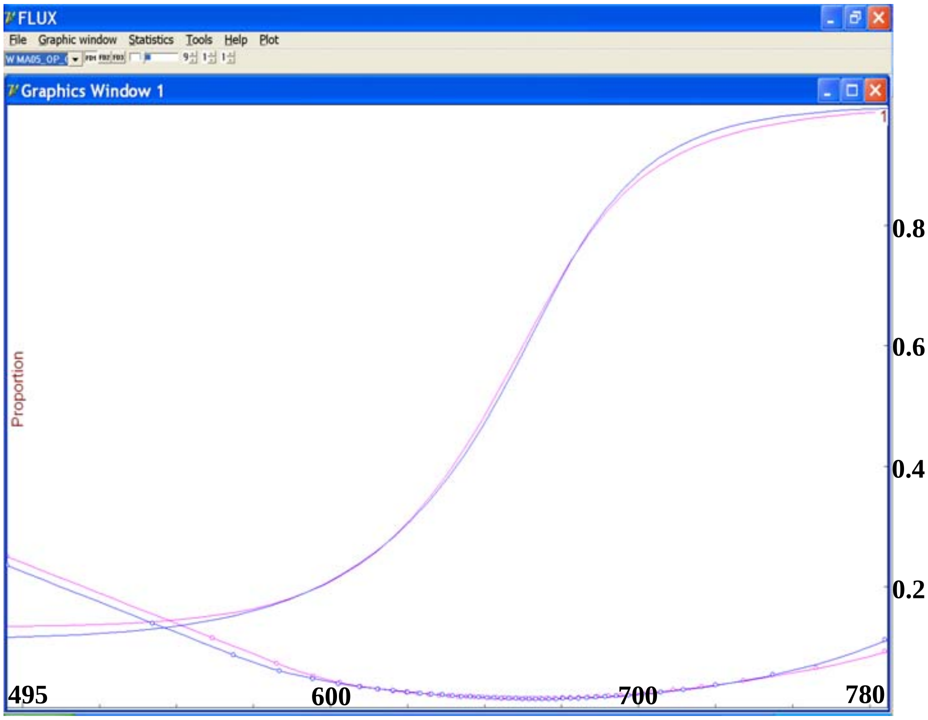


Figure 4. Grade 6 Mathematics 209 and 2010 OP TCCs and SE

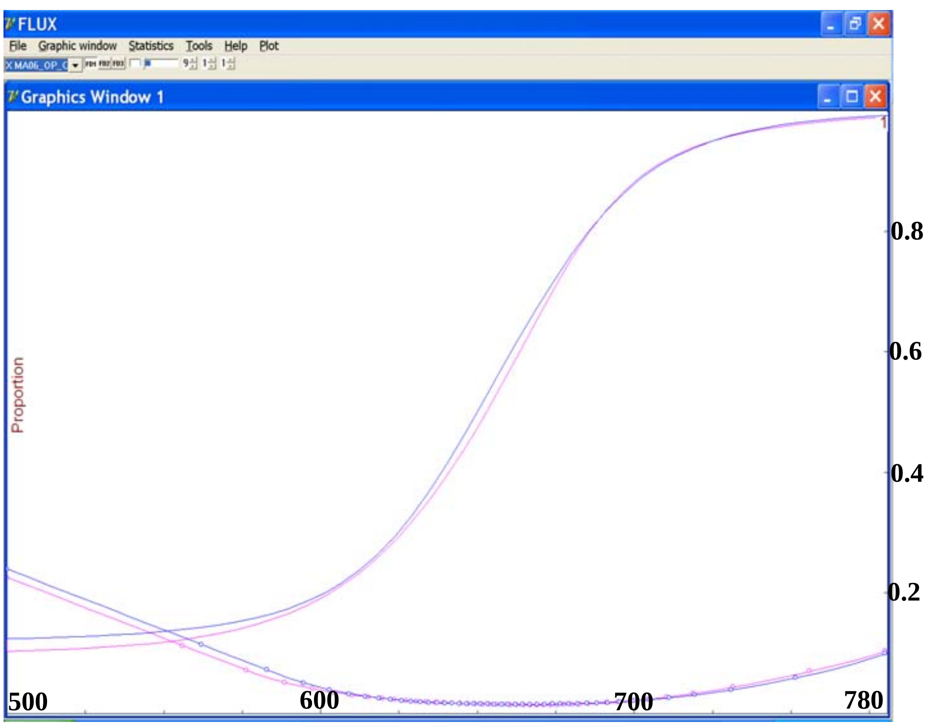


Figure 5. Grade 7 Mathematics 2009 and 2010 OP TCCs and SE

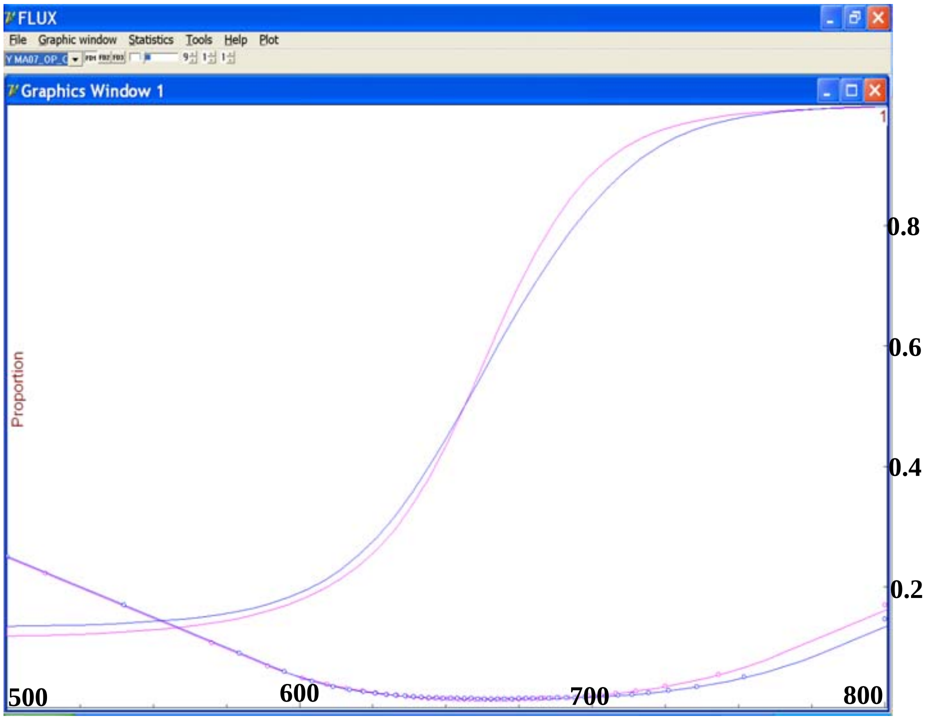
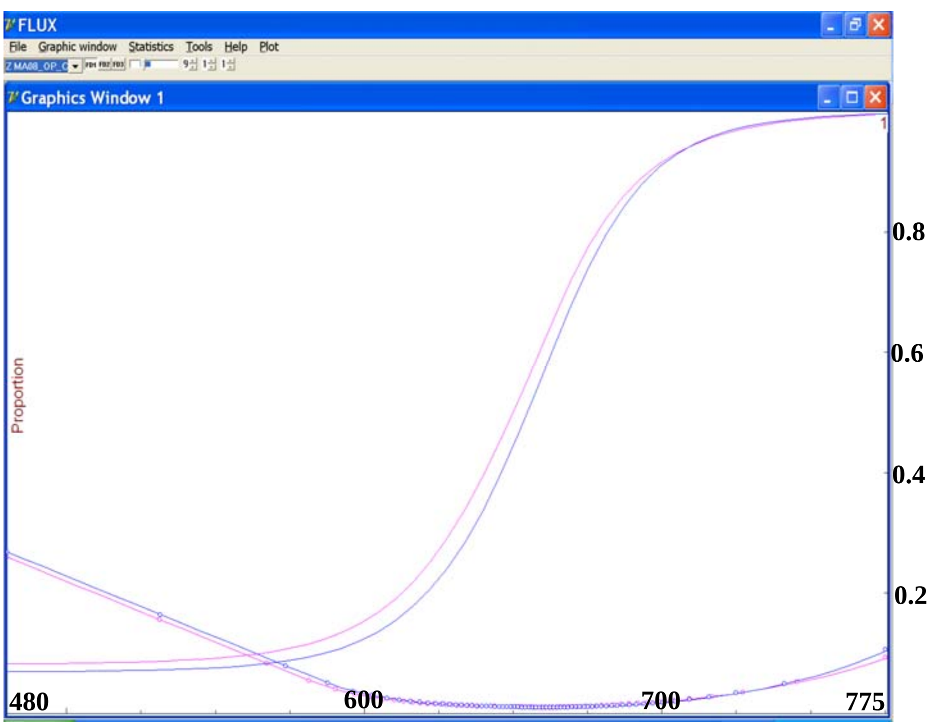


Figure 6. Grade 8 Mathematics 2009 and 2010 OP TCCs and SE



As seen in Figures 1, 3, 4, and 6, very good alignments of the 2009 and 2010 TCCs and SE curves were found for Grades 3, 5, 6, and 8. The TCCs for Grades 4 and 7 were somewhat less well-aligned at the lower and upper ends of the scale, indicating that the 2010 Grade 4 form tended to be slightly more difficult for lower-ability students and easier for the high-ability students, and the 2010 Grade 7 form tended to be slightly more difficult for high-ability students and easier for the 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 the 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:

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

where

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

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

Table 32. Grade 3 Raw Score-to-Scale Score (with Standard Error)

Raw Score	Scale Score	Standard Error
0	470	162
1	470	162
2	470	162
3	470	162
4	470	162
5	600	33
6	613	20
7	620	14
8	625	11
9	629	9
10	633	8
11	636	8
12	638	7
13	641	7
14	643	6
15	645	6

(Continued on next page)

Table 32. Grade 3 Raw Score-to-Scale Score (with Standard Error) (cont.)

Raw Score	Scale Score	Standard Error
16	647	6
17	649	6
18	651	6
19	653	5
20	655	5
21	657	5
22	659	5
23	660	5
24	662	5
25	664	5
26	666	5
27	668	5
28	670	5
29	672	5
30	674	5
31	676	6
32	678	6
33	681	6
34	684	7
35	687	7
36	691	8
37	697	10
38	707	15
39	770	78

Table 33. Grade 4 Raw Score-to-Scale Score (with Standard Error)

Raw Score	Scale Score	Standard Error
0	485	110
1	485	110
2	485	110
3	485	110
4	485	110
5	485	110
6	537	58
7	559	36
8	572	26
9	581	21
10	588	17
11	594	15
12	599	14

(Continued on next page)

Table 33. Grade 4 Raw Score-to-Scale Score (with Standard Error) (cont.)

Raw Score	Scale Score	Standard Error
13	603	12
14	607	12
15	611	11
16	614	10
17	617	10
18	620	9
19	623	9
20	625	8
21	628	8
22	630	8
23	632	8
24	634	7
25	636	7
26	638	7
27	640	7
28	641	7
29	643	7
30	645	6
31	646	6
32	648	6
33	650	6
34	651	6
35	653	6
36	654	6
37	655	6
38	657	6
39	658	6
40	660	6
41	661	6
42	663	6
43	664	6
44	665	6
45	667	6
46	668	6
47	670	6
48	671	6
49	673	6
50	675	6
51	676	6
52	678	6
53	680	6

(Continued on next page)

Table 33. Grade 4 Raw Score-to-Scale Score (with Standard Error) (cont.)

Raw Score	Scale Score	Standard Error
54	682	7
55	683	7
56	685	7
57	688	7
58	690	7
59	692	7
60	695	8
61	697	8
62	700	9
63	704	9
64	707	10
65	712	11
66	717	12
67	724	14
68	734	17
69	751	26
70	800	71

Table 34. Grade 5 Raw Score-to-Scale Score (with Standard Error)

Raw Score	Scale Score	Standard Error
0	495	116
1	495	116
2	495	116
3	495	116
4	495	116
5	495	116
6	542	69
7	568	43
8	583	30
9	594	24
10	603	20
11	609	17
12	615	15
13	620	14
14	625	13
15	629	12
16	633	11
17	636	10

(Continued on next page)

Table 34. Grade 5 Raw Score-to-Scale Score (with Standard Error) (cont.)

Raw Score	Scale Score	Standard Error
18	640	10
19	643	9
20	645	9
21	648	8
22	651	8
23	653	8
24	656	8
25	658	7
26	660	7
27	662	7
28	664	7
29	667	7
30	669	7
31	671	7
32	673	7
33	676	7
34	678	7
35	680	8
36	683	8
37	686	8
38	689	9
39	693	9
40	697	10
41	701	11
42	707	13
43	714	15
44	725	18
45	744	27
46	780	55

Table 35. Grade 6 Raw Score-to-Scale Score (with Standard Error)

Raw Score	Scale Score	Standard Error
0	500	119
1	500	119
2	500	119
3	500	119
4	500	119
5	500	119
6	500	119
7	562	57
8	583	36
9	595	25
10	603	19
11	609	16
12	614	14
13	619	13
14	622	12
15	626	11
16	629	10
17	632	10
18	635	9
19	637	9
20	640	9
21	642	8
22	644	8
23	647	8
24	649	8
25	651	8
26	653	8
27	655	8
28	658	8
29	660	8
30	662	8
31	664	8
32	667	8
33	669	8
34	671	8
35	674	8
36	676	8
37	679	8
38	682	8
39	685	8

(Continued on next page)

Table 35. Grade 6 Raw Score-to-Scale Score (with Standard Error) (cont.)

Raw Score	Scale Score	Standard Error
40	688	9
41	692	9
42	695	10
43	700	11
44	705	12
45	711	13
46	719	15
47	731	20
48	751	30
49	780	50

Table 36. Grade 7 Raw Score-to-Scale Score (with Standard Error)

Raw Score	Scale Score	Standard Error
0	500	124
1	500	124
2	500	124
3	500	124
4	500	124
5	500	124
6	500	124
7	540	84
8	579	45
9	595	29
10	604	21
11	611	17
12	617	15
13	622	13
14	626	12
15	630	11
16	633	10
17	636	9
18	639	9
19	642	8
20	644	8
21	647	8
22	649	8
23	652	8

(Continued on next page)

Table 36. Grade 7 Raw Score-to-Scale Score (with Standard Error) (cont.)

Raw Score	Scale Score	Standard Error
24	654	7
25	656	7
26	659	7
27	661	7
28	663	7
29	665	7
30	668	7
31	670	7
32	672	7
33	675	7
34	677	7
35	680	7
36	683	8
37	685	8
38	688	8
39	691	8
40	694	8
41	697	9
42	701	9
43	705	9
44	709	10
45	714	11
46	719	12
47	726	14
48	736	17
49	752	25
50	800	73

Table 37. Grade 8 Raw Score-to-Scale Score (with Standard Error)

Raw Score	Scale Score	Standard Error
0	480	133
1	480	133
2	480	133
3	480	133
4	480	133
5	532	81

(Continued on next page)

Table 37. Grade 8 Raw Score-to-Scale Score (with Standard Error) (cont.)

Raw Score	Scale Score	Standard Error
6	574	39
7	588	25
8	596	18
9	603	15
10	608	13
11	612	11
12	615	10
13	619	9
14	621	9
15	624	8
16	626	8
17	629	7
18	631	7
19	633	7
20	634	7
21	636	6
22	638	6
23	639	6
24	641	6
25	642	6
26	644	6
27	645	6
28	647	6
29	648	5
30	649	5
31	650	5
32	652	5
33	653	5
34	654	5
35	655	5
36	657	5
37	658	5
38	659	5
39	660	5
40	661	5
41	662	5
42	664	5
43	665	5
44	666	5
45	667	5
46	669	5
47	670	5

(Continued on next page)

Table 37. Grade 8 Raw Score-to-Scale Score (with Standard Error) (cont.)

