New York State Testing Program 2011: Mathematics, Grades 3–8

Technical Report

CTB/McGraw-Hill
Monterey, California 93940
2011

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 2011 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 NYSTP Grades 3–8 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 performances.

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 2011, nonpublic schools participated in all grade tests but were not well represented in the testing program. The New York State Education Department (NYSED) made a decision to exclude these schools from the data analyses. Public school students were required to take all State assessments administered at their grade level, except for a very small percentage of students with disabilities who took the New York State Alternate Assessment (NYSAA) for students with severe disabilities. For more detail on this exemption, please refer to the *School Administrator's Manual for Public and Nonpublic Schools* (SAM), available online at http://www.p12.nysed.gov/apda/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 examinations 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 the *New York State Testing Program 2010: Mathematics, Grades 3–8 Technical 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.

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/apda/math/samplers/chinese/ (Chinese)
http://www.p12.nysed.gov/apda/math/samplers/haitian/ (Haitian Creole)
http://www.p12.nysed.gov/apda/math/samplers/korean/ (Korean)
http://www.p12.nysed.gov/apda/math/samplers/russian/ (Russian)
http://www.p12.nysed.gov/apda/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 during May 2011 over a three-day period. The tests were printed in black and white and incorporated the concepts of universal design. Details on the administration and scoring of these tests can be found in Section IV, "Test Administration and Scoring."

Test Configuration

The OP tests books were administered, in order, on two consecutive days. 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 contained only CR items. The 2011 *Teacher's Directions* (http://www.p12.nysed.gov/apda/ei/directions/m3-5-td-11.pdf and http://www.p12.nysed.gov/apda/ei/directions/m6-8-td-11.pdf) as well as the 2011 *School Administrator's Manual* (http://www.p12.nysed.gov/apda/sam/math/mathei-sam-11.pdf) provide details on security, scheduling, classroom organization and preparation, test materials, and administration.

Table 1. NYSTP Mathematics 2011 Test Configuration

Grade	Day	Book	Number of Items				Allotted Time (minutes)	
			MC	SR	ER	Total	Testing	Prep
3	1	1	40	0	0	40	60	10
	2	2	0	4	2	6	40	10
	Totals		40	4	2	46	100	20
4	1	1	45	0	0	45	70	10
	2	2	0	8	4	12	70	10
	Totals		45	8	4	57	140	20
5	1	1	41	0	0	41	60	10
	2	2	0	4	4	8	50	10
	Totals		41	4	4	49	110	20
6	1	1	40	0	0	40	60	10
	2	2	0	6	4	10	60	10
	Totals		40	6	4	50	120	20

(Continued on next page)

Table 1. NYSTP Mathematics 2011 Test Configuration (cont.)

Grade	Day	Book	Number of Items				Allotted Time (minutes)	
			MC	SR	ER	Total	Testing	Prep
7	1	1	45	0	0	45	70	10
	2	2	0	4	4	8	55	10
	Totals		45	4	4	53	125	20
8	1	1	42	0	0	42	65	10
	1	2	0	8	4	12	70	10
	Totals		42	8	4	54	135	20

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 the 2011 OP tests.

Table 2. NYSTP Mathematics 2011 Test Blueprint

Grade	Total Points	Content Strand	Target Points	Selected Points	Target % of Test	Selected % of Test
3	54	Number Sense and Operations	26	25	48.0	46.0
		Algebra	7	7	13.0	13.0
		Geometry	7	7	13.0	13.0
		Measurement	7	7	13.0	13.0
		Statistics and Probability	7	8	13.0	15.0

(Continued on next page)

Table 2. NYSTP Mathematics 2011 Test Blueprint (cont.)

Grade	Total Points	Content Strand	Target Points	Selected Points	Target % of Test	Selected % of Test
4	73	Number Sense and Operations	33	34	45.0	46.0
		Algebra	10	9	14.0	12.0
		Geometry	9	10	12.0	14.0
		Measurement	12	10	17.0	14.0
		Statistics and Probability	9	10	12.0	14.0
5	61	Number Sense and Operations	24	20	39.0	33.0
		Algebra	7	9	11.0	15.0
		Geometry	15	15	25.0	24.0
		Measurement	9	11	14.0	18.0
		Statistics and Probability	7	6	11.0	10.0
6	64	Number Sense and Operations	24	25	37.0	39.0
		Algebra	12	10	19.0	16.0
		Geometry	11	12	16.5	19.0
		Measurement	7	6	11.0	9.0
		Statistics and Probability	11	11	16.5	17.0
7	65	Number Sense and Operations	20	17	30.0	26.0
		Algebra	8	9	12.0	14.0
		Geometry	9	8	14.0	12.5
		Measurement	9	8	14.0	12.5
		Statistics and Probability	20	23	30.0	35.0
8	70	Number Sense and Operations	8	10	11.0	14.0
		Algebra	31	32	44.0	46.0
		Geometry	25	22	35.0	31.0
		Measurement	7	6	10.0	9.0

New York State Educator's Involvement in Test Development

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

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

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

Content Rationale

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

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

The committees were composed of teachers from around the state selected for their grade-level expertise, were grouped by grade band (i.e., 3/4, 5/6, 7/8), and met for four days. The

committees were composed of approximately 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 response and distracters (MC items)
- the conciseness, preciseness, clarity, and readability of the items
- the existence of any ethnic, gender, regional, or other possible bias evident in the items

Upon completion of the committee work, NYSED reviewed the decisions of the committee members; NYSED either approved the changes to the items or suggested additional revisions so that the nature and format of the items were consistent across grades and with the format and style of the testing program. All approved changes were then incorporated into the items prior to field testing.

Materials Development

Following item review, CTB/McGraw-Hill staff assembled the approved items into field test (FT) forms and submitted the FT forms to NYSED for their review and approval. The FT forms were administered to students across New York State, using either the State Sampling Matrix (from 2005 to 2009) to ensure appropriate sampling of students or a census sample (in 2010). In addition, CTB/McGraw-Hill, in conjunction with NYSED test specialists, developed a combined *Teacher's Directions and School Administrator's Manual* so that the FT forms were administered in a uniform manner to all participating students. FT forms were assigned to participants at the school (grade) level while balancing the demographic statistics across forms, in order to proactively sample the students.

After administration of the FT forms, rangefinding sessions were conducted 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 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 the NYSTP Grades 3–8 Mathematics OP Tests were administered in May 2011. The test items were selected from the pool of items primarily field-tested in 2007, 2008, and 2009, using the data from those FT forms. The pool also included items owned by CTB/McGraw-Hill. These items consisted mostly of *TerraNova*™ items but also included items field-tested in New York State in 2010 and newly developed items that had not yet been field-tested. Using this extended pool, CTB/McGraw-Hill made preliminary selections for each grade. The selections were reviewed for alignment with the test design, blueprint, and research guidelines for item selection (Appendix B). Item selection for the Grades 3–8 Mathematics Tests was based on the classical and IRT statistics of the test items. Selection was conducted by content experts from CTB/McGraw-Hill and NYSED and reviewed by psychometricians at CTB/McGraw-Hill and at NYSED. Two criteria governed the item selection process. The first of these was to meet the content specifications provided by NYSED. Second, within the limits set by these requirements, developers selected items with the best psychometric characteristics from the FT item pool.

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

The program has three parts. The first part of the program selects a working item pool of manageable size from the larger pool. The second part uses this selected item pool to perform the final test selection. The third part of the program includes a table showing the expected

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

CTB/McGraw-Hill editors traveled to Albany, New York, in September 2010, to finalize item selection and test creation with the NYSED staff (including content and research experts). NYSED 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 12 participants across all grade levels. After the approval by NYSED, the tests were produced and administered in May 2011.

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

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 examinations led to changes in the proficiency cut scores after the 2010 test administration. The results were reviewed by the NYS Technical Advisory Group (TAG) 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 and 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 NYSTP, the content is defined by detailed blueprints that describe New York State content standards and define the skills that must be measured to assess these content standards (see Table 2 in Section II). The test development process requires specific attention to content representation and the balance within each test form. New York State educators were involved in test constructions in various test development stages. For example, during the item review process, they reviewed FTs for their alignment with the test blueprint. Educators also participated in a process of establishing scoring rubrics (during rangefinding sessions) for CR items. Section II, "Test Design and Development," contains more information specific to the item review process. An independent study of alignment between the New York State curriculum and the New York State Grades 3–8 Mathematics Tests was conducted using Norman Webb's method. The results of the study provided additional evidence of test content validity (refer to *An External Alignment Study for New York State's Assessment Program*, April 2006, Educational Testing Services).

Construct (Internal Structure) Validity

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

Internal Consistency

Empirical studies of the internal structure of the test provide one type of evidence of construct validity. For example, high internal consistency constitutes evidence of validity. This is because high coefficients imply that the test questions are measuring the same domain of skill and are reliable and consistent. Reliability coefficients of the tests for total populations and subgroups of students are presented in Section VII, "Reliability and Standard Error of Measurement." For the total populations, the reliability coefficients (Cronbach's alpha) ranged 0.90–0.94, and for all subgroups, the reliability coefficients are greater than 0.80. Overall, high internal consistency of the NYSTP 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 Q_1 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 4, 5, and 7 displayed good item-model fit. One item in Grade 3, two items in Grade 6, and one item 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 3.

Table 3. Factor Analysis Results for Mathematics Tests (Total Population)

Grade	Initial Eigenvalues			
	Component	Total	% of Variance	Cumulative %
3	**1**	**9.69**	**21.07**	**21.07**
	2	1.98	4.30	25.37
	3	1.15	2.49	27.86
	4	1.04	2.27	30.13
	5	1.03	2.24	32.37
4	**1**	**13.45**	**23.59**	**23.59**
	2	1.97	3.45	27.04
	3	1.23	2.15	29.19
	4	1.16	2.04	31.23
	5	1.03	1.80	33.03
	6	1.01	1.77	34.80
5	**1**	**10.49**	**21.41**	**21.41**
	2	1.58	3.23	24.63
	3	1.13	2.31	26.94
	4	1.07	2.18	29.13
	5	1.02	2.07	31.20
6	**1**	**12.09**	**24.17**	**24.17**
	2	1.92	3.85	28.02
	3	1.42	2.84	30.86
	4	1.12	2.23	33.09
	5	1.02	2.04	35.13
7	**1**	**11.33**	**21.79**	**21.79**
	2	1.74	3.35	25.13
	3	1.36	2.61	27.75
	4	1.06	2.04	29.79

(Continued on next page)

Table 3. Factor Analysis Results for Mathematics Tests (Total Population) (cont.)

Grade	Initial Eigenvalues			
	Component	Total	% of Variance	Cumulative %
8	**1**	**13.31**	**24.65**	**24.65**
	2	1.45	2.69	27.35
	3	1.30	2.40	29.75
	4	1.06	1.96	31.71

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 statuses are true and who use one or more ELL-related accommodations. 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 and Mille, 1979; Jensen, 1980). Thus, empirical studies were conducted.

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

Section IV: Test Administration and Scoring

Listed in this section are brief summaries of New York State test administration and scoring procedures. For further information, refer to the *New York State Scoring Leader Handbooks* and *School Administrator's Manual* (SAM). In addition, please refer to Scoring Site Operations Manual (2011) located at http://www.p12.nysed.gov/apda/ei/ssom/ssom-11.pdf.

Test Administration

NYSTP Grades 3–8 Mathematics Tests were administered at the classroom level, during May 2011. The testing window for Grades 3–8 (including the makeup test administration) was May 11–18, 2011. The makeup test administration window allowed students who were ill or otherwise unable to test during the assigned window to take the test.