Raw Score	Scale Score	Standard Error
48	671	5
49	672	5
50	674	5
51	675	6
52	677	6
53	678	6
54	680	6
55	681	6
56	683	6
57	685	7
58	687	7
59	689	7
60	691	7
61	694	8
62	697	8
63	700	9
64	704	10
65	709	11
66	716	13
67	725	17
68	741	25
69	775	52

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 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 2010 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. Table 38 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 38. SPI Target Ranges

Grade	Objective	# Items	Total Points	Level III Cut SPI Target Range
3	1	16	19	79–92
	2	3	4	87–95
	3	4	5	75–86
	4	4	4	90–97
	5	4	7	81–94
4	1	24	35	62–74
	2	7	11	68–78
	3	5	8	76–82
	4	8	10	74–82
	5	4	6	75–84
5	1	13	15	66–80
	2	6	8	69–77
	3	9	12	65–78
	4	3	6	48–64
	5	3	5	62–74
6	1	12	18	54–67
	2	8	12	71–84
	3	6	7	64–78
	4	4	5	79–90
	5	5	7	73–84
7	1	12	16	46–61
	2	6	7	61–69
	3	5	8	39–49
	4	4	5	66–79
	5	11	14	70–81
8	1	5	9	62–72
	2	19	27	66–78
	3	16	24	65–75
	4	5	9	73–81

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 to improve 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 & 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

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

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

The following groups were analyzed using the IRT-based DIF analysis: Female, Male, Asian, Black, Hispanic, White, High Needs districts (by NRC code), Low Needs districts (by NRC code), Spanish language test version, and ELLs. 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; one item was flagged on the Grade 4 test; two items were flagged on the Grade 5 test; three items were flagged on the Grade 6 test; five items were flagged on the Grade 7 test; and four items were flagged on the Grade 8 test, as is shown in Table 39.

Table 39. Number of Items Flagged for DIF by the Linn-Harnisch Method

Grade	Number of Flagged Items
3	0
4	1
5	2
6	3
7	5
8	4

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

Section VII: Proficiency Level Cut Score Adjustment

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

Proficiency Cut Score Adjustment Process

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

Although the score scales did not change, the NYSED together with TAG and CTB conducted a series of studies and surveys concerning student cut scores and student proficiency. The following steps were taken to set new 2010 cut scores:

1. Grade 8 Mathematics proficiency Level II cut score was raised to reflect 75% probability of achieving a Math A Regents score of 65 or above. Grade 8 Mathematics proficiency Level III cut score was raised to reflect 75% probability of achieving a Math A Regents score of 80 or above. As a result, a Grade 8 student scoring at or above the new Level 2 standard is on track to pass the math Regents exam required for high school graduation. A Grade 8 student scoring at or above the new Level 3 standard is on track to earn a college-ready score on the Regents exam. The alignment of Level II and Level III cut scores with student performance on the Regents exam was conducted by the NYSED, and the resulting cut scores were provided to CTB/McGraw-Hill. The Grade 8 Level II and Level III cut scores are 638 and 672 respectvely. Details on setting Grade 8 Level II and Level III cut scores are available online at http://usny.nysed.gov/scoring_changes/.

2. Grade 8 Mathematics proficiency Levels II and III cut scores were further adjusted to account for additional instructional time between 2009 and 2010 administration windows, as the 2010 test administration occurred in May instead of March administration in 2009. After the time adjustment the Grade 8 Level II and Level III cut scores are 639 and 673 as shown in Table 41.

3. Grades 3–7 Levels II and III cut scores were established to reflect the corresponding academic rigor applied to the Grade 8 adjusted cut scores by holding the national percentile rank associated with each grade's cut score equal to the national percentile rank associated with the Grade 8 cut score. The national percentile ranks were computed based on NYS student performance on nationally standardized and vertically scaled test items from CTB/McGraw-Hill's *TerraNova* test battery (CTB, 1999, 2000, 2006) that were administered as part of the Secure Anchor/Audit (SAA) test one week after OP tests. The percentile ranks for Grade 8 Levels II and III cut scores are 19 and 54, respectively, and the Levels II and III cut scores for the remaining grades were set to correspond to the same percentile ranks. The

concordance tables between OP scale scores and *TerraNova* scores were produced to aid the cut score adjustment process and relate the test scores on the NYS scale to the test scores on the *TerraNova* scale. The concordant scores are defined as those having the same percentile rank with respect to the group of students who took the on grade SAA tests. The concordance tables can be found in Appendix H. The national percentile ranks corresponding to the TerraNova scores are also presented in the concordance tables. Linear interpolation was used to locate the OP cut scores associated with national percentile ranks 19 and 54 if they are not available in the tables

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

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

Adjustment of 2009 Cut Scores to Reflect the 2010 Administration Window

In order to adjust the 2009 cut scores to reflect the 2010 test administration window, student growth within a school year was estimated using the data from the NYS student performance on the CTB/McGraw-Hill's *TerraNova* Mathematics items contained in the Secure Anchor/Audit test administered in 2010. An assumption was made that NYS student growth is similar to the growth pattern obtained from a national sample. The estimation was supported by the *TerraNova* norms available for all quarter-months of the school year. Growth between the 24th and 32nd quarter-months was computed based on NYS student performance on the *TerraNova* Mathematics items. The amount of growth on *TerraNova* items was then expressed in standard deviation units and translated back to NYS OP scales. As the last step, the number of scale score points reflecting amount of growth between the two administration windows on the NYS scales was computed and added to the 2009 OP cut scores to derive the time-adjusted cut scores.

The data analysis steps employed in this procedure are described in detail below and the results of each step are presented in Table 40.

1) The 2010 Anchor/Audit item responses were merged at the student level with the 2010 OP data. The NYS Mathematics OP items and the Mathematics items in the Anchor/Audit forms were equated to the *TerraNova* Mathematics scale by using *TerraNova* parameters for Anchor/Audit mathematics items as anchors in the Stocking and Lord (1983) equating method.
2) Item pattern scores were computed for all students who took both the Anchor/Audit forms and OP test based on their responses to the NYS OP items and Anchor/Audit items.
3) Student scores from step 2 were used to compute mean scale scores on the *TerraNova* scale (these scores are presented in column 1).
4) Mean scale scores from step 3 were used to find normative information (national percentile rank) based on the 2007 *TerraNova* national norms. These percentile ranks are presented in column 2 for the quarter-month in which the tests were administered.

The NYS Mathematics Test was administered in the 32nd quarter-month of the 2009-2010 school year.

5) *TerraNova* scale scores corresponding to the national percentile rank (from column 2) were found in *TerraNova* norms for the quarter-months in which the NYS Mathematics Test was administered in the 2008–2009 school year. These scores are presented in column 3. The NYS Mathematics Test was administered in the 24th quarter-month of the 2008–2009 school year.
6) *TerraNova* standard deviations from the nationally representative norming samples (presented in column 4) were used to compute standardized growth (growth in standard deviation units) between the old and new administration windows in a following manner:

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

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

Table 40. Input Data for and Results of Computing NYS Student Growth in Mathematics

	Mean scale scores on *TerraNova* scale from new (2010) administration window ($TN\_Mean_{\_new}$)	National percentile rank (from *TerraNova* norms)	*TerraNova* mean scale scores from old (2009) administration window ($TN\_Mean_{\_old}$)	*TerranNova* standard deviation (TN_SD)	Standardized growth (SG)	NYS operational test standard deviation	Growth on NYS scale between old and new administration windows
Column	1	2	3	4	5	6	7
Grade 3	623	60	615	46	0.1739	32.78	6
Grade 4	650	63	643	47	0.1489	34.56	5
Grade 5	669	67	665	45	0.0889	32.44	3
Grade 6	680	60	676	46	0.0870	33.75	3
Grade 7	687	60	685	47	0.0426	31.70	1
Grade 8	698	58	696	48	0.0417	32.29	1

Final 2010 Mathematics Cut Scores

The resulting 2010 Mathematics OP proficiency level cut scores are presented in Table 41, columns 4 through 6. The 2009 OP cut scores (in columns 1–3) are also shown for comparison purposes. The 2010 OP test cut scores were applied to OP scores for tests administered in the 2009–2010 school year. These cut scores were determined following the procedures outlined in this section of the report.

A "Maximum RS – 1" rule was implemented for Level IV cut scores in cases when it was not possible to adjust this score. The "Maximum RS – 1" rule is used to determine Level IV cut scores if a perfect test score is required for a student to be classified in the proficiency Level IV category. In such situations, a scale score associated with a penultimate raw score (maximum raw score minus 1) is considered a performance Level IV cut score. Information on the cut score adjustment using the "Maximum RS– 1" rule was posted on the NYSED web site at http://www.p12.nysed.gov/irs/ela-math/2008/2008ELAScaleScoretoPerformanceLevels.html. For example, a Level IV cut score for Grade 3 in 2009 was 703. This cut score adjusted for the 2010 OP administration window should be 709 as indicated by the amount of growth on NYS scale between old and new administration windows from Table 40 (column 7). Because there was no scale score of 709 in the 2010 Grade 3 Raw Score-to-Scale Score conversion table and the next higher scale score was the highest obtainable score (770) associated with a perfect raw score, the 2010 Level IV cut score for Grade 3 was set at a penultimate scale score of 707 associates with a penultimate raw score.

Table 41. NYS 2009 and 2010 Mathematics Proficiency Level Cut Scores

	2009 operational test cut scores			2010 operational test cut scores		
	Proficiency Level			Proficiency Level		
	II	III	IV	II	III	IV
Column	1	2	3	4	5	6
Grade 3	624	650	703	**661**	**684**	**707***
Grade 4	622	650	702	**636**	**676**	**707**
Grade 5	619	650	699	**640**	**674**	**702**
Grade 6	616	650	696	**640**	**674**	**699**
Grade 7	611	650	693	**639**	**670**	**694**
Grade 8	616	650	701	**639**	**673**	**702**

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

Section VIII: Reliability and Standard Error of Measurement

This section presents specific information on various test reliability statistics (RSs) 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 Collection 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 42 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 42. 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	198549	31	39	0.88	2.20	0.89	2.05
4	201418	48	70	0.94	3.53	0.95	3.27
5	199254	34	46	0.90	2.87	0.91	2.68
6	200415	35	49	0.90	3.02	0.91	2.85
7	202359	38	50	0.90	3.07	0.91	2.86
8	206346	45	69	0.94	3.80	0.95	3.51

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 the 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 coefficients 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 43 presents reliabilities for the MC subsets.

Table 43. 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	198549	25	0.85	1.49	0.85	1.47
4	201418	30	0.86	1.95	0.87	1.92
5	199254	26	0.86	1.84	0.87	1.82
6	200415	25	0.85	1.86	0.85	1.84
7	202359	30	0.86	2.11	0.86	2.09
8	206346	27	0.89	1.86	0.89	1.85

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 44.

Table 44. 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	198549	6	14	0.75	1.47	0.77	1.41
4	201418	18	40	0.91	2.70	0.92	2.62
5	199254	8	20	0.79	2.04	0.80	1.97
6	200415	10	24	0.83	2.21	0.84	2.17
7	202359	8	20	0.81	2.02	0.82	1.95
8	206346	18	41	0.92	3.02	0.93	2.97

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 a 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 45a–45f, 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 0.80. Feldt-Raju reliability coefficients, which tend to be larger than the Cronbach's alpha estimates for the same group, were all larger than 0.82. Overall, the New York State Mathematics Tests were found to have very good test internal consistency (reliability) for analyzed subgroups of examinees.

Table 45a. 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	198549	0.88	2.20	0.89	2.05
Gender	Female	96870	0.87	2.20	0.89	2.05
	Male	101679	0.88	2.21	0.90	2.05
Ethnicity	Asian	15837	0.87	1.82	0.89	1.69
	Black	37343	0.88	2.45	0.90	2.30
	Hispanic	44650	0.88	2.37	0.89	2.21
	American Indian	959	0.87	2.37	0.88	2.21
	Multi-Racial	1114	0.88	2.19	0.89	2.05
	Unknown	123	0.84	1.88	0.86	1.76
	White	98523	0.85	2.05	0.87	1.92
NRC	New York City	71212	0.89	2.28	0.90	2.11
	Big 4 Cites	8491	0.89	2.58	0.90	2.44
	High Needs Urban/Suburban	15548	0.88	2.34	0.89	2.19
	High Needs Rural	11570	0.86	2.29	0.87	2.15
	Average Needs	58797	0.85	2.10	0.87	1.98
	Low Needs	28278	0.82	1.84	0.84	1.74
	Charter	4117	0.82	2.11	0.84	2.00
SWD	All Codes	28296	0.90	2.64	0.91	2.48
SUA	All Codes	49195	0.90	2.55	0.91	2.38
SWD/ SUA	SUA=504 Plan Codes	24683	0.90	2.66	0.91	2.51
ELL/ SUA	SUA=ELL Codes	18297	0.89	2.53	0.90	2.37
ELL	English	16784	0.89	2.51	0.90	2.35
	Chinese	596	0.83	2.05	0.85	1.93
	Haitian Creole	90	0.89	2.80	0.91	2.64
	Korean	63	0.91	1.88	0.93	1.65
	Russian	79	0.92	2.53	0.93	2.29
	Spanish	3525	0.90	2.65	0.91	2.48
	All Translations	4353	0.90	2.60	0.92	2.41

Table 45b. 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	201418	0.94	3.53	0.95	3.27
Gender	Female	98271	0.93	3.55	0.94	3.29
	Male	103147	0.94	3.51	0.95	3.24
Ethnicity	Asian	17023	0.93	2.97	0.94	2.75
	Black	37879	0.93	3.82	0.94	3.55
	Hispanic	43650	0.93	3.74	0.94	3.47
	American Indian	927	0.93	3.68	0.94	3.43
	Multi-Racial	1002	0.93	3.53	0.94	3.28
	Unknown	114	0.92	3.28	0.93	3.05
	White	100823	0.93	3.34	0.94	3.12
NRC	New York City	71973	0.94	3.60	0.95	3.31
	Big 4 Cites	8276	0.94	3.92	0.94	3.64
	High Needs Urban/Suburban	15385	0.93	3.73	0.94	3.47
	High Needs Rural	11569	0.93	3.66	0.94	3.43
	Average Needs	60389	0.92	3.43	0.93	3.21
	Low Needs	29816	0.91	3.06	0.92	2.89
	Charter	3455	0.91	3.58	0.92	3.39
SWD	All Codes	29723	0.94	3.95	0.95	3.66
SUA	All Codes	50224	0.94	3.90	0.95	3.61
SWD/ SUA	SUA=504 Plan Codes	26813	0.93	3.96	0.94	3.68
ELL/ SUA	SUA=ELL Codes	16381	0.93	3.90	0.94	3.63
ELL	English	14782	0.93	3.90	0.94	3.64
	Chinese	640	0.91	3.22	0.92	3.04
	Haitian Creole	118	0.94	4.03	0.95	3.71
	Korean	79	0.82	2.81	0.84	2.67
	Russian	70	0.95	3.84	0.95	3.55
	Spanish	3319	0.93	3.96	0.94	3.69
	All Translations	4226	0.94	3.92	0.95	3.61

Table 45c. 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	199254	0.90	2.87	0.91	2.68
Gender	Female	97021	0.89	2.87	0.91	2.69
	Male	102233	0.90	2.87	0.91	2.67
Ethnicity	Asian	15798	0.89	2.46	0.90	2.29
	Black	37962	0.89	3.08	0.90	2.91
	Hispanic	42946	0.89	3.02	0.91	2.83
	American Indian	919	0.89	3.02	0.90	2.87
	Multi-Racial	870	0.89	2.89	0.91	2.70
	Unknown	98	0.86	2.78	0.88	2.63
	White	100661	0.88	2.74	0.89	2.58
NRC	New York City	69240	0.90	2.91	0.92	2.71
	Big 4 Cites	7999	0.89	3.18	0.91	3.00
	High Needs Urban/Suburban	14913	0.89	3.02	0.90	2.84
	High Needs Rural	11620	0.88	2.99	0.89	2.83
	Average Needs	60495	0.88	2.80	0.89	2.65
	Low Needs	29825	0.85	2.54	0.87	2.41
	Charter	4585	0.86	2.94	0.88	2.80
SWD	All Codes	30360	0.89	3.18	0.91	3.00
SUA	All Codes	48591	0.90	3.15	0.91	2.95
SWD/ SUA	SUA=504 Plan Codes	27760	0.89	3.18	0.90	3.00
ELL/ SUA	SUA=ELL Codes	13278	0.90	3.14	0.91	2.95
ELL	English	11770	0.89	3.14	0.91	2.96
	Chinese	558	0.87	2.69	0.89	2.51
	Haitian Creole	115	0.90	3.14	0.91	2.95
	Korean	64	0.82	1.92	0.84	1.77
	Russian	79	0.91	3.03	0.92	2.83
	Spanish	3214	0.89	3.18	0.91	2.98
	All Translations	4030	0.91	3.14	0.92	2.91