Scoring Procedures of Operational Tests

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

Scoring Models

For the 2010−11 school year, schools and school districts used local decision-making processes to select the model that best met their needs for the scoring of the Grades 3−8 Mathematics Tests. Schools were able to score these tests regionally, districtwide, or individually. Schools were required to enter one of the following scoring model codes on student answer sheets:

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

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

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

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

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

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

For further information, refer to the following link for a brief comparison between regional, district, and local scoring: http://www.emsc.nysed.gov/3-8/update-rev-dec05.htm (see Attachment C).

Scoring of Constructed-Response Items

The scoring of CR items was based primarily on the scoring guides, which were created by CTB/McGraw-Hill handscoring and content development specialists during rangefinding sessions. In 2011, the CTB/McGraw-Hill mathematics handscoring team was composed of six team leaders, each representing one grade. Team Leaders were selected on the basis of their handscoring experiences along with their educational and professional backgrounds.

Sets of actual FT student responses were reviewed and discussed openly, and consensus scores were agreed upon. Scoring guides were developed based on rangefinding decisions. 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 the 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 the audio files that highlighted important elements of the scoring guides. Each test book was scored by three separate scoring committee members, who scored three distinct sections of the test book. After test books were completed, the table facilitator or 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 paper 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 or fax. 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 2011. These data were used for all data analyses. Phase 2 involved submitting "straggler files" to CTB/McGraw-Hill in late June 2011. 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. A list of the data cleaning procedures conducted by research and accompanying case counts is presented in Tables 4a–4f.

Table 4a. NYSTP Mathematics Data Cleaning, Grade 3

Exclusion Rule	# Deleted	# Cases Remain
Initial N		197405
Out of grade	87	197318
No grade	0	197318
Duplicate record	0	197318
Non-public schools	2586	194732
Less than 5 items attempted	8	194724
Out-of-range CR scores	0	194724

Table 4b. NYSTP Mathematics Data Cleaning, Grade 4

Exclusion Rule	# Deleted	# Cases Remain
Initial N		200316
Out of grade	55	200261
No grade	0	200261
Duplicate record	0	200261
Non-public schools	8294	191967
Less than 5 items attempted	7	191960
Out-of-range CR scores	0	191960

Table 4c. NYSTP Mathematics Data Cleaning, Grade 5

Exclusion Rule	# Deleted	# Cases Remain
Initial N		197324
Out of grade	54	197270
No grade	0	197270
Duplicate record	0	197270
Non-public schools	1956	195314
Less than 5 items attempted	4	195310
Out-of-range CR scores	0	195310

Table 4d. NYSTP Mathematics Data Cleaning, Grade 6

Exclusion Rule	# Deleted	# Cases Remain
Initial N		199846
Out of grade	137	199709
No grade	0	199709
Duplicate record	0	199709
Non-public schools	8149	191560
Less than 5 items attempted	5	191555
Out-of-range CR scores	0	191555

Table 4e. NYSTP Mathematics Data Cleaning, Grade 7

Exclusion Rule	# Deleted	# Cases Remain
Initial N		196977
Out of grade	78	196899
No grade	0	196899
Duplicate record	0	196899
Non-public schools	2240	194659
Less than 5 items attempted	6	194653
Out-of-range CR scores	0	194653

Table 4f. NYSTP Mathematics Data Cleaning, Grade 8

Exclusion Rule	# Deleted	# Cases Remain
Initial N		202533
Out of grade	92	202441
No grade	0	202441
Duplicate record	0	202441
Non-public schools	9891	192550
Less than 5 items attempted	14	192536
Out-of-range CR scores	0	192536

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 5a. Grade 3 Sample Characteristics (N = 194724)

Demographic Category		N-count	% of Total N-count
NRC	NYC	72437	37.24
	Big 4 Cities	8243	4.24
	Urban/Suburban	15207	7.82
	Rural	10385	5.34
	Average needs	55778	28.67
	Low needs	27638	14.21
	Charter	4840	2.49
Ethnicity	Asian	16368	8.41
	Black	36220	18.60
	Hispanic	46220	23.74
	American Indian	1091	0.56
	Multi-Racial	1542	0.79
	Unknown	284	0.15
	White	92999	47.76
ELL	No	178378	91.61
	Yes	16346	8.39
SWD	No	167154	85.84
	Yes	27570	14.16
SUA	No	145430	74.69
	Yes	49294	25.31

Table 5b. Grade 4 Sample Characteristics (N = 191960)

Demographic Category		N-count	% of Total N-count
NRC	NYC	71650	37.37
	Big 4 Cities	8404	4.38
	Urban/Suburban	14714	7.67
	Rural	9576	4.99
	Average needs	55818	29.11
	Low needs	27811	14.51
	Charter	3754	1.96
Ethnicity	Asian	15913	8.29
	Black	36533	19.03
	Hispanic	44974	23.43
	American Indian	852	0.44
	Multi-Racial	1321	0.69
	Unknown	268	0.14
	White	92099	47.98
ELL	No	177340	92.38
	Yes	14620	7.62

(Continued on next page)

Table 5b. Grade 4 Sample Characteristics (N = 191960) (cont.)

Demographic Category		N-count	% of Total N-count
SWD	No	163093	84.96
	Yes	28867	15.04
SUA	No	142341	74.15
	Yes	49619	25.85

Table 5c. Grade 5 Sample Characteristics (N = 195310)

Demographic Category		N-count	% of Total N-count
NRC	NYC	70933	36.36
	Big 4 Cities	8022	4.11
	Urban/Suburban	14683	7.53
	Rural	9816	5.03
	Average needs	56691	29.06
	Low needs	29764	15.26
	Charter	5168	2.65
Ethnicity	Asian	17042	8.73
	Black	37070	18.98
	Hispanic	44680	22.88
	American Indian	878	0.45
	Multi-Racial	1253	0.64
	Unknown	256	0.13
	White	94131	48.20
ELL	No	182950	93.67
	Yes	12360	6.33
SWD	No	165679	84.83
	Yes	29631	15.17
SUA	No	146095	74.80
	Yes	49215	25.20

Table 5d. Grade 6 Sample Characteristics (N = 191555)

Demographic Category		N-count	% of Total N-count
NRC	NYC	69122	36.14
	Big 4 Cities	7823	4.09
	Urban/Suburban	14010	7.32
	Rural	10627	5.56
	Average needs	55961	29.26
	Low needs	29090	15.21
	Charter	4644	2.43

(Continued on next page)

Table 5d. Grade 6 Sample Characteristics (N = 191555) (cont.)

Demographic Category		N-count	% of Total N-count
Ethnicity	Asian	15478	8.08
	Black	37049	19.34
	Hispanic	43731	22.83
	American Indian	910	0.48
	Multi-Racial	1194	0.62
	White	266	0.14
	Unknown	92927	48.51
ELL	No	181393	94.69
	Yes	10162	5.31
SWD	No	162533	84.85
	Yes	29022	15.15
SUA	No	146968	76.72
	Yes	44587	23.28

Table 5e. Grade 7 Sample Characteristics (N = 194653)

Demographic Category		N-count	% of Total N-count
NRC	NYC	70120	36.09
	Big 4 Cities	7611	3.92
	Urban/Suburban	14013	7.21
	Rural	10412	5.36
	Average needs	57351	29.52
	Low needs	31260	16.09
	Charter	3502	1.80
Ethnicity	Asian	15667	8.05
	Black	37361	19.19
	Hispanic	43486	22.34
	American Indian	902	0.46
	Multi-Racial	1054	0.54
	Unknown	283	0.15
	White	95900	49.27
ELL	No	185649	95.37
	Yes	9004	4.63
SWD	No	165155	84.85
	Yes	29498	15.15
SUA	No	151256	77.71
	Yes	43397	22.29

Table 5f. Grade 8 Sample Characteristics (N = 192536)

Demographic Category		N-count	% of Total N-count
NRC	NYC	71541	37.25
	Big 4 Cities	7627	3.97
	Urban/Suburban	13068	6.80
	Rural	10190	5.31
	Average needs	55884	29.10
	Low needs	31072	16.18
	Charter	2680	1.40
Ethnicity	Asian	16145	8.39
	Black	37092	19.26
	Hispanic	42627	22.14
	American Indian	848	0.44
	Multi-Racial	917	0.48
	Unknown	258	0.13
	White	94649	49.16
ELL	No	183830	95.48
	Yes	8706	4.52
SWD	No	164265	85.32
	Yes	28271	14.68
SUA	No	150892	78.37
	Yes	41644	21.63

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 VII, "Reliability and Standard Error of Measurement").

Item Suppression

One item in Grade 7 test was suppressed from scoring. It was found that only 11% of students responded correctly to item 26 on the Grade 7 test. This item also showed negative point biserial correlation for the correct response (-0.15) and positive correlation for a distractor (0.22) indicating that higher ability students tended to select the distractor rather than the correct response option. This item was reviewed by CTB's Mathematics content

expert who confirmed that the item key was correct. Nevertheless, the NYSED instructed CTB to suppress it. The suppression did not affect the test blueprint.

Item Difficulty and Response Distribution

Item difficulty and response distribution tables (Tables 6a–6f) 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.

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 0.22 to 0.97. For Grade 3, the item p-values were between 0.22 and 0.97 with a mean of 0.72. For Grade 4, the item p-values were between 0.23 and 0.96 with a mean of 0.70. For Grade 5, the item p-values were between 0.37 and 0.96 with a mean of 0.67. For Grade 6, the item p-values were between 0.24 and 0.95 with a mean of 0.65. For Grade 7, the item p-values were between 0.31 and 0.98 with a mean of 0.63. For Grade 8, the item p-values were between 0.33 and 0.82 with a mean of 0.59. These statistics are provided in Tables 6a–6f, along with other classical test summary statistics.

Table 6a. Item Analysis, Grade 3

Item	Item Type	N-count	P-value	% Omit	Pbis Key
1	MC	194580	0.82	0.02	0.25
2	MC	194553	0.91	0.05	0.35
3	MC	194481	0.84	0.08	0.45
4	MC	194276	0.69	0.19	0.44
5	MC	194462	0.94	0.07	0.32
6	MC	194380	0.80	0.11	0.48
7	MC	194459	0.94	0.07	0.33
8	MC	194480	0.87	0.06	0.43
9	MC	194371	0.35	0.13	0.30
10	MC	194516	0.92	0.07	0.39
11	MC	194530	0.78	0.07	0.41
12	MC	194546	0.94	0.06	0.29
13	MC	194507	0.91	0.07	0.41
14	MC	194392	0.92	0.09	0.36
15	MC	194344	0.53	0.14	0.38
16	MC	194415	0.43	0.11	0.36
17	MC	194470	0.87	0.09	0.40
18	MC	194256	0.73	0.17	0.49
19	MC	194381	0.66	0.11	0.27
20	MC	194495	0.93	0.09	0.33

(Continued on next page)

Table 6a. Item Analysis, Grade 3 (cont.)