Table 45d. 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	200415	0.90	3.02	0.91	2.85
Gender	Female	98143	0.90	3.01	0.91	2.85
	Male	102272	0.91	3.02	0.92	2.85
Ethnicity	Asian	15732	0.89	2.62	0.91	2.46
	Black	38306	0.89	3.29	0.90	3.12
	Hispanic	42544	0.89	3.24	0.90	3.06
	American Indian	968	0.90	3.12	0.91	2.95
	Multi-Racial	775	0.90	3.05	0.91	2.88
	Unknown	103	0.88	2.77	0.89	2.62
	White	101987	0.88	2.83	0.89	2.69
NRC	New York City	69397	0.91	3.17	0.92	2.96
	Big 4 Cites	7661	0.89	3.34	0.90	3.17
	High Needs Urban/Suburban	14676	0.89	3.17	0.90	3.01
	High Needs Rural	11628	0.88	3.08	0.89	2.94
	Average Needs	62056	0.88	2.89	0.89	2.76
	Low Needs	30473	0.87	2.61	0.88	2.49
	Charter	3859	0.87	3.02	0.88	2.89
SWD	All Codes	30788	0.89	3.43	0.90	3.23
SUA	All Codes	44734	0.89	3.40	0.91	3.20
SWD/ SUA	SUA=504 Plan Codes	27694	0.89	3.43	0.90	3.24
ELL/ SUA	SUA=ELL Codes	10581	0.89	3.45	0.90	3.25
ELL	English	9848	0.88	3.45	0.90	3.26
	Chinese	729	0.87	3.00	0.88	2.84
	Haitian Creole	175	0.88	3.51	0.89	3.29
	Korean	66	0.80	2.56	0.82	2.41
	Russian	70	0.92	3.36	0.93	3.05
	Spanish	2962	0.88	3.46	0.90	3.27
	All Translations	4002	0.90	3.42	0.91	3.21

Table 45e. 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	202359	0.90	3.07	0.91	2.86
Gender	Female	98671	0.90	3.06	0.91	2.86
	Male	103688	0.90	3.07	0.92	2.86
Ethnicity	Asian	16147	0.90	2.82	0.92	2.59
	Black	38559	0.89	3.17	0.90	3.02
	Hispanic	42126	0.89	3.16	0.90	3.00
	American Indian	963	0.89	3.10	0.90	2.94
	Multi-Racial	736	0.89	3.10	0.91	2.90
	Unknown	78	0.88	2.95	0.90	2.74
	White	103750	0.88	2.95	0.90	2.76
NRC	New York City	70122	0.91	3.13	0.92	2.92
	Big 4 Cites	7760	0.88	3.21	0.89	3.07
	High Needs Urban/Suburban	14573	0.89	3.14	0.90	2.99
	High Needs Rural	11870	0.88	3.06	0.89	2.90
	Average Needs	61776	0.88	2.97	0.89	2.80
	Low Needs	32384	0.87	2.82	0.89	2.65
	Charter	2954	0.88	3.07	0.89	2.92
SWD	All Codes	30432	0.88	3.22	0.89	3.09
SUA	All Codes	43323	0.89	3.22	0.90	3.08
SWD/ SUA	SUA=504 Plan Codes	27210	0.87	3.22	0.88	3.09
ELL/ SUA	SUA=ELL Codes	10403	0.89	3.25	0.90	3.10
ELL	English	8850	0.88	3.25	0.89	3.12
	Chinese	893	0.88	2.94	0.89	2.77
	Haitian Creole	181	0.87	3.19	0.88	3.05
	Korean	63	0.90	2.91	0.91	2.69
	Russian	79	0.88	3.23	0.89	3.06
	Spanish	3197	0.88	3.26	0.89	3.12
	All Translations	4413	0.90	3.25	0.91	3.07

Table 45f. 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	206346	0.94	3.80	0.95	3.51
Gender	Female	100529	0.94	3.76	0.95	3.48
	Male	105817	0.94	3.83	0.95	3.52
Ethnicity	Asian	16459	0.94	3.17	0.95	2.90
	Black	38687	0.94	4.05	0.95	3.77
	Hispanic	43053	0.94	4.01	0.95	3.72
	American Indian	927	0.94	3.97	0.95	3.70
	Multi-Racial	614	0.95	3.87	0.95	3.54
	Unknown	95	0.92	3.66	0.93	3.39
	White	106511	0.93	3.62	0.94	3.39
NRC	New York City	72544	0.95	3.90	0.96	3.55
	Big 4 Cites	7673	0.94	4.07	0.95	3.82
	High Needs Urban/Suburban	14516	0.93	4.01	0.94	3.76
	High Needs Rural	11979	0.93	3.91	0.93	3.68
	Average Needs	62954	0.93	3.69	0.94	3.46
	Low Needs	33081	0.92	3.34	0.93	3.15
	Charter	2392	0.93	3.90	0.93	3.66
SWD	All Codes	30662	0.93	4.06	0.94	3.81
SUA	All Codes	43648	0.94	4.07	0.95	3.79
SWD/ SUA	SUA=504 Plan Codes	27435	0.93	4.06	0.94	3.82
ELL/ SUA	SUA=ELL Codes	10594	0.94	4.05	0.95	3.75
ELL	English	8544	0.94	4.06	0.95	3.79
	Chinese	1034	0.94	3.41	0.95	3.12
	Haitian Creole	167	0.94	4.01	0.94	3.77
	Korean	73	0.92	3.26	0.94	2.98
	Russian	106	0.94	3.95	0.95	3.61
	Spanish	3236	0.94	4.08	0.94	3.80
	All Translations	4616	0.95	4.01	0.96	3.67

Standard Error of Measurement

SEMs, as computed from Cronbach's alpha and the Feldt-Raju reliability statistics, are presented in Table 42. SEMs based on Cronbach's alpha ranged from 2.20–3.80, 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 45a–45f. The SEMs associated with all reliability estimates for all subpopulations are in the range of 1.65–4.08, 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 & Lewis, 1995).

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

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

Classification consistency and accuracy were estimated using the IRT procedure suggested by Lee, Hanson & 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 46 and 47 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 46 depicts the consistency study results based on the range of performance levels for all grades. Overall, between 67% and 79% 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 0.54–0.69.

Table 46. Decision Consistency (All Cuts)

Grade	N-count	Agreement	Inconsistency	Kappa
3	198549	0.6690	0.3310	0.5398
4	201418	0.7824	0.2176	0.6867
5	199254	0.7226	0.2774	0.6018
6	200415	0.7192	0.2808	0.6065
7	202359	0.7338	0.2662	0.6274
8	206346	0.7857	0.2143	0.6945

Table 47 depicts the consistency study results based on two performance levels (passing and not passing) as defined by the Level III cut. Overall, about 86%–91% 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 0.70–0.82.

Table 47. Decision Consistency (Level III Cut)

Grade	N-count	Agreement	Inconsistency	Kappa
3	198549	0.8557	0.1443	0.7018
4	201418	0.9118	0.0882	0.8099
5	199254	0.8911	0.1089	0.7630
6	200415	0.8854	0.1146	0.7589
7	202359	0.8902	0.1098	0.7670
8	206346	0.9125	0.0875	0.8233

Accuracy

The results of classification accuracy are presented in Table 48. 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, including "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 73%–84% of the time across all performance levels and approximately 89%–94% of the time in regards to the Level III cut score.

Table 48. 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	198549	**0.7306**	0.2029	0.0664	**0.8944**	0.0639	0.0418
4	201418	**0.8440**	0.0872	0.0689	**0.9379**	0.0261	0.0360
5	199254	**0.7966**	0.1277	0.0756	**0.9224**	0.0424	0.0351
6	200415	**0.7902**	0.1385	0.0712	**0.9181**	0.0462	0.0357
7	202359	**0.8106**	0.1000	0.0894	**0.9222**	0.0409	0.0368
8	206346	**0.8445**	0.0933	0.0623	**0.9385**	0.0269	0.0346

Section IX: Summary of Operational Test Results

This section summarizes the distribution of OP scale score results on the New York State 2010 Grades 3–8 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, ELLs, SWDs, SUAs, and test language variables (Test Language) were used to calculate the results of subgroups required for federal reporting and test equity purposes. Especially, the ELL/SUA subgroup is defined as examinees whose ELL status is true and use one or more ELL-related accommodation. The SWD/SUA subgroup includes examinees who are classified with disabilities and use one or more disability-related accommodations. Data include examinees with valid scores from all public and charter schools. Note that complete scale score frequency distribution tables are located in Appendix I.

Scale Score Distribution Summary

Scale score distribution summaries are presented and discussed in Table 49. 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; and ELLs, taking the mathematics test in English, SWDs, and/or SUAs 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 I.

Table 49. Mathematics Scale Score Distribution Summary Grades 3–8

Grade	N-count	SS Mean	SS Std Dev	10th %tile	25th %tile	50th %tile	75th %tile	90th %tile
3	198549	692.72	32.85	662	674	687	697	770
4	201418	686.99	34.69	646	665	685	707	724
5	199254	684.79	32.48	648	667	686	701	725
6	200415	680.25	33.85	642	662	682	700	719
7	202359	676.91	31.78	642	659	677	697	714
8	206346	677.18	32.37	641	658	677	694	716

Grade 3

Scale score statistics and N-counts of demographic groups for Grade 3 are presented in Table 50. The population scale score mean was 692.72 with a standard deviation of 32.85. The gender subgroups performed the same, with a mean difference of 0.05 scale score points. Asian 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 and the Charter schools. The lowest performing NRC subgroup was the Big 4 Cities, with a mean of 675.95, and the lowest performing ethnic subgroup was Black (mean scale score of 681.78). 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 661 (Haitian Creole subgroup) to 697 (Korean subgroup), a difference that exceeds a standard deviation. The subgroup that used the Haitian Creole translation had a scale score mean of 34 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 687: Asian (697), White (691), Low Needs (697), students who used the Chinese (691) translations, and students who used the Korean (697) translations.

Table 50. 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	198549	692.72	32.85	662	674	687	697	770
Gender	Female	96870	692.75	32.33	662	674	687	697	770
	Male	101679	692.70	33.35	662	674	687	697	770
Ethnicity	Asian	15837	708.45	37.28	672	684	697	707	770
	Black	37343	681.78	28.75	655	666	678	691	707
	Hispanic	44650	685.44	29.67	657	670	681	691	707
	American Indian	959	685.95	28.68	659	670	681	697	707
	Multi-Racial	1114	692.29	31.95	662	674	687	697	770
	Unknown	123	705.07	36.17	672	681	697	707	770
	White	98523	697.70	32.93	668	678	691	707	770
NRC	New York City	71212	690.41	33.39	659	672	684	697	770
	Big 4 Cites	8491	675.95	27.58	649	662	674	687	697
	High Needs Urban/Suburban	15548	686.84	30.49	659	670	681	697	707
	High Needs Rural	11570	688.52	29.27	662	672	684	697	707
	Average Needs	58797	694.94	31.45	666	676	687	707	770
	Low Needs	28278	704.56	34.15	674	684	697	707	770
	Charter	4117	693.83	29.49	668	676	687	697	770
SWD	All Codes	28296	671.95	27.46	645	659	672	684	697
SUA	All Codes	49195	677.26	29.23	649	662	676	687	707
SWD/SUA	SUA=504 Plan Codes	24683	670.31	26.23	643	657	670	684	691
ELL/SUA	SUA=ELL Codes	18297	678.26	27.87	651	664	676	687	707

(Continued on next page)

Table 50. 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	16784	678.67	27.93	651	664	676	687	707
	Chinese	596	696.24	30.45	670	678	691	707	770
	Haitian Creole	90	658.77	35.16	637	649	661	676	684
	Korean	63	707.95	38.64	674	687	697	707	770
	Russian	79	678.47	36.99	643	664	681	691	707
	Spanish	3525	672.38	27.58	645	659	672	684	697
	All Translations	4353	675.99	30.02	647	660	674	687	707

Grade 4

Scale score statistics and N-counts of demographic groups for Grade 4 are presented in Table 51. The population scale score mean was 686.99 with a standard deviation of 34.69. The gender subgroups performed very similarly, with a mean difference of less than two scale score points. Asian, Multi-Racial, 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, Hispanic, 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 665.29, 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 685: Asian (707), Multi-Racial (688), White (692), Average Needs (690), Low Needs (700), and students who used the Chinese (695) and Korean (707) translations.

Table 51. 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	201418	686.99	34.69	646	665	685	707	724
Gender	Female	98271	686.63	33.83	648	665	685	704	724
	Male	103147	687.32	35.48	646	667	685	707	724
Ethnicity	Asian	17023	708.16	39.07	665	685	707	724	751
	Black	37879	672.53	31.26	636	654	673	692	707
	Hispanic	43650	676.56	31.81	640	657	676	695	712
	American Indian	927	678.39	30.14	641	658	678	697	717
	Multi-Racial	1002	687.56	32.59	651	667	688	707	724
	Unknown	114	696.82	32.95	660	673	695	717	734
	White	100823	693.42	32.82	657	675	692	712	734
NRC	New York City	71973	684.39	36.79	641	661	682	704	724
	Big 4 Cites	8276	665.29	31.63	628	646	665	685	704
	High Needs Urban/Suburban	15385	677.47	30.99	641	658	676	695	712
	High Needs Rural	11569	679.81	30.18	646	663	680	697	717
	Average Needs	60389	689.85	31.46	654	671	690	707	724
	Low Needs	29816	702.59	32.63	667	683	700	717	734
	Charter	3455	684.45	27.31	653	667	683	700	717
SWD	All Codes	29723	658.99	32.53	620	640	660	680	697
SUA	All Codes	50224	665.53	33.37	625	646	667	685	704
SWD/SUA	SUA=504 Plan Codes	26813	657.31	31.86	617	640	658	678	695
ELL/SUA	SUA=ELL Codes	16381	666.06	31.70	628	648	667	685	704
ELL	English	14782	665.71	31.38	630	648	667	683	700
	Chinese	640	697.16	31.55	664	680	695	712	734
	Haitian Creole	118	646.59	33.35	603	625	652	671	685
	Korean	79	711.27	30.11	676	692	707	724	751
	Russian	70	670.31	39.46	627	655	669	697	715
	Spanish	3319	658.32	31.63	620	640	660	678	695
	All Translations	4226	665.06	35.37	623	645	665	685	707

Grade 5

Grade 5 demographic group N-counts and scale score statistics are presented in Table 52. The population scale score mean was 684.79 with a standard deviation of 32.48. The gender subgroups performed very similarly, with a mean difference of less than one scale score point. Asian, Multi-Racial, 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 19 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 661.31, nearly one-half of a standard deviation below the second lowest performing NRC subgroup (High Needs, Urban/Suburban: 676.11) and close to 40 scale score units below the Low Needs subgroup mean. SWD, SUA, and ELL without testing in an alternate language subgroups scored consistently below the Statewide percentile scale score rankings. The Haitian Creole translation subgroup, which had a scale score mean (645.82) of more than 38 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 716.54, about one standard deviation above the population mean. At the 50th percentile, the following groups exceeded the population scale score of 686: Asian (701), White (689), Low Needs (697), and students who used the Chinese (693) and Korean (714) translations.

Table 52. 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	199254	684.79	32.48	648	667	686	701	725
Gender	Female	97021	685.00	31.84	648	667	683	701	725
	Male	102233	684.59	33.07	648	667	686	701	725
Ethnicity	Asian	15798	703.16	34.49	667	683	701	725	744
	Black	37962	671.03	30.63	636	653	671	689	707
	Hispanic	42946	676.10	31.38	640	658	676	693	714
	American Indian	919	675.71	31.93	643	660	678	693	707
	Multi-Racial	870	685.64	32.44	645	667	683	701	725
	Unknown	98	688.29	28.83	660	673	686	701	725
	White	100661	690.87	30.22	658	673	689	707	725
NRC	New York City	69240	682.57	34.42	643	662	680	701	725
	Big 4 Cites	7999	661.31	32.43	625	645	662	680	697
	High Needs Urban/Suburban	14913	676.11	29.98	643	660	676	693	707
	High Needs Rural	11620	678.02	28.29	645	662	678	693	707
	Average Needs	60495	687.91	29.48	656	671	686	701	725
	Low Needs	29825	698.72	28.99	667	680	697	714	744
	Charter	4585	680.58	26.33	648	664	680	697	714
SWD	All Codes	30360	657.51	33.18	620	640	660	678	693
SUA	All Codes	48591	663.37	33.41	625	645	664	683	701
SWD/SUA	SUA=504 Plan Codes	27760	656.43	32.86	615	640	660	676	693
ELL/SUA	SUA=ELL Codes	13278	662.68	32.87	625	645	664	683	697
ELL	English	11770	662.73	32.12	625	645	664	680	697
	Chinese	558	692.83	29.78	656	676	693	707	725
	Haitian Creole	115	645.82	36.88	603	625	653	669	689
	Korean	64	716.52	28.13	686	697	714	725	744
	Russian	79	668.65	39.54	629	648	671	686	714
	Spanish	3214	655.43	32.91	615	640	658	676	693
	All Translations	4030	661.56	35.90	620	643	662	683	701

Grade 6

Grade 6 scale score statistics and N-counts of demographic groups are presented in Table 53. The population scale score mean was 680.251 with a standard deviation of 33.85. 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 660.65. New York City, High Needs Urban/Suburban, High Needs Rural, and Charter subgroups had similar scale score means (ranging from approximately 670–680). SWD, SUA, and ELL without testing in an alternate language subgroups scored consistently below the Statewide percentile scale score rankings. The Haitian Creole translation subgroup, which had a scale score mean (638.05) more than 42 units below the population mean, was the lowest performing group analyzed. Asian students (the highest performing subgroup with a mean of 700.84) exceeded the State mean by over 20 scale score points. At the 50th percentile, the following groups exceeded the population scale score of 682: Asian (700), White (688), Average Needs (685), Low Needs (695), and students who used the Korean (694) translations.