Item	Item Type	N-count	P-value	% Omit	Pbis Key
21	MC	194513	0.72	0.07	0.45
22	MC	194448	0.60	0.10	0.55
23	MC	194376	0.82	0.14	0.51
24	MC	194037	0.63	0.29	0.46
25	MC	194141	0.56	0.20	0.39
26	MC	193915	0.78	0.31	0.39
27	MC	194412	0.97	0.13	0.28
28	MC	194323	0.89	0.15	0.41
29	MC	194280	0.56	0.17	0.45
30	MC	194265	0.76	0.20	0.47
31	MC	194377	0.74	0.14	0.52
32	MC	194180	0.57	0.21	0.33
33	MC	193774	0.33	0.39	0.32
34	MC	194016	0.70	0.26	0.34
35	MC	193941	0.61	0.33	0.27
36	MC	193756	0.22	0.41	0.17
37	MC	193726	0.64	0.42	0.39
38	MC	193708	0.82	0.48	0.36
39	MC	193453	0.75	0.61	0.49
40	MC	192602	0.41	1.06	0.25
41	CR	194596	0.75	0.07	
42	CR	194614	0.81	0.06	
43	CR	194469	0.62	0.13	
44	CR	194436	0.49	0.15	
45	CR	194508	0.52	0.11	
46	CR	194557	0.87	0.09	

Table 6b. Item Analysis, Grade 4

Item	Item Type	N-count	P-value	% Omit	Pbis Key
1	MC	191897	0.92	0.02	0.35
2	MC	191858	0.76	0.03	0.40
3	MC	191811	0.96	0.03	0.32
4	MC	191798	0.92	0.04	0.32
5	MC	191723	0.81	0.09	0.55
6	MC	191790	0.77	0.05	0.38
7	MC	191656	0.67	0.12	0.51
8	MC	191812	0.96	0.04	0.27
9	MC	191753	0.81	0.08	0.57
10	MC	191711	0.72	0.09	0.52
11	MC	191685	0.75	0.08	0.49
12	MC	191803	0.93	0.06	0.35
13	MC	191829	0.88	0.05	0.48

(Continued on next page)

Table 6b. Item Analysis, Grade 4 (cont.)

Item	Item Type	N-count	P-value	% Omit	Pbis Key
14	MC	191774	0.70	0.05	0.48
15	MC	191740	0.73	0.09	0.43
16	MC	191681	0.53	0.12	0.36
17	MC	191837	0.95	0.05	0.16
18	MC	191746	0.68	0.06	0.40
19	MC	191598	0.36	0.16	0.42
20	MC	191712	0.33	0.08	0.37
21	MC	191779	0.93	0.07	0.27
22	MC	191551	0.59	0.15	0.52
23	MC	191620	0.96	0.09	0.26
24	MC	191740	0.90	0.08	0.32
25	MC	191768	0.57	0.07	0.48
26	MC	191464	0.65	0.21	0.52
27	MC	191700	0.74	0.10	0.57
28	MC	191754	0.90	0.09	0.30
29	MC	191699	0.40	0.11	0.36
30	MC	191699	0.70	0.10	0.41
31	MC	191724	0.71	0.10	0.43
32	MC	191611	0.62	0.13	0.51
33	MC	191611	0.66	0.12	0.36
34	MC	191574	0.60	0.14	0.42
35	MC	191457	0.54	0.21	0.42
36	MC	191487	0.66	0.21	0.51
37	MC	191489	0.69	0.22	0.55
38	MC	191544	0.89	0.17	0.38
39	MC	191384	0.23	0.22	0.30
40	MC	191426	0.60	0.24	0.27
41	MC	191290	0.76	0.27	0.37
42	MC	191332	0.62	0.28	0.51
43	MC	191306	0.50	0.30	0.40
44	MC	190881	0.62	0.53	0.41
45	MC	190744	0.66	0.61	0.43
46	CR	191827	0.79	0.07	
47	CR	191778	0.69	0.09	
48	CR	191831	0.84	0.07	
49	CR	191574	0.70	0.20	
50	CR	191628	0.66	0.17	
51	CR	191302	0.47	0.34	
52	CR	191555	0.73	0.21	
53	CR	191599	0.61	0.19	
54	CR	191334	0.33	0.33	
55	CR	191547	0.77	0.22	
56	CR	191196	0.72	0.40	
57	CR	190584	0.76	0.72	

Table 6c. Item Analysis, Grade 5

Item	Item Type	N-count	P-value	% Omit	Pbis Key
1	MC	195134	0.82	0.08	0.39
2	MC	195151	0.61	0.05	0.50
3	MC	194829	0.37	0.21	0.07
4	MC	195071	0.74	0.08	0.40
5	MC	194865	0.78	0.21	0.53
6	MC	195151	0.96	0.06	0.23
7	MC	194977	0.58	0.12	0.38
8	MC	195129	0.63	0.05	0.30
9	MC	195026	0.37	0.11	0.35
10	MC	195025	0.64	0.13	0.52
11	MC	195125	0.77	0.08	0.54
12	MC	194973	0.59	0.14	0.45
13	MC	195132	0.73	0.07	0.46
14	MC	194760	0.41	0.25	0.10
15	MC	194960	0.76	0.14	0.48
16	MC	195070	0.82	0.08	0.37
17	MC	194872	0.64	0.19	0.38
18	MC	195055	0.93	0.08	0.33
19	MC	194886	0.59	0.19	0.49
20	MC	195106	0.87	0.09	0.40
21	MC	195155	0.94	0.05	0.21
22	MC	195034	0.79	0.10	0.51
23	MC	194734	0.43	0.26	0.40
24	MC	195025	0.85	0.12	0.42
25	MC	194947	0.75	0.15	0.54
26	MC	194941	0.78	0.16	0.35
27	MC	194550	0.52	0.34	0.39
28	MC	194899	0.83	0.17	0.50
29	MC	194777	0.55	0.24	0.27
30	MC	194798	0.86	0.24	0.42
31	MC	194680	0.72	0.29	0.41
32	MC	194719	0.68	0.26	0.33
33	MC	194425	0.51	0.40	0.44
34	MC	194468	0.43	0.39	0.35
35	MC	194419	0.57	0.41	0.34
36	MC	194264	0.69	0.50	0.37
37	MC	194045	0.61	0.61	0.39
38	MC	193863	0.40	0.70	0.34
39	MC	193761	0.70	0.75	0.49
40	MC	193659	0.92	0.82	0.33
41	MC	193312	0.85	1.01	0.38
42	CR	194979	0.59	0.17	

(Continued on next page)

Table 6c. Item Analysis, Grade 5 (cont.)

Item	Item Type	N-count	P-value	% Omit	Pbis Key
43	CR	194713	0.59	0.31	
44	CR	195059	0.56	0.13	
45	CR	195140	0.61	0.09	
46	CR	194450	0.56	0.44	
47	CR	195070	0.69	0.12	
48	CR	194878	0.52	0.22	
49	CR	194985	0.66	0.17	

Table 6d. Item Analysis, Grade 6

Item	Item Type	N-count	P-value	% Omit	Pbis Key
1	MC	191502	0.77	0.02	0.34
2	MC	191459	0.92	0.02	0.35
3	MC	191298	0.67	0.12	0.40
4	MC	191487	0.95	0.02	0.29
5	MC	191355	0.69	0.08	0.51
6	MC	191256	0.52	0.14	0.57
7	MC	191412	0.74	0.06	0.48
8	MC	191232	0.57	0.15	0.51
9	MC	191402	0.91	0.06	0.18
10	MC	191361	0.73	0.08	0.37
11	MC	191284	0.63	0.12	0.41
12	MC	191444	0.93	0.03	0.28
13	MC	191424	0.75	0.05	0.50
14	MC	191276	0.75	0.12	0.49
15	MC	191370	0.63	0.08	0.43
16	MC	191387	0.38	0.07	0.44
17	MC	191347	0.50	0.08	0.56
18	MC	191367	0.62	0.07	0.44
19	MC	191380	0.57	0.07	0.31
20	MC	191363	0.72	0.08	0.56
21	MC	191326	0.52	0.10	0.43
22	MC	191442	0.89	0.04	0.36
23	MC	191098	0.51	0.20	0.33
24	MC	191265	0.39	0.12	0.42
25	MC	191191	0.55	0.16	0.54
26	MC	191214	0.69	0.15	0.52
27	MC	191231	0.73	0.13	0.46
28	MC	191313	0.94	0.11	0.31

(Continued on next page)

Table 6d. Item Analysis, Grade 6 (cont.)

Item	Item Type	N-count	P-value	% Omit	Pbis Key
29	MC	191112	0.72	0.21	0.52
30	MC	191288	0.72	0.12	0.33
31	MC	191304	0.48	0.11	0.42
32	MC	191277	0.67	0.11	0.60
33	MC	191052	0.59	0.23	0.53
34	MC	191070	0.74	0.22	0.50
35	MC	191156	0.88	0.19	0.42
36	MC	190854	0.33	0.33	0.16
37	MC	190970	0.76	0.26	0.32
38	MC	190583	0.45	0.48	0.05
39	MC	190707	0.70	0.42	0.54
40	MC	190255	0.24	0.65	0.37
41	CR	191303	0.71	0.13	
42	CR	191204	0.68	0.18	
43	CR	191328	0.83	0.12	
44	CR	189703	0.58	0.97	
45	CR	191227	0.72	0.17	
46	CR	190597	0.30	0.50	
47	CR	190753	0.62	0.42	
48	CR	191305	0.80	0.13	
49	CR	191209	0.50	0.18	
50	CR	191002	0.48	0.29	

Table 6e. Item Analysis, Grade 7

Item	Item Type	N-count	P-value	% Omit	Pbis Key
1	MC	194567	0.96	0.04	0.23
2	MC	194541	0.73	0.05	0.47
3	MC	194346	0.31	0.14	0.43
4	MC	194548	0.87	0.04	0.34
5	MC	194433	0.56	0.08	0.17
6	MC	194393	0.64	0.11	0.51
7	MC	194278	0.41	0.18	0.37
8	MC	194435	0.40	0.10	0.23
9	MC	194495	0.88	0.06	0.33
10	MC	194459	0.80	0.08	0.30
11	MC	194497	0.73	0.06	0.52
12	MC	194485	0.74	0.07	0.16
13	MC	193598	0.36	0.53	0.12

(Continued on next page)

Table 6e. Item Analysis, Grade 7 (cont.)

Item	Item Type	N-count	P-value	% Omit	Pbis Key
14	MC	194514	0.74	0.06	0.34
15	MC	194420	0.63	0.10	0.41
16	MC	194336	0.49	0.14	0.41
17	MC	194467	0.77	0.08	0.50
18	MC	194392	0.65	0.11	0.37
19	MC	194453	0.61	0.09	0.49
20	MC	194279	0.74	0.17	0.39
21	MC	194385	0.51	0.12	0.39
22	MC	194525	0.98	0.05	0.16
23	MC	194392	0.74	0.12	0.48
24	MC	194394	0.84	0.11	0.40
25	MC	194333	0.72	0.15	0.41
26	MC	194299	0.11	0.16	-0.15
27	MC	194083	0.46	0.28	0.49
28	MC	194429	0.72	0.10	0.46
29	MC	194396	0.80	0.10	0.45
30	MC	194383	0.41	0.12	0.51
31	MC	194350	0.58	0.13	0.44
32	MC	194427	0.82	0.10	0.44
33	MC	194121	0.37	0.24	0.36
34	MC	194181	0.58	0.22	0.45
35	MC	194290	0.68	0.16	0.32
36	MC	194258	0.72	0.18	0.50
37	MC	194282	0.58	0.16	0.32
38	MC	194229	0.44	0.20	0.45
39	MC	194189	0.69	0.21	0.44
40	MC	194001	0.57	0.31	0.45
41	MC	194161	0.76	0.23	0.46
42	MC	194073	0.63	0.26	0.31
43	MC	194060	0.58	0.28	0.40
44	MC	194044	0.63	0.29	0.43
45	MC	193848	0.38	0.40	0.48
46	CR	193807	0.55	0.43	
47	CR	191151	0.51	1.80	
48	CR	193752	0.55	0.46	
49	CR	188489	0.50	3.17	
50	CR	194051	0.76	0.31	
51	CR	193723	0.72	0.48	
52	CR	192327	0.46	1.19	
53	CR	190620	0.57	2.07	