Table 53. 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	200415	680.25	33.85	642	662	682	700	719
Gender	Female	98143	681.48	32.91	644	662	682	700	719
	Male	102272	679.06	34.68	640	660	679	700	719
Ethnicity	Asian	15732	700.84	35.75	660	679	700	719	751
	Black	38306	665.47	33.04	632	649	667	685	700
	Hispanic	42544	669.35	32.56	635	653	671	688	705
	American Indian	968	673.53	32.76	637	658	675	692	705
	Multi-Racial	775	678.83	32.36	642	662	679	695	719
	Unknown	103	689.97	29.84	658	671	688	705	731
	White	101987	687.23	30.61	653	671	688	705	719
NRC	New York City	69397	675.39	36.61	635	655	674	695	719
	Big 4 Cites	7661	660.65	33.35	626	644	662	679	695
	High Needs Urban/Suburban	14676	670.45	30.98	637	655	671	688	705
	High Needs Rural	11628	676.12	28.69	644	660	676	692	705
	Average Needs	62056	684.24	29.87	651	667	685	700	719
	Low Needs	30473	695.84	29.95	662	679	695	711	731
	Charter	3859	679.87	27.66	649	664	679	695	711
SWD	All Codes	30788	648.31	36.53	609	635	653	669	685
SUA	All Codes	44734	653.78	36.40	614	637	658	674	692
SWD/SUA	SUA=504 Plan Codes	27694	647.31	36.24	609	632	651	669	685
ELL/SUA	SUA=ELL Codes	10581	652.48	35.62	614	637	655	671	688

(Continued on next page)

Table 53. 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	9848	650.58	35.01	614	635	653	671	685
	Chinese	729	686.17	30.28	653	669	682	705	719
	Haitian Creole	175	638.05	41.87	595	622	647	664	676
	Korean	66	700.89	28.19	669	682	694	719	731
	Russian	70	651.10	49.44	589	642	659	679	695
	Spanish	2962	648.84	35.16	609	635	653	669	685
	All Translations	4002	656.07	38.27	614	640	658	676	695

Grade 7

N-counts and scale score statistics of demographic groups for Grade 7 are presented in Table 54. The population scale score mean was 676.91 with a standard deviation of 31.78. The gender subgroups performed very similarly, with a mean difference of less than three 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 652.47, while the Low Needs subgroup's scale score mean was 691.72. 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 677: Asian (697), White (685), Average Needs (683), Low Needs (691), and students who used the Chinese (683) and Korean (688) translations.

Table 54. 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	202359	676.91	31.78	642	659	677	697	714
Gender	Female	98671	678.11	31.15	642	661	677	697	714
	Male	103688	675.77	32.33	639	659	677	694	714
Ethnicity	Asian	16147	696.02	34.88	656	677	697	714	736
	Black	38559	661.29	29.88	630	647	663	680	694
	Hispanic	42126	665.48	29.91	633	649	668	683	701
	American Indian	963	667.54	28.99	636	652	668	685	701
	Multi-Racial	736	676.03	31.46	644	659	675	694	709
	Unknown	78	686.46	30.46	654	670	683	701	736
	White	103750	684.47	28.49	654	668	685	701	714

(Continued on next page)

Table 54. 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	70122	671.30	33.63	636	652	670	691	709
	Big 4 Cites	7760	652.47	32.51	617	639	654	672	688
	High Needs Urban/Suburban	14573	666.74	29.15	636	652	668	685	701
	High Needs Rural	11870	672.97	25.94	644	659	672	688	705
	Average Needs	61776	682.64	27.78	652	668	683	697	714
	Low Needs	32384	691.72	27.65	661	675	691	709	726
	Charter	2954	674.79	26.33	644	659	675	691	709
SWD	All Codes	30432	647.33	33.35	611	633	652	668	683
SUA	All Codes	43323	651.96	33.53	617	636	654	672	688
SWD/SUA	SUA=504 Plan Codes	27210	646.55	33.10	611	633	652	665	680
ELL/SUA	SUA=ELL Codes	10403	650.50	33.98	617	636	654	670	685
ELL	English	8850	647.14	33.77	611	633	649	668	683
	Chinese	893	683.91	27.70	654	670	683	701	714
	Haitian Creole	181	632.21	41.22	579	622	642	656	670
	Korean	63	689.08	29.90	654	668	688	709	726
	Russian	79	659.10	27.18	633	639	661	677	694
	Spanish	3197	648.05	32.65	611	633	652	668	683
	All Translations	4413	655.44	35.56	617	639	659	677	694

Grade 8

Grade 8 scale score statistics and N-counts of demographic groups are presented in Table 55. The population scale score mean was 677.18 with a standard deviation of 32.37. The gender subgroups performed similarly, with a mean difference of less than 5 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 9 scale score points below the population mean. The lowest performing NRC subgroup, Big 4 Cities, had a scale score mean of 651.77, while the Low Needs subgroup's scale score mean was 692.23, which indicated a large performance discrepancy by school district NRC designation. SWD, SUA, and ELL without testing in an alternate language subgroups scored consistently below the Statewide percentile scale score rankings. At the 50th percentile, the following groups exceeded the population scale score of 677: Female (678), Asian (697), White (683), Average Needs (681), Low Needs (691), and students who used the Chinese (689) and Korean (691) translations.

Table 55. 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	206346	677.18	32.37	641	658	677	694	716
Gender	Female	100529	679.40	31.89	644	660	678	697	716
	Male	105817	675.07	32.68	638	657	675	694	716
Ethnicity	Asian	16459	700.14	36.31	659	678	697	725	741
	Black	38687	661.64	29.80	629	645	661	678	697
	Hispanic	43053	666.06	29.97	633	649	666	683	700
	American Indian	927	667.92	29.65	634	653	667	685	700
	Multi-Racial	614	673.94	34.53	636	655	674	694	709
	Unknown	95	690.02	32.50	657	666	685	704	741
	White	106511	683.86	29.33	652	667	683	700	716
NRC	New York City	72544	672.98	34.86	634	650	670	691	716
	Big 4 Cites	7673	651.77	31.30	619	634	653	670	685
	High Needs Urban/Suburban	14516	667.23	27.16	636	652	667	683	700
	High Needs Rural	11979	672.52	26.70	644	658	672	687	704
	Average Needs	62954	681.55	28.49	650	665	681	697	716
	Low Needs	33081	692.23	28.39	661	675	691	709	725
	Charter	2392	674.22	25.65	644	658	674	689	704
SWD	All Codes	30662	647.73	30.53	615	633	650	666	680
SUA	All Codes	43648	653.61	31.61	619	636	655	672	689
SWD/SUA	SUA=504 Plan Codes	27435	647.28	30.13	615	633	650	665	680
ELL/SUA	SUA =ELL Codes	10594	656.24	32.16	621	638	657	674	694
ELL	English	8544	652.39	30.66	619	636	653	670	687
	Chinese	1034	691.34	32.19	655	672	689	709	725
	Haitian Creole	167	645.61	26.69	612	626	647	664	681
	Korean	73	694.77	27.71	661	681	691	709	725
	Russian	106	668.94	32.13	633	649	667	685	704
	Spanish	3236	650.39	30.10	615	633	653	669	683
	All Translations	4616	660.52	35.23	621	639	660	681	700

Performance Level Distribution Summary

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

Table 56 shows the mathematics cut scores used for classification of students to the four performance levels in 2010.

Table 56. Mathematics Grades 3–8 Performance Level Cut Scores

Grade	Level II Cut	Level III Cut	Level IV Cut
3	661	684	707
4	636	676	707
5	640	674	702
6	640	674	699
7	639	670	694
8	639	673	702

Tables 57–63 show the performance level distributions for all examinees from public and charter schools with valid scores. Table 57 presents performance level data for total populations of students in Grades 3–8. Tables 58–63 contain performance level data for selected subgroups of students. In general, these summaries reflect the same achievement trends as in the scale score summary discussion. Male and Female students performed similarly across grades. 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 57. 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	198549	9.30	31.50	35.16	24.05	59.20
4	201418	5.26	30.84	38.14	25.75	63.90
5	199254	5.99	29.25	40.85	23.91	64.76
6	200415	7.96	30.58	34.27	27.19	61.46
7	202359	8.10	29.40	33.32	29.18	62.49
8	206346	9.19	35.94	36.60	18.27	54.87

Grade 3

Performance level summaries and N-counts of demographic groups for Grade 3 are presented in Table 58. Statewide, 59.20% 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 77% of students from Low Needs districts were classified in Levels III and IV, whereas only about 33% of Big 4 Cities students were in Levels III and IV. Less than 40% 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 67% in Levels III and IV, with Korean students having the greatest percentage of more than 79%.

Table 58. 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	198549	9.30	31.50	35.16	24.05	59.20
Gender	Female	96870	8.77	32.19	35.32	23.72	59.04
	Male	101679	9.80	30.85	35.00	24.36	59.35
Ethnicity	Asian	15837	4.09	17.00	36.19	42.72	78.90
	Black	37343	16.93	41.27	28.54	13.26	41.80
	Hispanic	44650	13.40	38.38	32.10	16.12	48.22
	American Indian	959	12.30	40.35	31.49	15.85	47.34
	Multi-Racial	1114	9.61	29.53	36.27	24.60	60.86
	Unknown	123	2.44	26.02	32.52	39.02	71.54
	White	98523	5.36	26.95	38.91	28.78	67.69
NRC	New York City	71212	11.86	33.76	32.26	22.13	54.38
	Big 4 Cites	8491	24.85	42.40	23.48	9.27	32.75
	High Needs Urban/Suburban	15548	12.21	38.02	32.37	17.40	49.77
	High Needs Rural	11570	9.41	37.76	34.56	18.26	52.83
	Average Needs	58797	6.07	29.78	38.68	25.47	64.15
	Low Needs	28278	2.91	19.75	40.44	36.90	77.34
	Charter	4117	4.69	33.71	38.35	23.25	61.60
SWD	All Codes	28296	30.17	42.10	20.78	6.94	27.72
SUA	All Codes	49195	23.42	41.09	24.89	10.61	35.49
SWD/SUA	SUA=504 Plan Codes	24683	31.95	42.94	19.39	5.72	25.11
ELL/SUA	SUA=ELL Codes	18297	20.39	42.87	26.46	10.29	36.75
ELL	ELL status = Y	20494	21.41	42.43	26.09	10.07	36.16
ELL Test Language	English	16784	19.96	42.50	27.00	10.54	37.54
	Chinese	596	4.87	27.52	41.44	26.17	67.62
	Haitian Creole	90	50.00	36.67	12.22	1.11	13.33
	Korean	63	6.35	14.29	36.51	42.86	79.37
	Russian	79	21.52	35.44	29.11	13.92	43.04
	Spanish	3525	29.11	42.55	20.60	7.74	28.34
	All Translations	4353	25.75	39.83	23.66	10.75	34.41

Grade 4

Performance level summaries and N-counts of demographic groups for Grade 4 are presented in Table 59. Statewide, 63.90% of the fourth-grade population was placed in Levels III and IV. Around 6%–10% of American Indian, Black, and Hispanic students were Level I, as compared to only about 2.39% of Asian students and 2.90% of White students. American Indian, Black, and Hispanic ethnic subgroups had percentages of students in Levels III and IV ranging from 44%–54%, but the percentages of the Multi-Racial, White, and Asian subgroup students meeting standards for Levels III and IV (64.67%, 73.25%, and 83.36%, respectively) exceeded the population. Student achievement also varied widely by NRC subgroup. About 83% of students from Low Needs districts were meeting standards for Levels III and IV, but only about 37% of Big 4 Cities students were. Less than 40% of SWD or SUA status students or those who took translated test forms met or exceeded the Level III cut score; however, the Chinese translation subgroup had a very high percentage of students in Levels III and IV (80.31%). 91.14% 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: Male, Asian, Multi-Racial, White, Average Needs, Low Needs, Chinese translation, and Korean translation.

Table 59. 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	201418	5.26	30.84	38.14	25.75	63.90
Gender	Female	98271	4.87	31.68	38.55	24.90	63.45
	Male	103147	5.64	30.04	37.75	26.57	64.32
Ethnicity	Asian	17023	2.39	14.25	33.35	50.01	83.36
	Black	37879	9.75	44.82	33.11	12.32	45.43
	Hispanic	43650	7.97	41.12	35.78	15.14	50.92
	American Indian	927	6.36	39.81	38.40	15.43	53.83
	Multi-Racial	1002	4.19	31.14	38.82	25.85	64.67
	Unknown	114	0.88	26.32	34.21	38.60	72.81
	White	100823	2.90	23.85	41.86	31.38	73.25
NRC	New York City	71973	6.82	34.63	34.24	24.30	58.55
	Big 4 Cites	8276	14.98	48.51	27.72	8.78	36.50
	High Needs Urban/Suburban	15385	6.92	40.64	37.35	15.09	52.44
	High Needs Rural	11569	5.41	38.25	39.87	16.48	56.34
	Average Needs	60389	3.20	27.39	42.65	26.76	69.41
	Low Needs	29816	1.45	15.40	40.83	42.33	83.16
	Charter	3455	2.78	33.46	45.24	18.52	63.76
SWD	All Codes	29723	20.26	50.50	23.35	5.89	29.24
SUA	All Codes	50224	15.41	47.40	27.97	9.22	37.19
SWD/SUA	SUA=504 Plan Codes	26813	21.22	51.57	22.46	4.75	27.21
ELL/SUA	SUA=ELL Codes	16381	13.71	49.26	28.64	8.39	37.02

(Continued on next page)

Table 59. 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	18455	14.72	49.42	27.72	8.14	35.85
ELL Test Language	English	14782	13.69	49.74	28.68	7.88	36.56
	Chinese	640	2.50	17.19	45.63	34.69	80.31
	Haitian Creole	118	34.75	45.76	17.80	1.69	19.49
	Korean	79	0.00	8.86	37.97	53.16	91.14
	Russian	70	14.29	42.86	25.71	17.14	42.86
	Spanish	3319	20.22	52.18	22.39	5.21	27.60
	All Translations	4226	17.46	45.74	26.12	10.67	36.80

Grade 5

Performance level summaries and N-counts of demographic groups for Grade 5 are presented in Table 60. Statewide, 64.76% of the fifth-grade population was placed in Levels III and IV. There was little performance differentiation by gender subgroup, with less than 1% 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 45%–55%), as compared to the percentage of Asian and White students meeting standards for Levels III and IV (84% and 74% respectively). Over 83% of students from Low Needs districts were in Levels III or IV, but only about 33% of the Big 4 Cities students were. Only about 5%–8% of SWD or SUA subgroups were placed in Level IV, compared to the population's 23.91% in Level IV. 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 Russian (12.66%) and Chinese and Korean translation subgroups that had very high percentages of students in Level IV (32.44% and 64.06%, respectively). 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 60. 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	199254	5.99	29.25	40.85	23.91	64.76
Gender	Female	97021	5.49	29.73	40.95	23.82	64.77
	Male	102233	6.46	28.79	40.76	23.99	64.75
Ethnicity	Asian	15798	2.62	13.39	37.08	46.90	83.99
	Black	37962	11.08	43.22	34.66	11.04	45.70
	Hispanic	42946	8.93	38.01	37.82	15.24	53.06
	American Indian	919	7.40	37.65	41.68	13.28	54.95