Table 6f. Item Analysis, Grade 8

Item	Item Type	N-count	P-value	% Omit	Pbis Key
1	MC	192179	0.39	0.17	0.49
2	MC	192306	0.77	0.10	0.37
3	MC	191704	0.36	0.42	0.44
4	MC	192300	0.70	0.11	0.44
5	MC	192342	0.73	0.09	0.49
6	MC	192423	0.65	0.04	0.52
7	MC	192203	0.64	0.16	0.30
8	MC	192376	0.69	0.07	0.52
9	MC	192366	0.69	0.07	0.48
10	MC	192264	0.78	0.12	0.41
11	MC	192160	0.69	0.18	0.45
12	MC	192346	0.68	0.09	0.46
13	MC	192052	0.51	0.24	0.51
14	MC	192256	0.60	0.12	0.41
15	MC	192378	0.62	0.07	0.30
16	MC	192405	0.82	0.06	0.28
17	MC	192118	0.47	0.20	0.29
18	MC	192146	0.58	0.19	0.51
19	MC	192033	0.51	0.24	0.33
20	MC	192379	0.82	0.06	0.32
21	MC	192243	0.39	0.13	0.29
22	MC	191998	0.36	0.26	0.43
23	MC	192349	0.75	0.08	0.44
24	MC	191844	0.58	0.34	0.44
25	MC	192021	0.66	0.25	0.49
26	MC	192258	0.78	0.12	0.42
27	MC	191753	0.50	0.38	0.44
28	MC	192049	0.66	0.23	0.49
29	MC	191666	0.46	0.43	0.41
30	MC	192166	0.71	0.17	0.44
31	MC	191764	0.61	0.37	0.43
32	MC	191642	0.45	0.44	0.29
33	MC	191975	0.63	0.27	0.39
34	MC	191981	0.79	0.27	0.24
35	MC	191741	0.33	0.39	0.21
36	MC	191854	0.70	0.32	0.38
37	MC	191772	0.59	0.37	0.53
38	MC	191443	0.56	0.54	0.52
39	MC	191413	0.52	0.56	0.36
40	MC	191635	0.75	0.45	0.54
41	MC	191501	0.80	0.51	0.39
42	MC	191336	0.70	0.61	0.52

(Continued on next page)

Table 6f. Item Analysis, Grade 8 (cont.)

Item	Item Type	N-count	P-value	% Omit	Pbis Key
43	CR	191148	0.47	0.72	
44	CR	191789	0.52	0.39	
45	CR	191122	0.49	0.73	
46	CR	189392	0.52	1.63	
47	CR	190943	0.50	0.83	
48	CR	191565	0.52	0.50	
49	CR	189672	0.47	1.49	
50	CR	189760	0.42	1.44	
51	CR	191339	0.42	0.62	
52	CR	190596	0.50	1.01	
53	CR	189820	0.43	1.41	
54	CR	189693	0.57	1.48	

Point-Biserial Correlation Coefficients

Point biserial statistics are used to examine item-test correlations, or item discrimination. As shown in Tables 6a–6f, point biserial correlation coefficients were computed for the correct answers of MC items. 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. Point biserials for correct answer options on the Mathematics Tests ranged 0.17–0.60. For Grade 3, the point biserials were between 0.17 and 0.55. For Grade 4, the point biserials were between 0.16 and 0.57. For Grade 5, the point biserials were between 0.07 and 0.54. For Grade 6, the point biserials were between 0.05 and 0.60. For Grade 7, the point biserials were between 0.12 and 0.52. For Grade 8, the point biserials were between 0.21 and 0.54.

Test Statistics and Reliability Coefficients

Test statistics, including raw-score mean and standard deviation, are presented in Table 7. Reliability coefficients provide measures of internal consistency that range from zero to one. Two reliability coefficients, Cronbach's alpha and Feldt-Raju, were computed for the Grades 3–8 Mathematics Tests. Both types of reliability estimates are appropriate to use when a test contains both MC and CR items. Calculated Cronbach's alpha reliabilities ranged 0.91–0.94. Feldt-Raju reliability coefficients ranged 0.91–0.94. The lowest reliability was observed for Grade 3; however, as that test had the lowest number of score points, it was reasonable that its reliability would not be as high as the other grade-level tests. The highest reliability was observed for Grades 4 and 8. All reliabilities exceeded 0.90 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 VII, "Reliability and Standard Error of Measurement.")

Table 7. NYSTP Mathematics 2011 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	54	38.30	9.68	0.71	0.22	0.96	0.91	0.91
4	73	50.46	14.40	0.70	0.23	0.96	0.94	0.94
5	61	39.87	11.48	0.67	0.37	0.96	0.91	0.92
6	64	41.25	13.51	0.65	0.24	0.95	0.93	0.94
7	64	39.82	13.26	0.62	0.11	0.98	0.92	0.93
8	70	39.36	15.89	0.59	0.33	0.82	0.94	0.94

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 6a–6f 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, and Bleistein, 1992) compares the mean scores of reference and focal groups, after adjusting for ability differences. A moderate amount of significant DIF, for or against the focal group, is represented by an SMD with an absolute value between 0.10 and 0.19, inclusive. A large amount of practically significant DIF is represented by an SMD with an absolute value of 0.20 or greater. Then, the Mantel-Haenszel method is employed to compute DIF statistics for MC items. This non-parametric DIF method partitions the sample of examinees into categories based on total raw test scores. It then compares the log-odds ratio of keyed responses for the focal and reference groups. The Mantel-Haenszel method has a critical value of 6.63 (degrees of freedom = 1 for MC items; alpha = 0.01) and is compared to its corresponding delta-value (significant when absolute value of delta > 1.50) to factor in effect size (Zwick, Donoghue, and Grima, 1993). It is important to recognize that the two methods differ in assumptions and computation; therefore, the results from both methods may not be in agreement. It should be noted that two methods of classical DIF computation and one method of IRT DIF computation (described in Section VI) were employed because no single method can identify all DIF items on a test (Hambleton, Clauser, Mazer, and Jones, 1993).