(Continued on next page)

Table 60. 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	870	5.40	30.46	39.66	24.48	64.14
	Unknown	98	4.08	25.51	46.94	23.47	70.41
	White	100661	3.34	22.64	45.07	28.94	74.01
NRC	New York City	69240	7.46	32.67	36.61	23.26	59.87
	Big 4 Cites	7999	19.53	47.22	26.40	6.85	33.25
	High Needs Urban/Suburban	14913	8.05	38.38	39.51	14.05	53.56
	High Needs Rural	11620	6.14	36.83	42.38	14.65	57.02
	Average Needs	60495	3.74	25.85	45.50	24.91	70.41
	Low Needs	29825	1.64	14.77	45.33	38.26	83.59
	Charter	4585	4.25	36.01	44.45	15.29	59.74
SWD	All Codes	30360	22.42	48.45	24.07	5.05	29.12
SUA	All Codes	48591	17.78	45.70	28.48	8.03	36.52
SWD/SUA	SUA=504 Plan Codes	27760	23.23	49.06	23.19	4.52	27.71
ELL/SUA	SUA=ELL Codes	13278	17.44	47.67	27.51	7.38	34.89
ELL	ELL status = Y	15153	18.38	47.69	26.68	7.25	33.93
ELL Test Language	English	11770	17.12	48.23	27.61	7.03	34.65
	Chinese	558	2.33	21.68	43.55	32.44	75.99
	Haitian Creole	115	37.39	44.35	15.65	2.61	18.26
	Korean	64	0.00	6.25	29.69	64.06	93.75
	Russian	79	12.66	46.84	27.85	12.66	40.51
	Spanish	3214	24.11	48.41	23.12	4.36	27.47
	All Translations	4030	20.87	43.90	25.93	9.31	35.24

Grade 6

Performance level summaries and N-counts of demographic groups for Grade 6 are presented in Table 61. Statewide, 61.46% of the sixth-grade population was placed in Levels III and IV. There was a slight performance differentiation by gender subgroup with less than 2% difference between each level. There were marked differences across ethnic and test translation subgroups. About 10%–15% of American Indian, Black, and Hispanic students were in Level I, as compared to less than 5% 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 41%–54%), as compared to the percentage of Asian and White students meeting standards for Levels III and IV (82.13% and 72.05%, respectively). About 82% of students from Low Needs districts were in Levels III or IV, but only about 35% of the Big 4 Cities students were. Only about 4%–7% of SWD and SUA subgroups were placed in Level IV, compared to the population's 27.19% in Level IV. 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 very high percentages of students in Level IV (31.28% and 43.94%, respectively). The following subgroups had a higher percentage of students meeting standards for Levels III and IV than the State population: Female, Asian, Multi-Racial, White, Average Needs, Low Needs, Chinese translation, and Korean translation.

Table 61. 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	200415	7.96	30.58	34.27	27.19	61.46
Gender	Female	98143	6.97	30.52	34.60	27.92	62.52
	Male	102272	8.91	30.64	33.96	26.49	60.45
Ethnicity	Asian	15732	3.12	14.75	29.55	52.57	82.13
	Black	38306	15.03	43.55	28.99	12.44	41.42
	Hispanic	42544	12.48	40.91	31.08	15.54	46.62
	American Indian	968	10.43	36.16	34.61	18.80	53.41
	Multi-Racial	775	8.39	29.81	37.29	24.52	61.81
	Unknown	103	3.88	21.36	38.84	35.92	74.76
	White	101987	4.14	23.81	38.29	33.76	72.05
NRC	New York City	69397	11.41	35.41	29.77	23.41	53.18
	Big 4 Cites	7661	18.25	46.33	26.39	9.03	35.43
	High Needs Urban/Suburban	14676	10.81	40.34	33.41	15.43	48.84
	High Needs Rural	11628	6.51	36.38	38.15	18.96	57.11
	Average Needs	62056	4.71	27.01	38.87	29.42	68.28
	Low Needs	30473	2.38	15.69	36.35	45.58	81.93
	Charter	3859	5.57	33.09	37.29	24.05	61.34
SWD	All Codes	30788	30.46	47.79	17.55	4.20	21.75
SUA	All Codes	44734	25.45	46.75	21.07	6.74	27.80
SWD/SUA	SUA=504 Plan Codes	27694	31.38	48.15	16.92	3.54	20.47
ELL/SUA	SUA=ELL Codes	10581	26.67	48.48	18.47	6.38	24.85
ELL	ELL status = Y	13014	28.15	48.57	17.5	5.78	23.27
ELL Test Language	English	9848	28.04	49.71	17.27	4.99	22.26
	Chinese	729	3.98	25.65	39.09	31.28	70.37
	Haitian Creole	175	39.43	46.86	12.00	1.71	13.71
	Korean	66	0.00	15.15	40.91	43.94	84.85
	Russian	70	21.43	47.14	22.86	8.57	31.43
	Spanish	2962	29.64	49.29	17.12	3.95	21.07
	All Translations	4002	24.76	44.28	21.39	9.57	30.96

Grade 7

Performance level summaries and N-counts of demographic groups for Grade 7 are presented in Table 62. Statewide, 62.49% 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 39%–50% of students meeting standards for Levels III and IV, with less than 18% of those students in Level IV, whereas over 82% of Asian students were meeting standards for Levels III and IV (and over 54% were in Level IV.) About 29% of Big 4 Cities students were meeting standards for Levels III and IV, with less than 8% in Level IV, yet over 83% of students from Low Needs districts were meeting standards for Levels III and IV (with about 48% in Level IV). Less than 8% of SWD and SUA subgroups were placed in Level IV, and about 30% were in Level I. Less than 12% 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 (35.50% and 41.27%, respectively). Across all subgroups, the Haitian Creole translation subgroup had the largest percentage of students placed in Level I (45.86%) and the Korean translation subgroup had the largest percentage of students (41.27%) 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, Average Needs, Low Needs, Chinese translation, and Korean translation.

Table 62. 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	202359	8.10	29.40	33.32	29.18	62.49
Gender	Female	98671	7.35	28.60	33.75	30.29	64.05
	Male	103688	8.81	30.17	32.90	28.11	61.02
Ethnicity	Asian	16147	3.57	14.36	27.22	54.85	82.07
	Black	38559	15.93	44.09	28.36	11.62	39.98
	Hispanic	42126	13.32	39.87	31.33	15.48	46.81
	American Indian	963	11.42	39.46	31.78	17.34	49.12
	Multi-Racial	736	6.39	33.42	32.74	27.45	60.19
	Unknown	78	3.85	19.23	43.59	33.33	76.92
	White	103750	3.77	21.92	36.93	37.38	74.31
NRC	New York City	70122	11.75	35.39	29.57	23.29	52.87
	Big 4 Cites	7760	23.93	47.32	21.55	7.20	28.75
	High Needs Urban/Suburban	14573	11.63	40.64	31.69	16.04	47.73
	High Needs Rural	11870	6.54	35.17	37.90	20.39	58.29
	Average Needs	61776	3.94	23.88	38.07	34.10	72.18
	Low Needs	32384	2.12	14.80	34.80	48.28	83.08
	Charter	2954	6.13	34.73	36.15	22.99	59.14
SWD	All Codes	30432	29.95	47.25	18.41	4.39	22.80
SUA	All Codes	43323	25.67	45.63	21.58	7.13	28.70
SWD/SUA 3	SUA=504 Plan Codes	27210	30.69	47.75	17.64	3.92	21.56
ELL/SUA 2	SUA=ELL Codes	10403	27.72	45.32	20.03	6.92	26.95
ELL	ELL status = Y	12557	29.43	45.56	18.8	6.21	25.01
ELL Test Language	English	8850	30.55	47.21	17.56	4.68	22.24
	Chinese	893	4.26	20.60	39.64	35.50	75.14
	Haitian Creole	181	45.86	43.09	9.39	1.66	11.05
	Korean	63	3.17	23.81	31.75	41.27	73.02
	Russian	79	24.05	40.51	25.32	10.13	35.44
	Spanish	3197	29.34	46.01	20.27	4.38	24.65
	All Translations	4413	24.47	40.34	24.00	11.19	35.19

Grade 8

Performance level summaries and N-counts of demographic groups for Grade 8 are presented in Table 63. Statewide, 54.87% of the eighth-grade population was placed in Levels III and IV. Overall, there was little performance differentiation by gender subgroup, with less than 4% difference between each level percentage. Across ethnic and test translation subgroups, there were marked differences in performance. Around 12%–18% of Black, Hispanic, and American Indian students were in Level I, compared to less than 5% of Asian and White students. American Indian, Black, Hispanic, and Multi-Racial ethnic subgroups had around 32%–40% of students meeting standards for Levels III and IV, respectively, whereas about 80% of Asian students were meeting Levels III and IV standards. About 21% of Big 4 Cities students were in Levels III and IV, yet over 77% of students from Low Needs districts were classified in these proficiency levels. Approximately 26%–32% 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 Russian, Chinese, and Korean translation subgroups that had a very high percentage of students in Level IV (11.32%, 29.98% and 38.36%, respectively). Across all subgroups, the Haitian Creole translation subgroup had the largest percentage of students placed in Level I (40.12%), and the Korean translation subgroup had the largest percentage of students placed in Level IV (38.36%). 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 63. 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	206346	9.19	35.94	36.60	18.27	54.87
Gender	Female	100529	7.76	34.62	37.80	19.83	57.63
	Male	105817	10.56	37.20	35.45	16.79	52.24
Ethnicity	Asian	16459	3.38	16.81	35.52	44.29	79.81
	Black	38687	18.17	49.63	25.36	6.84	32.20
	Hispanic	43053	14.76	46.72	29.27	9.25	38.52
	American Indian	927	12.51	47.36	30.74	9.39	40.13
	Multi-Racial	614	10.91	37.79	33.88	17.43	51.30
	Unknown	95	2.11	32.63	36.84	28.42	65.26
	White	106511	4.55	29.46	43.87	22.12	65.99
NRC	New York City	72544	13.05	40.52	29.30	17.14	46.43
	Big 4 Cites	7673	29.04	49.88	17.56	3.53	21.09
	High Needs Urban/Suburban	14516	11.39	49.47	31.07	8.07	39.14
	High Needs Rural	11979	7.34	44.11	38.31	10.24	48.55
	Average Needs	62954	4.81	32.41	43.72	19.06	62.78
	Low Needs	33081	2.22	20.26	46.32	31.21	77.52
	Charter	2392	7.07	42.52	39.13	11.29	50.42
SWD	All Codes	30662	32.48	50.92	14.78	1.82	16.60
SUA	All Codes	43648	26.93	49.41	19.46	4.20	23.66

(Continued on next page)

Table 63. 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	27435	32.97	51.19	14.20	1.63	15.84
ELL/SUA	SUA=ELL Codes	10594	25.54	47.93	20.57	5.96	26.52
ELL	ELL status = Y	12491	27.60	48.03	19.05	5.32	24.37
ELL Test Language	English	8544	28.07	50.42	17.59	3.92	21.51
	Chinese	1034	4.35	22.34	43.33	29.98	73.31
	Haitian Creole	167	40.12	44.91	13.77	1.20	14.97
	Korean	73	2.74	16.44	42.47	38.36	80.82
	Russian	106	13.21	50.00	25.47	11.32	36.79
	Spanish	3236	31.18	48.83	17.55	2.44	19.99
	All Translations	4616	24.63	42.27	23.77	9.34	33.10

Section X: Longitudinal Comparison of Results

This section provides a longitudinal comparison of OP scale score results on the New York State 2007–2010 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 64.

Table 64. 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	2010	198549	692.72	32.85	9.30	31.50	35.16	24.05	59.20
	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	2010	201418	686.99	34.69	5.26	30.84	38.14	25.75	63.90
	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	2010	199254	684.79	32.48	5.99	29.25	40.85	23.91	64.76
	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	2010	200415	680.25	33.85	7.96	30.58	34.27	27.19	61.46
	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	2010	202359	676.91	31.78	8.10	29.40	33.32	29.18	62.49
	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	2010	206346	677.18	32.37	9.19	35.94	36.60	18.27	54.87
	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

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

As seen in Table 64, an increase in scale score means was observed for all mathematics grades between 2006 and 2010. The least gain was observed for Grades 3 and 4, for which total gain was 15 and 10 scale score points, respectively, between 2006 and 2010 test administrations. The greatest gain in scale score points between 2006 and 2010 test administrations was noted for Grades 6, 7, and 8 (24, 26, and 26 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 32–34 scale score points in 2010. The scale score standard deviation for Grades 3 and 4 only decreased slightly between years 2006 and 2009 (less than 3 scale score points) and then decreased about 4 scale score points between years 2009 and 2010.

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 2010 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 the item pool. The pools of items eligible for 2010 item selection will include 2005, 2006, 2007, 2008, and 2009 FT items. All items for each grade will be on the same (grade-specific) scale.

Here are general guidelines for item selection:

- Satisfy the content specifications in terms of objective coverage and the number and percentage of MC and CR items on the test. An often used criterion for objective coverage is within 5% difference of the 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 2010 OP forms and the 2009 OP forms.
- From the ITEMWIN output, evaluate expected percentage of maximum raw score at each scale score and difference between reference set (2009) and working set (2010)—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, 2010 full selection, 2010 MC items only)
    - Item number in 2010 OP book
    - Item unique identification number, item type, FT year, FT form, and FT item number
    - Item classical statistics (p-values, point biserials, etc.)
    - ITEMWIN output (including TCCs)
    - Summary file with IRT item parameters for selected items

Appendix C—Factor Analysis Results

As described in Section III, "Validity," a principal component factor analysis was conducted on Grades 3–8 Mathematics Tests data. The analyses were conducted for the total population of students and selected subpopulations: English language learners (ELLs), students with disabilities (SWDs), students using accommodations (SUA), SWD students using disability accommodations (SWD/SUA), and ELL students using ELL related accommodations (ELL/SUA). Table C1 contains 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**	7.76	25.03	25.03
		2	1.49	4.82	29.85
	SWD	**1**	8.38	27.02	27.02
		2	1.43	4.63	31.65
	SUA	**1**	8.35	26.95	26.95
		2	1.45	4.68	31.62
		3	1.01	3.24	34.87
	SWD/SUA	**1**	8.18	26.40	26.40
		2	1.42	4.59	30.98
		3	1.00	3.23	34.21
	ELL/SUA	**1**	7.75	24.99	24.99
		2	1.51	4.87	29.86
4	ELL	**1**	12.18	25.38	25.38
		2	1.56	3.26	28.63
		3	1.20	2.50	31.13
		4	1.04	2.17	33.30
		5	1.01	2.11	35.42
	SWD	**1**	12.77	26.61	26.61
		2	1.59	3.32	29.93
		3	1.18	2.45	32.38
		4	1.05	2.19	34.57
		5	1.02	2.13	36.71
	SWA	**1**	13.05	27.20	27.20
		2	1.59	3.30	30.50
		3	1.17	2.43	32.93
		4	1.03	2.15	35.09

(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**	12.49	26.02	26.02
		2	1.59	3.31	29.33
		3	1.19	2.47	31.80
		4	1.06	2.21	34.01
		5	1.04	2.16	36.17
	ELL/SUA	**1**	12.29	25.61	25.61
		2	1.57	3.27	28.88
		3	1.20	2.50	31.37
		4	1.04	2.16	33.53
5	ELL	**1**	8.02	23.60	23.60
		2	1.31	3.86	27.46
		3	1.04	3.07	30.53
		4	1.01	2.98	33.51
	SWD	**1**	8.09	23.79	23.79
		2	1.36	4.01	27.80
		3	1.02	3.00	30.80
	SUA	**1**	8.43	24.79	24.79
		2	1.32	3.89	28.68
		3	1.01	2.97	31.65
	SWD/SUA	**1**	7.97	23.43	23.43
		2	1.37	4.03	27.46
		3	1.02	3.00	30.46
	ELL/SUA	**1**	8.19	24.09	24.09
		2	1.31	3.87	27.95
		3	1.03	3.03	30.98
		4	1.00	2.95	33.93
6	ELL	**1**	7.50	21.42	21.42
		2	1.39	3.97	25.39
		3	1.21	3.45	28.84
		4	1.03	2.95	31.79
	SWD	**1**	7.80	22.27	22.27
		2	1.37	3.92	26.20
		3	1.23	3.52	29.72
		4	1.01	2.90	32.62
	SUA	**1**	8.16	23.31	23.31
		2	1.39	3.96	27.27
		3	1.20	3.42	30.69
	SWD/SUA	**1**	7.63	21.81	21.81
		2	1.37	3.90	25.71
		3	1.24	3.55	29.26
		4	1.02	2.91	32.17

(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**	7.77	22.19	22.19
		2	1.42	4.05	26.23
		3	1.20	3.41	29.65
		4	1.02	2.90	32.55
7	ELL	**1**	7.17	18.86	18.86
		2	1.43	3.77	22.63
		3	1.24	3.26	25.89
		4	1.09	2.86	28.75
		5	1.06	2.80	31.55
		6	1.03	2.70	34.25
	SWD	**1**	7.26	19.10	19.10
		2	1.56	4.11	23.22
		3	1.24	3.27	26.49
		4	1.08	2.85	29.34
		5	1.03	2.71	32.05
		6	1.01	2.65	34.70
	SUA	**1**	7.78	20.47	20.47
		2	1.56	4.09	24.56
		3	1.23	3.24	27.80
		4	1.08	2.84	30.64
		5	1.01	2.67	33.30
	SWD/SUA	**1**	7.09	18.65	18.65
		2	1.56	4.11	22.76
		3	1.25	3.28	26.04
		4	1.09	2.87	28.91
		5	1.04	2.73	31.64
		6	1.02	2.68	34.31
	ELL/SUA	**1**	7.68	20.20	20.20
		2	1.41	3.71	23.92
		3	1.28	3.38	27.29
		4	1.08	2.83	30.12
		5	1.04	2.74	32.86
		6	1.01	2.66	35.52
8	ELL	**1**	12.91	28.68	28.68
		2	1.42	3.15	31.83
		3	1.19	2.63	34.46
		4	1.13	2.52	36.98
		5	1.01	2.25	39.23
	SWD	**1**	12.23	27.18	27.18
		2	1.41	3.14	30.32
		3	1.18	2.63	32.94
		4	1.07	2.39	35.33

(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	SWD	**5**	1.05	2.33	37.66
		6	1.00	2.23	39.88
8	SUA	**1**	13.28	29.51	29.51
		2	1.41	3.12	32.63
		3	1.15	2.56	35.19
		4	1.05	2.33	37.52
		5	1.01	2.24	39.76
	SWD/SUA	**1**	12.03	26.74	26.74
		2	1.41	3.14	29.89
		3	1.19	2.63	32.52
		4	1.08	2.40	34.91
		5	1.06	2.35	37.26
		6	1.00	2.23	39.49
	ELL/SUA	**1**	13.78	30.62	30.62
		2	1.41	3.13	33.75
		3	1.15	2.55	36.30
		4	1.10	2.45	38.74

Appendix D—Items Flagged for DIF

Tables D1 and D2 support the DIF information in Section V, "Operational Test Data Collection and Classical Analysis," and Section VI, "IRT Scaling and Equating." They include item numbers, focal groups, 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 in Table D1 positive values of SMD and Delta 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 2010 Classical DIF Item Flags

Grade	Item #	Subgroup	DIF	SMD	Mantel-Haenszel	Delta
3	4	Spanish	Against	-0.101	No Flag	No Flag
3	20	Spanish	Against	No Flag	225.135	-1.640
3	20	Asian	Against	No Flag	226.572	-1.506
3	25	Spanish	Against	-0.129	371.937	-1.687
3	25	ELL	Against	-0.109	1457.450	-1.611
3	25	Hispanic	Against	No Flag	2134.950	-1.79
3	28	Spanish	Against	-0.107	No Flag	No Flag
3	29	Black	In Favor	0.105	No Flag	No Flag
4	32	Spanish	Against	-0.108	No Flag	No Flag
4	39	Spanish	Against	-0.103	No Flag	No Flag
4	47	Female	In Favor	0.125	No Flag	No Flag
4	48	Spanish	Against	-0.138	No Flag	No Flag
5	11	Spanish	Against	-0.125	287.775	-1.558
5	11	Hispanic	Against	No Flag	1774.496	-1.685
5	11	ELL	Against	-0.115	1119.900	-1.560
5	15	Spanish	Against	-0.122	No Flag	No Flag
5	29	Spanish	Against	-0.129	No Flag	No Flag
5	29	Female	In Favor	0.107	No Flag	No Flag
5	31	Spanish	Against	-0.134	No Flag	No Flag
5	33	Spanish	Against	-0.120	No Flag	No Flag
5	34	Spanish	Against	-0.156	No Flag	No Flag
5	34	Female	Against	-0.160	No Flag	No Flag
5	34	Black	Against	-0.103	No Flag	No Flag
5	34	Hispanic	Against	-0.120	No Flag	No Flag
5	34	ELL	Against	-0.127	No Flag	No Flag
6	11	Spanish	In Favor	0.107	No Flag	No Flag
6	20	Female	Against	-0.116	4202.640	-1.762
6	29	Black	Against	-0.105	No Flag	No Flag
6	29	Female	In Favor	0.121	No Flag	No Flag
6	32	Spanish	Against	-0.195	No Flag	No Flag
6	32	High needs	Against	-0.190	No Flag	No Flag

(Continued on next page)

Table D1. NYSTP Mathematics 2010 Classical DIF Item Flags (cont.)

Grade	Item #	Subgroup	DIF	SMD	Mantel-Haenszel	Delta
6	33	Female	In Favor	0.125	No Flag	No Flag
6	34	Asian	In Favor	0.151	No Flag	No Flag
7	1	Female	Against	No Flag	3510.370	-1.949
7	2	Spanish	Against	-0.114	No Flag	No Flag
7	2	Female	Against	No Flag	3769.410	-2.014
7	8	Asian	Against	No Flag	516.307	-2.104
7	8	ELL	Against	-0.117	1218.840	-1.952
7	11	Spanish	Against	-0.161	530.110	-2.104
7	11	ELL	Against	-0.124	1132.540	-1.710
7	18	Female	Against	-0.107	No Flag	No Flag
7	33	Spanish	Against	-0.217	No Flag	No Flag
7	33	ELL	Against	-0.233	No Flag	No Flag
7	34	Black	Against	-0.177	No Flag	No Flag
7	35	Spanish	Against	-0.190	No Flag	No Flag
7	35	ELL	Against	-0.112	No Flag	No Flag
7	36	Asian	In Favor	0.140	No Flag	No Flag
7	36	Black	In Favor	0.168	No Flag	No Flag
7	36	Hispanic	In Favor	0.141	No Flag	No Flag
7	36	Female	In Favor	0.170	No Flag	No Flag
7	36	High needs	In Favor	0.101	No Flag	No Flag
7	37	Spanish	Against	-0.106	No Flag	No Flag
7	38	Spanish	Against	-0.112	No Flag	No Flag
7	38	Asian	Against	-0.111	No Flag	No Flag
7	38	Black	Against	-0.180	No Flag	No Flag
7	38	Hispanic	Against	-0.163	No Flag	No Flag
7	38	Female	In Favor	0.105	No Flag	No Flag
7	38	High needs	Against	-0.190	No Flag	No Flag
8	18	Spanish	Against	-0.121	391.689	-2.056
8	18	ELL	Against	No Flag	710.086	-1.594
8	31	Spanish	Against	-0.117	No Flag	No Flag
8	33	Black	Against	-0.139	No Flag	No Flag
8	34	High needs	Against	-0.122	No Flag	No Flag
8	36	Asian	In Favor	0.168	No Flag	No Flag
8	36	Black	In Favor	0.104	No Flag	No Flag
8	36	High needs	In Favor	0.110	No Flag	No Flag
8	38	Spanish	Against	-0.113	No Flag	No Flag
8	39	Spanish	Against	-0.144	No Flag	No Flag
8	40	Spanish	Against	-0.113	No Flag	No Flag
8	42	Spanish	Against	-0.123	No Flag	No Flag
8	43	Spanish	In Favor	0.104	No Flag	No Flag
8	43	Spanish	Against	-0.330	No Flag	No Flag
8	45	ELL	Against	-0.279	No Flag	No Flag

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

Grade	Item	Focal Group	Direction	Magnitude
4	43	Spanish	In Favor	0.108
5	28	Spanish	In Favor	0.129
5	34	ELL	Against	-0.105
6	11	Spanish	In Favor	0.125
6	27	Spanish	In Favor	0.110
6	32	Spanish	Against	-0.102
7	11	Spanish	Against	-0.121
7	33	ELL	Against	-0.147
7	33	Spanish	Against	-0.125
7	34	Black	Against	-0.130
7	34	Spanish	In Favor	0.107
7	36	ELL	In Favor	0.113
7	36	Spanish	In Favor	0.133
7	38	Black	Against	-0.115
8	29	Spanish	In Favor	0.104
8	35	Spanish	In Favor	0.108
8	43	Spanish	In Favor	0.132
8	45	Spanish	Against	-0.271

Appendix E—Item-Model Fit Statistics

Tables E1–E6 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 Grades 3–8 Mathematics Tests was acceptable (critical *Z* > obtained *Z*).

Table E1. Mathematics Grade 3 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	Z_{Q1}	Z_{Q1} critical	Fit OK?
1	3PL	84.34	7	172362	20.67	459.632	Y
2	3PL	108.47	7	172362	27.12	459.632	Y
3	3PL	660.65	7	172362	174.69	459.632	Y
4	3PL	250.83	7	172362	65.17	459.632	Y
5	3PL	53.75	7	172362	12.50	459.632	Y
6	3PL	145.51	7	172362	37.02	459.632	Y
7	3PL	166.43	7	172362	42.61	459.632	Y
8	3PL	963.03	7	172362	255.51	459.632	Y
9	3PL	399.55	7	172362	104.91	459.632	Y
10	3PL	86.42	7	172362	21.23	459.632	Y
11	3PL	57.52	7	172362	13.50	459.632	Y
12	3PL	397.58	7	172362	104.39	459.632	Y
13	3PL	480.87	7	172362	126.65	459.632	Y
14	3PL	164.55	7	172362	42.11	459.632	Y
15	3PL	88.63	7	172362	21.82	459.632	Y
16	3PL	182.22	7	172362	46.83	459.632	Y
17	3PL	187.35	7	172362	48.20	459.632	Y
18	3PL	172.77	7	172362	44.30	459.632	Y
19	3PL	926.20	7	172362	245.67	459.632	Y
20	3PL	51.61	7	172362	11.92	459.632	Y
21	3PL	67.81	7	172362	16.25	459.632	Y
22	3PL	87.68	7	172362	21.56	459.632	Y
23	3PL	310.04	7	172362	80.99	459.632	Y
24	3PL	570.38	7	172362	150.57	459.632	Y
25	3PL	322.81	7	172362	84.40	459.632	Y
26	2PPC	416.11	17	172362	68.45	459.632	Y
27	2PPC	1437.38	17	172362	243.59	459.632	Y
28	2PPC	1051.83	17	172362	177.47	459.632	Y
29	2PPC	2251.01	17	172362	383.13	459.632	Y
30	2PPC	889.10	26	172362	119.69	459.632	Y
31	2PPC	2286.07	26	172362	313.42	459.632	Y

Table E2. Mathematics Grade 4 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	Z_{Q1}	Z_{Q1} critical	Fit OK?
1	3PL	46.78	7	193793	10.63	516.781	Y
2	3PL	62.08	7	193793	14.72	516.781	Y
3	3PL	95.16	7	193793	23.56	516.781	Y
4	3PL	339.33	7	193793	88.82	516.781	Y
5	3PL	12.87	7	193793	1.57	516.781	Y
6	3PL	36.59	7	193793	7.91	516.781	Y
7	3PL	82.81	7	193793	20.26	516.781	Y
8	3PL	42.19	7	193793	9.40	516.781	Y
9	3PL	234.65	7	193793	60.84	516.781	Y
10	3PL	30.66	7	193793	6.32	516.781	Y
11	3PL	17.22	7	193793	2.73	516.781	Y
12	3PL	28.35	7	193793	5.71	516.781	Y
13	3PL	54.97	7	193793	12.82	516.781	Y
14	3PL	341.58	7	193793	89.42	516.781	Y
15	3PL	15.80	7	193793	2.35	516.781	Y
16	3PL	18.63	7	193793	3.11	516.781	Y
17	3PL	52.16	7	193793	12.07	516.781	Y
18	3PL	90.52	7	193793	22.32	516.781	Y
19	3PL	76.29	7	193793	18.52	516.781	Y
20	3PL	24.28	7	193793	4.62	516.781	Y
21	3PL	246.69	7	193793	64.06	516.781	Y
22	3PL	128.47	7	193793	32.46	516.781	Y
23	3PL	1108.83	7	193793	294.48	516.781	Y
24	3PL	139.18	7	193793	35.33	516.781	Y
25	3PL	68.48	7	193793	16.43	516.781	Y
26	3PL	286.94	7	193793	74.82	516.781	Y
27	3PL	44.16	7	193793	9.93	516.781	Y
28	3PL	20.90	7	193793	3.71	516.781	Y
29	3PL	101.21	7	193793	25.18	516.781	Y
30	3PL	91.84	7	193793	22.67	516.781	Y
31	2PPC	419.08	17	193793	68.96	516.781	Y
32	2PPC	1909.87	17	193793	324.62	516.781	Y
33	2PPC	173.70	17	193793	26.87	516.781	Y
34	2PPC	179.33	17	193793	27.84	516.781	Y
35	2PPC	843.25	17	193793	141.70	516.781	Y
36	2PPC	1191.87	17	193793	201.49	516.781	Y
37	2PPC	633.59	17	193793	105.74	516.781	Y
38	2PPC	1061.66	26	193793	143.62	516.781	Y
39	2PPC	1024.97	26	193793	138.53	516.781	Y
40	2PPC	5403.40	17	193793	923.76	516.781	N
41	2PPC	720.56	17	193793	120.66	516.781	Y
42	2PPC	943.08	17	193793	158.82	516.781	Y
43	2PPC	279.93	17	193793	45.09	516.781	Y
44	2PPC	1744.48	17	193793	296.26	516.781	Y

(Continued on next page)

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

Item	Model	Chi Square	DF	Total N	Z_{Q1}	Z_{Q1} critical	Fit OK?
45	2PPC	855.27	17	193793	143.76	516.781	Y
46	2PPC	3133.55	17	193793	534.48	516.781	N
47	2PPC	661.06	26	193793	88.07	516.781	Y
48	2PPC	1363.49	26	193793	185.48	516.781	Y

Table E3. Mathematics Grade 5 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	Z_{Q1}	Z_{Q1} critical	Fit OK?
1	3PL	33.56	7	192496	7.10	513.323	Y
2	3PL	124.16	7	192496	31.31	513.323	Y
3	3PL	286.68	7	192496	74.75	513.323	Y
4	3PL	51.94	7	192496	12.01	513.323	Y
5	3PL	40.63	7	192496	8.99	513.323	Y
6	3PL	218.55	7	192496	56.54	513.323	Y
7	3PL	39.28	7	192496	8.63	513.323	Y
8	3PL	33.14	7	192496	6.99	513.323	Y
9	3PL	289.23	7	192496	75.43	513.323	Y
10	3PL	35.11	7	192496	7.51	513.323	Y
11	3PL	107.50	7	192496	26.86	513.323	Y
12	3PL	139.14	7	192496	35.32	513.323	Y
13	3PL	521.18	7	192496	137.42	513.323	Y
14	3PL	245.08	7	192496	63.63	513.323	Y
15	3PL	36.66	7	192496	7.93	513.323	Y
16	3PL	31.91	7	192496	6.66	513.323	Y
17	3PL	208.85	7	192496	53.95	513.323	Y
18	3PL	41.31	7	192496	9.17	513.323	Y
19	3PL	138.37	7	192496	35.11	513.323	Y
20	3PL	148.88	7	192496	37.92	513.323	Y
21	3PL	239.89	7	192496	62.24	513.323	Y
22	3PL	94.50	7	192496	23.39	513.323	Y
23	3PL	167.97	7	192496	43.02	513.323	Y
24	3PL	80.84	7	192496	19.74	513.323	Y
25	3PL	104.48	7	192496	26.05	513.323	Y
26	3PL	230.19	7	192496	59.65	513.323	Y
27	2PPC	494.60	17	192496	81.91	513.323	Y
28	2PPC	593.50	17	192496	98.87	513.323	Y
29	2PPC	512.69	17	192496	85.01	513.323	Y
30	2PPC	2036.35	17	192496	346.32	513.323	Y
31	2PPC	633.92	26	192496	84.30	513.323	Y
32	2PPC	1292.39	26	192496	175.62	513.323	Y
33	2PPC	3312.12	26	192496	455.70	513.323	Y
34	2PPC	1448.44	26	192496	197.26	513.323	Y

Table E4. Mathematics Grade 6 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	Z_{Q1}	Z_{Q1} critical	Fit OK?
1	3PL	75.49	7	194773	18.30	519.395	Y
2	3PL	77.31	7	194773	18.79	519.395	Y
3	3PL	329.19	7	194773	86.11	519.395	Y
4	3PL	41.73	7	194773	9.28	519.395	Y
5	3PL	298.58	7	194773	77.93	519.395	Y
6	3PL	52.85	7	194773	12.25	519.395	Y
7	3PL	57.80	7	194773	13.58	519.395	Y
8	3PL	369.17	7	194773	96.79	519.395	Y
9	3PL	173.33	7	194773	44.45	519.395	Y
10	3PL	74.94	7	194773	18.16	519.395	Y
11	3PL	22.10	7	194773	4.04	519.395	Y
12	3PL	513.44	7	194773	135.35	519.395	Y
13	3PL	109.99	7	194773	27.53	519.395	Y
14	3PL	48.76	7	194773	11.16	519.395	Y
15	3PL	38.82	7	194773	8.50	519.395	Y
16	3PL	200.76	7	194773	51.79	519.395	Y
17	3PL	90.64	7	194773	22.35	519.395	Y
18	3PL	43.51	7	194773	9.76	519.395	Y
19	3PL	138.84	7	194773	35.23	519.395	Y
20	3PL	294.11	7	194773	76.73	519.395	Y
21	3PL	48.68	7	194773	11.14	519.395	Y
22	3PL	190.61	7	194773	49.07	519.395	Y
23	3PL	154.49	7	194773	39.42	519.395	Y
24	3PL	89.86	7	194773	22.14	519.395	Y
25	3PL	45.66	7	194773	10.33	519.395	Y
26	2PPC	1527.60	17	194773	259.07	519.395	Y
27	2PPC	1400.51	17	194773	237.27	519.395	Y
28	2PPC	3612.36	17	194773	616.60	519.39	N
29	2PPC	746.19	17	194773	125.06	519.395	Y
30	2PPC	753.31	17	194773	126.28	519.395	Y
31	2PPC	711.82	17	194773	119.16	519.395	Y
32	2PPC	2966.97	26	194773	407.84	519.395	Y
33	2PPC	1217.46	26	194773	165.23	519.395	Y
34	2PPC	1237.67	26	194773	168.03	519.395	Y
35	2PPC	521.12	26	194773	68.66	519.395	Y

Table E5. Mathematics Grade 7 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	Z_{Q1}	Z_{Q1} critical	Fit OK?
1	3PL	261.94	7	197405	68.14	526.413	Y
2	3PL	48.28	7	197405	11.03	526.413	Y
3	3PL	28.80	7	197405	5.83	526.413	Y
4	3PL	55.06	7	197405	12.84	526.413	Y
5	3PL	269.45	7	197405	70.14	526.413	Y
6	3PL	145.79	7	197405	37.09	526.413	Y
7	3PL	64.15	7	197405	15.27	526.413	Y
8	3PL	214.02	7	197405	55.33	526.413	Y
9	3PL	148.41	7	197405	37.79	526.413	Y
10	3PL	293.51	7	197405	76.57	526.413	Y
11	3PL	58.72	7	197405	13.82	526.413	Y
12	3PL	92.53	7	197405	22.86	526.413	Y
13	3PL	21.53	7	197405	3.88	526.413	Y
14	3PL	63.74	7	197405	15.16	526.413	Y
15	3PL	19.17	7	197405	3.25	526.413	Y
16	3PL	154.96	7	197405	39.54	526.413	Y
17	3PL	293.97	7	197405	76.70	526.413	Y
18	3PL	100.60	7	197405	25.01	526.413	Y
19	3PL	264.68	7	197405	68.87	526.413	Y
20	3PL	242.62	7	197405	62.97	526.413	Y
21	3PL	94.06	7	197405	23.27	526.413	Y
22	3PL	669.06	7	197405	176.94	526.413	Y
23	3PL	205.65	7	197405	53.09	526.413	Y
24	3PL	82.28	7	197405	20.12	526.413	Y
25	3PL	210.76	7	197405	54.46	526.413	Y
26	3PL	188.69	7	197405	48.56	526.413	Y
27	3PL	127.55	7	197405	32.22	526.413	Y
28	3PL	134.94	7	197405	34.19	526.413	Y
29	3PL	89.35	7	197405	22.01	526.413	Y
30	3PL	1277.16	7	197405	339.46	526.413	Y
31	2PPC	1524.11	17	197405	258.47	526.413	Y
32	2PPC	235.64	17	197405	37.50	526.413	Y
33	2PPC	339.78	17	197405	55.36	526.413	Y
34	2PPC	421.49	17	197405	69.37	526.413	Y
35	2PPC	5455.65	26	197405	752.96	526.413	N
36	2PPC	1305.86	26	197405	177.48	526.413	Y
37	2PPC	2844.88	26	197405	390.91	526.413	Y
38	2PPC	1558.41	26	197405	212.51	526.413	Y

Table E6. Mathematics Grade 8 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	Z_{Q1}	Z_{Q1} critical	Fit OK?
1	3PL	68.56	7	199307	16.45	531.485	Y
2	3PL	55.65	7	199307	13.00	531.485	Y
3	3PL	55.19	7	199307	12.88	531.485	Y
4	3PL	227.82	7	199307	59.02	531.485	Y
5	3PL	20.49	7	199307	3.60	531.485	Y
6	3PL	24.93	7	199307	4.79	531.485	Y
7	3PL	59.26	7	199307	13.97	531.485	Y
8	3PL	131.26	7	199307	33.21	531.485	Y
9	3PL	66.78	7	199307	15.98	531.485	Y
10	3PL	74.07	7	199307	17.93	531.485	Y
11	3PL	35.34	7	199307	7.57	531.485	Y
12	3PL	77.18	7	199307	18.76	531.485	Y
13	3PL	160.54	7	199307	41.03	531.485	Y
14	3PL	107.18	7	199307	26.77	531.485	Y
15	3PL	30.08	7	199307	6.17	531.485	Y
16	3PL	17.05	7	199307	2.69	531.485	Y
17	3PL	49.03	7	199307	11.23	531.485	Y
18	3PL	60.84	7	199307	14.39	531.485	Y
19	3PL	21.57	7	199307	3.89	531.485	Y
20	3PL	55.30	7	199307	12.91	531.485	Y
21	3PL	30.58	7	199307	6.30	531.485	Y
22	3PL	144.80	7	199307	36.83	531.485	Y
23	3PL	113.92	7	199307	28.57	531.485	Y
24	3PL	129.01	7	199307	32.61	531.485	Y
25	3PL	32.72	7	199307	6.87	531.485	Y
26	3PL	90.26	7	199307	22.25	531.485	Y
27	3PL	22.90	7	199307	4.25	531.485	Y
28	2PPC	401.30	17	199307	65.91	531.485	Y
29	2PPC	6258.79	17	199307	1070.46	531.485	N
30	2PPC	1213.99	17	199307	205.28	531.485	Y
31	2PPC	1219.23	17	199307	206.18	531.485	Y
32	2PPC	333.67	26	199307	42.67	531.485	Y
33	2PPC	317.93	26	199307	40.48	531.485	Y
34	2PPC	318.23	17	199307	51.66	531.485	Y
35	2PPC	4631.25	17	199307	791.34	531.485	N
36	2PPC	3457.77	17	199307	590.09	531.485	N
37	2PPC	198.44	17	199307	31.12	531.485	Y
38	2PPC	1059.12	17	199307	178.72	531.485	Y
39	2PPC	468.62	17	199307	77.45	531.485	Y
40	2PPC	617.91	17	199307	103.05	531.485	Y
41	2PPC	194.91	17	199307	30.51	531.485	Y
42	2PPC	1172.34	26	199307	158.97	531.485	Y
43	2PPC	1928.09	26	199307	263.77	531.485	Y
44	2PPC	377.46	26	199307	48.74	531.485	Y
45	2PPC	274.90	26	199307	34.52	531.485	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:

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

where

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

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

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

with

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

and

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

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

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

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

Estimation of the Prior Distribution of T_j

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

The 3PL model is

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

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, & Sykes, 1996). For a CR item with 1_i score levels, integer scores were assigned that ranged from 0 to $1_i - 1$:

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

where

$$z_{ig} = \alpha_i (m-1)\theta - \sum_{h=0}^{m-1} \gamma_{ih}{}' \tag{7}$$

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

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

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 $g(\hat{T}_j \,|\, \hat{\theta})$ with mean $\mu(\hat{T}_j | \theta)$ and variance $\sigma^2(\hat{T}_j \,|\, \theta)$. 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 $[\mu(\hat{T}_j \,|\, \theta)]$ and variance $[\sigma^2(\hat{T}_j \,|\, \theta)]$ 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 (Novick & Jackson, 1974, p. 113) produces

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

and

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

Solving these equations for r_j and s_j produces

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

and

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

where

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

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

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

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

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

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 3PL IRT and 2PPC 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 & 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, & 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 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

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

and

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

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 the 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

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

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—Concordance Tables

Table H1. Grade 3 Mathematics 2010 and TerraNova Scale Score Concordance Table

Raw Score OP	Scale Score OP	Scale Score TERRANOVA	NP	NCE
4	470	385	1	1
5	600	477	1	2
6	613	503	2	6
7	620	513	2	7
8	625	522	2	9
9	629	529	3	10
10	633	534	4	12
11	636	539	4	14
12	638	543	5	15
13	641	548	6	17
14	643	551	7	18
15	645	555	8	20
16	647	558	9	21
17	649	561	10	23
18	651	564	11	24
19	653	567	12	25
20	655	570	14	27
21	657	573	15	28
22	659	576	17	30
23	660	579	18	31
24	662	582	20	32
25	664	585	22	34
26	666	588	25	36
27	668	591	27	37
28	670	594	30	39
29	672	598	33	41
30	674	601	36	43
31	676	605	40	45
32	678	609	45	47
33	681	613	49	49
34	684	618	54	52
35	687	624	61	56
36	691	632	69	60
37	697	642	77	66
38	707	660	88	75
39	770	740	99	99

Table H2. Grade 4 Mathematics 2010 and *TerraNova* Scale Score Concordance Table

Raw Score OP	Scale Score OP	Scale Score TERRANOVA	NP	NCE
5	485	403	1	1
6	537	507	2	4
7	559	522	2	8
8	572	535	3	10
9	581	542	4	12
10	588	550	4	14
11	594	556	5	16
12	599	561	6	17
13	603	565	7	19
14	607	570	8	20
15	611	573	9	22
16	614	577	10	23
17	617	580	11	24
18	620	582	12	25
19	623	585	13	26
20	625	588	14	27
21	628	590	15	28
22	630	592	16	29
23	632	594	17	30
24	634	596	18	30
25	636	598	19	31
26	638	600	20	32
27	640	602	21	33
28	641	604	22	34
29	643	605	23	34
30	645	607	24	35
31	646	608	25	35
32	648	610	26	36
33	650	612	27	37
34	651	613	28	38
35	653	615	30	39
36	654	616	30	39
37	655	618	32	40
38	657	619	33	41
39	658	621	34	42
40	660	622	35	42
41	661	624	37	43
42	663	625	38	44

(Continued on next page)

Table H2. Grade 4 Mathematics 2010 and *TerraNova* Scale Score Concordance Table (cont.)

Raw Score OP	Scale Score OP	Scale Score TERRA NOVA	NP	NCE
43	664	627	40	45
44	665	628	41	45
45	667	630	43	46
46	668	632	45	47
47	670	633	46	48
48	671	635	48	49
49	673	636	49	49
50	675	638	51	50
51	676	640	53	51
52	678	642	55	53
53	680	644	57	54
54	682	645	58	54
55	683	647	60	55
56	685	650	63	57
57	688	652	65	58
58	690	654	67	59
59	692	656	69	61
60	695	659	72	62
61	697	662	75	64
62	700	665	78	66
63	704	669	81	68
64	707	673	84	71
65	712	677	87	73
66	717	683	90	77
67	724	691	93	82
68	734	703	96	87
69	751	726	98	94
70	800	770	99	99

Table H3. Grade 5 Mathematics 2010 and TerraNova Scale Score Concordance Table

Raw Score OP	Scale Score OP	Scale Score TERRANOVA	NP	NCE
5	495	430	1	1
6	542	517	1	3
7	568	538	2	6
8	583	553	3	9
9	594	565	4	13
10	603	575	5	16
11	609	582	6	18
12	615	589	8	20
13	620	594	9	22
14	625	599	11	24
15	629	604	13	26
16	633	608	15	28
17	636	612	17	30
18	640	615	19	31
19	643	619	21	33
20	645	622	23	34
21	648	625	25	36
22	651	628	27	37
23	653	631	30	39
24	656	634	32	40
25	658	637	35	42
26	660	640	37	43
27	662	642	39	44
28	664	645	42	46
29	667	648	45	47
30	669	650	47	48
31	671	653	50	50
32	673	656	53	52
33	676	659	56	53
34	678	662	60	55
35	680	665	63	57
36	683	668	66	59
37	686	672	70	61
38	689	676	74	63
39	693	680	77	66
40	697	686	82	69
41	701	692	86	73
42	707	699	90	77
43	714	709	94	83
44	725	724	97	90
45	744	753	99	99
46	780	797	99	99

Table H4. Grade 6 Mathematics 2010 and *TerraNova* Scale Score Concordance Table

Raw Score OP	Scale Score OP	Scale Score TERRANOVA	NP	NCE
6	500	538	1	4
7	562	560	2	8
8	583	574	3	11
9	595	583	4	13
10	603	590	5	16
11	609	597	7	18
12	614	602	8	20
13	619	607	9	22
14	622	611	11	24
15	626	614	12	25
16	629	618	13	27
17	632	621	15	28
18	635	624	16	29
19	637	627	18	30
20	640	630	19	32
21	642	633	21	33
22	644	636	23	34
23	647	639	25	36
24	649	641	26	36
25	651	644	28	38
26	653	647	30	39
27	655	650	33	41
28	658	652	34	41
29	660	655	37	43
30	662	658	39	44
31	664	661	42	46
32	667	664	45	47
33	669	667	48	49
34	671	670	50	50
35	674	674	54	52
36	676	677	57	54
37	679	680	60	55
38	682	684	63	57
39	685	688	67	59
40	688	692	70	61
41	692	697	75	64
42	695	702	79	67
43	700	708	83	70

(Continued on next page)

Table H4. Grade 6 Mathematics 2010 and *TerraNova* Scale Score Concordance Table (cont.)

Raw Score OP	Scale Score OP	Scale Score TERRANOVA	NP	NCE
44	705	715	88	74
45	711	723	92	79
46	719	734	96	86
47	731	749	98	95
48	751	781	99	99
49	780	820	99	99

Table H5. Grade 7 Mathematics 2010 and *TerraNova* Scale Score Concordance Table

Raw Score OP	Scale Score OP	Scale Score TERRANOVA	NP	NCE
6	500	520	1	1
7	540	552	2	5
8	579	573	3	9
9	595	586	4	13
10	604	596	6	16
11	611	603	7	18
12	617	610	8	21
13	622	616	10	23
14	626	621	11	25
15	630	625	13	26
16	633	629	15	28
17	636	633	17	30
18	639	637	19	32
19	642	641	22	34
20	644	644	24	35
21	647	647	26	37
22	649	651	29	38
23	652	654	31	40
24	654	657	34	41
25	656	660	36	43
26	659	663	39	44
27	661	666	41	45
28	663	669	44	47
29	665	672	46	48
30	668	675	49	49

(Continued on next page)

Table H5. Grade 7 Mathematics 2010 and *TerraNova* Scale Score Concordance Table (cont.)

Raw Score OP	Scale Score OP	Scale Score TERRANOVA	NP	NCE
31	670	679	53	51
32	672	682	55	53
33	675	685	58	54
34	677	689	62	56
35	680	692	64	58
36	683	696	68	60
37	685	700	71	62
38	688	704	75	64
39	691	708	78	66
40	694	712	81	68
41	697	717	84	71
42	701	722	87	74
43	705	727	90	77
44	709	733	93	81
45	714	740	95	85
46	719	748	97	90
47	726	758	98	95
48	736	772	99	99
49	752	797	99	99
50	800	850	99	99

Table H6. Grade 8 Mathematics 2010 and *TerraNova* Scale Score Concordance Table

Raw Score OP	Scale Score OP	Scale Score TERRANOVA	NP	NCE
4	480	502	1	1
5	532	552	2	4
6	574	572	2	7
7	588	585	2	9
8	596	594	3	11
9	603	600	4	12
10	608	606	5	14
11	612	611	5	16
12	615	615	6	18
13	619	620	8	20
14	621	623	8	21

(Continued on next page)

Table H6. Grade 8 Mathematics 2010 and *TerraNova* Scale Score Concordance Table (cont.)

Raw Score OP	Scale Score OP	Scale Score TERRANOVA	NP	NCE
15	624	627	10	23
16	626	630	11	24
17	629	633	12	25
18	631	635	13	26
19	633	638	14	28
20	634	641	16	29
21	636	643	17	30
22	638	645	18	31
23	639	647	19	32
24	641	649	20	32
25	642	651	21	33
26	644	653	23	34
27	645	655	24	35
28	647	656	24	35
29	648	658	26	36
30	649	660	27	37
31	650	662	28	38
32	652	663	29	38
33	653	665	30	39
34	654	667	32	40
35	655	668	33	41
36	657	670	34	41
37	658	672	36	42
38	659	673	36	43
39	660	675	38	44
40	661	676	39	44
41	662	678	40	45
42	664	680	42	46
43	665	681	43	46
44	666	683	45	47
45	667	685	46	48
46	669	686	47	49
47	670	688	49	49
48	671	690	51	50
49	672	692	53	51
50	674	694	54	52
51	675	696	56	53
52	677	698	58	54

(Continued on next page)

Table H6. Grade 8 Mathematics 2010 and *TerraNova* Scale Score Concordance Table (cont.)

Raw Score OP	Scale Score OP	Scale Score TERRANOVA	NP	NCE
53	678	700	60	55
54	680	702	62	56
55	681	705	64	58
56	683	708	67	59
57	685	710	69	60
58	687	713	71	62
59	689	716	74	63
60	691	720	77	65
61	694	724	80	67
62	697	728	82	69
63	700	733	85	72
64	704	739	88	75
65	709	746	91	78
66	716	755	94	83
67	725	768	97	90
68	741	793	99	99
69	775	872	99	99

Appendix I—Scale Score Frequency Distributions

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

Table I1. Grade 3 Mathematics 2010 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
470	128	0.06	128	0.06
600	95	0.05	223	0.11
613	196	0.10	419	0.21
620	241	0.12	660	0.33
625	310	0.16	970	0.49
629	404	0.20	1374	0.69
633	444	0.22	1818	0.92
636	509	0.26	2327	1.17
638	567	0.29	2894	1.46
641	651	0.33	3545	1.79
643	762	0.38	4307	2.17
645	878	0.44	5185	2.61
647	997	0.50	6182	3.11
649	1158	0.58	7340	3.70
651	1321	0.67	8661	4.36
653	1426	0.72	10087	5.08
655	1736	0.87	11823	5.95
657	1894	0.95	13717	6.91
659	2229	1.12	15946	8.03
660	2515	1.27	18461	9.30
662	2933	1.48	21394	10.78
664	3366	1.70	24760	12.47
666	4020	2.02	28780	14.50
668	4522	2.28	33302	16.77
670	5352	2.70	38654	19.47
672	6057	3.05	44711	22.52
674	7111	3.58	51822	26.10

(Continued on next page)

Table I1. Grade 3 Mathematics 2010 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
676	8353	4.21	60175	30.31
678	9553	4.81	69728	35.12
681	11276	5.68	81004	40.80
684	13295	6.70	94299	47.49
687	15799	7.96	110098	55.45
691	18881	9.51	128979	64.96
697	21827	10.99	150806	75.95
707	24277	12.23	175083	88.18
770	23466	11.82	198549	100.00

Table I2. Grade 4 Mathematics 2010 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
485	89	0.04	89	0.04
537	104	0.05	193	0.10
559	112	0.06	305	0.15
572	165	0.08	470	0.23
581	195	0.10	665	0.33
588	268	0.13	933	0.46
594	308	0.15	1241	0.62
599	361	0.18	1602	0.80
603	390	0.19	1992	0.99
607	516	0.26	2508	1.25
611	505	0.25	3013	1.50
614	605	0.30	3618	1.80
617	603	0.30	4221	2.10
620	648	0.32	4869	2.42
623	801	0.40	5670	2.82
625	828	0.41	6498	3.23
628	886	0.44	7384	3.67
630	993	0.49	8377	4.16
632	1062	0.53	9439	4.69

(Continued on next page)

Table I2. Grade 4 Mathematics 2010 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
634	1165	0.58	10604	5.26
636	1157	0.57	11761	5.84
638	1267	0.63	13028	6.47
640	1274	0.63	14302	7.10
641	1391	0.69	15693	7.79
643	1430	0.71	17123	8.50
645	1524	0.76	18647	9.26
646	1593	0.79	20240	10.05
648	1741	0.86	21981	10.91
650	1773	0.88	23754	11.79
651	1893	0.94	25647	12.73
653	2001	0.99	27648	13.73
654	2160	1.07	29808	14.80
655	2243	1.11	32051	15.91
657	2320	1.15	34371	17.06
658	2415	1.20	36786	18.26
660	2563	1.27	39349	19.54
661	2699	1.34	42048	20.88
663	2694	1.34	44742	22.21
664	2949	1.46	47691	23.68
665	3067	1.52	50758	25.20
667	3232	1.60	53990	26.80
668	3349	1.66	57339	28.47
670	3593	1.78	60932	30.25
671	3663	1.82	64595	32.07
673	4071	2.02	68666	34.09
675	4049	2.01	72715	36.10
676	4363	2.17	77078	38.27
678	4669	2.32	81747	40.59
680	4909	2.44	86656	43.02
682	5076	2.52	91732	45.54
683	5323	2.64	97055	48.19

(Continued on next page)

Table I2. Grade 4 Mathematics 2010 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
685	5539	2.75	102594	50.94
688	5878	2.92	108472	53.85
690	6137	3.05	114609	56.90
692	6448	3.20	121057	60.10
695	6679	3.32	127736	63.42
697	7174	3.56	134910	66.98
700	7210	3.58	142120	70.56
704	7424	3.69	149544	74.25
707	7937	3.94	157481	78.19
712	8226	4.08	165707	82.27
717	8241	4.09	173948	86.36
724	8459	4.20	182407	90.56
734	7965	3.95	190372	94.52
751	6918	3.43	197290	97.95
800	4128	2.05	201418	100.00

Table I3. Grade 5 Mathematics 2010 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
495	233	0.12	233	0.12
542	208	0.10	441	0.22
568	318	0.16	759	0.38
583	464	0.23	1223	0.61
594	593	0.30	1816	0.91
603	751	0.38	2567	1.29
609	819	0.41	3386	1.70
615	986	0.49	4372	2.19
620	1120	0.56	5492	2.76
625	1344	0.67	6836	3.43
629	1492	0.75	8328	4.18
633	1658	0.83	9986	5.01
636	1953	0.98	11939	5.99
640	2100	1.05	14039	7.05

(Continued on next page)

Table I3. Grade 5 Mathematics 2010 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
643	2365	1.19	16404	8.23
645	2589	1.30	18993	9.53
648	2790	1.40	21783	10.93
651	3032	1.52	24815	12.45
653	3314	1.66	28129	14.12
656	3537	1.78	31666	15.89
658	3635	1.82	35301	17.72
660	4106	2.06	39407	19.78
662	4351	2.18	43758	21.96
664	4467	2.24	48225	24.20
667	5007	2.51	53232	26.72
669	5276	2.65	58508	29.36
671	5708	2.86	64216	32.23
673	5997	3.01	70213	35.24
676	6429	3.23	76642	38.46
678	7160	3.59	83802	42.06
680	7618	3.82	91420	45.88
683	8193	4.11	99613	49.99
686	8935	4.48	108548	54.48
689	9667	4.85	118215	59.33
693	10379	5.21	128594	64.54
697	11200	5.62	139794	70.16
701	11821	5.93	151615	76.09
707	12471	6.26	164086	82.35
714	12095	6.07	176181	88.42
725	10825	5.43	187006	93.85
744	8290	4.16	195296	98.01
780	3958	1.99	199254	100.00

Table I4. Grade 6 Mathematics 2010 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
500	931	0.46	931	0.46
562	530	0.26	1461	0.73
583	661	0.33	2122	1.06
595	673	0.34	2795	1.39
603	804	0.40	3599	1.80
609	902	0.45	4501	2.25
614	1043	0.52	5544	2.77
619	1160	0.58	6704	3.35
622	1145	0.57	7849	3.92
626	1326	0.66	9175	4.58
629	1408	0.70	10583	5.28
632	1560	0.78	12143	6.06
635	1811	0.90	13954	6.96
637	1996	1.00	15950	7.96
640	2125	1.06	18075	9.02
642	2435	1.21	20510	10.23
644	2503	1.25	23013	11.48
647	2853	1.42	25866	12.91
649	2972	1.48	28838	14.39
651	3170	1.58	32008	15.97
653	3637	1.81	35645	17.79
655	4032	2.01	39677	19.80
658	4275	2.13	43952	21.93
660	4642	2.32	48594	24.25
662	5056	2.52	53650	26.77
664	5363	2.68	59013	29.45
667	5709	2.85	64722	32.29
669	6147	3.07	70869	35.36
671	6367	3.18	77236	38.54
674	6929	3.46	84165	42.00
676	7456	3.72	91621	45.72
679	7818	3.90	99439	49.62
682	8358	4.17	107797	53.79
685	8926	4.45	116723	58.24

(Continued on next page)

Table I4. Grade 6 Mathematics 2010 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
688	9286	4.63	126009	62.87
692	9694	4.84	135703	67.71
695	10224	5.10	145927	72.81
700	10316	5.15	156243	77.96
705	10370	5.17	166613	83.13
711	9992	4.99	176605	88.12
719	8966	4.47	185571	92.59
731	7423	3.70	192994	96.30
751	5021	2.51	198015	98.80
780	2400	1.20	200415	100.00

Table I5. Grade 7 Mathematics 2010 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
500	621	0.31	621	0.31
540	503	0.25	1124	0.56
579	680	0.34	1804	0.89
595	873	0.43	2677	1.32
604	1033	0.51	3710	1.83
611	1177	0.58	4887	2.42
617	1404	0.69	6291	3.11
622	1564	0.77	7855	3.88
626	1801	0.89	9656	4.77
630	2020	1.00	11676	5.77
633	2198	1.09	13874	6.86
636	2523	1.25	16397	8.10
639	2768	1.37	19165	9.47
642	3013	1.49	22178	10.96
644	3303	1.63	25481	12.59
647	3606	1.78	29087	14.37
649	3911	1.93	32998	16.31

(Continued on next page)

Table I5. Grade 7 Mathematics 2010 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
652	4286	2.12	37284	18.42
654	4678	2.31	41962	20.74
656	4805	2.37	46767	23.11
659	5249	2.59	52016	25.70
661	5535	2.74	57551	28.44
663	5866	2.90	63417	31.34
665	6111	3.02	69528	34.36
668	6368	3.15	75896	37.51
670	6766	3.34	82662	40.85
672	7080	3.50	89742	44.35
675	7310	3.61	97052	47.96
677	7444	3.68	104496	51.64
680	7618	3.76	112114	55.40
683	7657	3.78	119771	59.19
685	7844	3.88	127615	63.06
688	7841	3.87	135456	66.94
691	7863	3.89	143319	70.82
694	7819	3.86	151138	74.69
697	7888	3.90	159026	78.59
701	7460	3.69	166486	82.27
705	7281	3.60	173767	85.87
709	6603	3.26	180370	89.13
714	6129	3.03	186499	92.16
719	5081	2.51	191580	94.67
726	4240	2.10	195820	96.77
736	3274	1.62	199094	98.39
752	2092	1.03	201186	99.42
800	1173	0.58	202359	100.00

Table I6. Grade 8 Mathematics 2010 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
480	269	0.13	269	0.13
532	242	0.12	511	0.25
574	373	0.18	884	0.43
588	501	0.24	1385	0.67
596	645	0.31	2030	0.98
603	678	0.33	2708	1.31
608	834	0.40	3542	1.72
612	925	0.45	4467	2.16
615	986	0.48	5453	2.64
619	1058	0.51	6511	3.16
621	1135	0.55	7646	3.71
624	1251	0.61	8897	4.31
626	1213	0.59	10110	4.90
629	1297	0.63	11407	5.53
631	1350	0.65	12757	6.18
633	1453	0.70	14210	6.89
634	1490	0.72	15700	7.61
636	1593	0.77	17293	8.38
638	1676	0.81	18969	9.19
639	1643	0.80	20612	9.99
641	1701	0.82	22313	10.81
642	1799	0.87	24112	11.69
644	1921	0.93	26033	12.62
645	1979	0.96	28012	13.58
647	2037	0.99	30049	14.56
648	2141	1.04	32190	15.60
649	2228	1.08	34418	16.68
650	2240	1.09	36658	17.77
652	2343	1.14	39001	18.90
653	2409	1.17	41410	20.07
654	2582	1.25	43992	21.32
655	2589	1.25	46581	22.57
657	2669	1.29	49250	23.87
658	2776	1.35	52026	25.21

(Continued on next page)

Table I6. Grade 8 Mathematics 2010 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
659	2855	1.38	54881	26.60
660	3016	1.46	57897	28.06
661	3020	1.46	60917	29.52
662	3186	1.54	64103	31.07
664	3270	1.58	67373	32.65
665	3391	1.64	70764	34.29
666	3455	1.67	74219	35.97
667	3564	1.73	77783	37.70
669	3684	1.79	81467	39.48
670	3804	1.84	85271	41.32
671	3873	1.88	89144	43.20
672	3987	1.93	93131	45.13
674	4119	2.00	97250	47.13
675	4236	2.05	101486	49.18
677	4398	2.13	105884	51.31
678	4609	2.23	110493	53.55
680	4823	2.34	115316	55.88
681	5026	2.44	120342	58.32
683	5183	2.51	125525	60.83
685	5305	2.57	130830	63.4
687	5550	2.69	136380	66.09
689	5866	2.84	142246	68.94
691	6203	3.01	148449	71.94
694	6412	3.11	154861	75.05
697	6696	3.25	161557	78.29
700	7089	3.44	168646	81.73
704	7217	3.50	175863	85.23
709	7234	3.51	183097	88.73
716	7056	3.42	190153	92.15
725	6621	3.21	196774	95.36
741	5574	2.70	202348	98.06
775	3998	1.94	206346	100.00

References

American Educational Research Association, American Psychological Association & 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. & 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.

CTB/McGraw-Hill. (1996). TerraNova™ *assessment series (1st ed.)*. Monterey, CA: CTB/McGraw-Hill.

CTB/McGraw-Hill. (2000). TerraNova™ *assessment series (2nd ed.)*. Monterey, CA: CTB/McGraw-Hill.

CTB/McGraw-Hill. (2006). TerraNova™ *assessment series (3rd ed.)*. Monterey, CA: CTB/McGraw-Hill.

Dorans, N.J., Schmitt, A.P. & Bleistein, C.A. (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. & Julian, M.W. (1996). *Two studies comparing the parameter estimates produced by PARDUX and PARSCLE*. Monterey, CA: CTB/McGraw-Hill**.**

Fitzpatrick, A.R., Link, V., Yen, W.M., Burket, G., Ito, K. & Sykes, R. (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., Yen, W.M., & Burket, G.R.. (1989). Experiences in the application of item response theory in test construction. *Applied Measurement in Education* 2:297–312.

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

Huynh, H. & Schneider, C. (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. & Kotz, S.. (1970). *Distributions in statistics: Continuous univariate distributions* (Vol. 2). New York: John Wiley.

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

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

Lee, W., Hanson, B.A. & Brennan, R.L. (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. & Harnisch, D. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement* 18:109–118.

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

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

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

Mehrens, W.A. & Lehmann, I.J. (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. & Bock, R.D. (1991). *PARSCALE: Parameter scaling of rating data* [Computer program]. Chicago: Scientific Software, Inc.

Novick, M.R. & Jackson, P.H. (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. & Mille, M.P. (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. & Lord, F.M. (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., Kolen, J. & Harris, D.J. (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. & Linacre, J.M. (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., Sykes, R.C., Ito, K. & Julian, M. (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., Donoghue, J.R. & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement* 36:225–33.