Classical DIF analyses were conducted on subgroups of 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 ELL (focal group: ELL; reference group: Non-ELL). All cases in clean data sets were used to compute DIF statistics. Table 8 shows the number of students in each focal and reference group.

Table 8. NYSTP Mathematics 2011 Classical DIF Sample N-Counts

Grade	Ethnicity				Gender		Needs Resource Category		Test Language	
	Black	Hispanic	Asian	White	Female	Male	High	Low	Spanish	English
3	36220	46220	16368	92999	95237	99487	106272	83416	3227	191497
4	36533	44974	15913	92099	93855	98105	104344	83629	2988	188972
5	37070	44680	17042	94131	95452	99858	103454	86455	3122	192188
6	37049	43731	15478	92927	93305	98250	101582	85051	3321	188234
7	37361	43486	15667	95900	95432	99221	102156	88611	3276	191377
8	37092	42627	16145	94649	94072	98464	102426	86956	3224	189312

Table 9 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 9. Number of Items Flagged by SMD and Mantel-Haenszel DIF Methods

Grade	Number of Flagged Items
3	2
4	9
5	3
6	5
7	14
8	10

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

Section VI: IRT Scaling and Equating

IRT Models and Rationale for Use

Item response theory (IRT) allows comparisons between items and examinees, even those from different test forms, by using a common scale for all items and examinees (i.e., as if there were a hypothetical test that contained items from all forms). The three-parameter logistic (3PL) model (Lord and Novick, 1968; Lord, 1980) was used to analyze item responses on the MC items. For analysis of the CR items, the two-parameter partial credit 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 NYSTP 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 and Novick, 1968; Lord, 1980) was used in the analysis of MC items. In this model, the probability that a student with ability θ responds correctly to item *i* is

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

where

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

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

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

where

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

and

k is the item response category ($k = 1, 2, \ldots. m_j$).

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 2011 OP tests to NYSTP OP scales as described in the "Scaling and Equating" subsection.

The sample representativeness of these FT anchor forms was 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. As shown in Tables 10 through 12, the 2011 OP test samples were comparable to 2010 populations in terms of NRC, student ethnicity, proportions of ELL, proportions of SWD, and proportions of SUA. In addition, as shown in Tables 13 through 15, the samples of students who responded to items in FT anchor forms were also representative of the state student population in each grade.

Table 10. Grades 3 and 4 Demographic Statistics

Demographics	2010 Grade 3 Population	2011 Grade 3 Sample	2010 Grade 4 Population	2011 Grade 4 Sample
	%	%	%	%
NRC SUBGROUPS				
NYC	36.82	37.24	36.54	37.37
Big 4 Cities	4.38	4.24	4.10	4.38
Urban/Suburban	8.09	7.82	8.09	7.67
Rural	5.78	5.34	5.75	4.99
Average needs	29.22	28.67	29.59	29.11
Low needs	13.59	14.21	14.18	14.51
Charter	2.12	2.49	1.75	1.96
ETHNICITY				
Asian	7.93	8.41	8.36	8.29
Black	18.97	18.60	18.98	19.03
Hispanics	22.57	23.74	21.84	23.43
American Indian	0.49	0.56	0.47	0.44
Multi-Racial	0.56	0.79	0.50	0.69
White	49.41	47.76	49.79	47.98
Unknown	0.06	0.15	0.06	0.14
ELL STATUS				
No	91.55	91.61	92.60	92.38
Yes	8.45	8.39	7.40	7.62
DISABILITY				
No	85.75	85.84	85.25	84.96
Yes	14.25	14.16	14.75	15.04
ACCOMMODATIONS				
No	75.15	74.69	75.00	74.15
Yes	24.85	25.31	25.00	25.85

Table 11. Grades 5 and 6 Demographic Statistics

Demographics	2010 Grade 5 Population	2011 Grade 5 Sample	2010 Grade 6 Population	2011 Grade 6 Sample
	%	%	%	%
NRC SUBGROUPS				
NYC	35.46	36.36	35.09	36.14
Big 4 Cities	4.07	4.11	3.90	4.09
Urban/Suburban	7.80	7.53	7.75	7.32
Rural	5.77	5.03	5.73	5.56
Average needs	29.94	29.06	30.73	29.26
Low needs	14.66	15.26	14.88	15.21
Charter	2.30	2.65	1.93	2.43
ETHNICITY				
Asian	7.92	8.73	7.84	8.08
Black	19.12	18.98	19.20	19.34
Hispanics	21.68	22.88	21.30	22.83
American Indian	0.47	0.45	0.49	0.48
Multi-Racial	0.44	0.64	0.39	0.62
White	50.33	48.20	50.74	48.51
Unknown	0.05	0.13	0.05	0.14
ELL STATUS				
No	94.07	93.67	95.05	94.69
Yes	5.93	6.33	4.95	5.31
DISABILITY				
No	84.78	84.83	84.69	84.85
Yes	15.22	15.17	15.31	15.15
ACCOMMODATIONS				
No	75.56	74.80	77.68	76.72
Yes	24.44	25.20	22.32	23.28

Table 12. Grades 7 and 8 Demographic Statistics

Demographics	2010 Grade 7 Population	2011 Grade 7 Sample	2010 Grade 8 Population	2011 Grade 8 Sample
	%	%	%	%
NRC SUBGROUPS				
NYC	35.34	36.09	35.95	37.25
Big 4 Cities	3.90	3.92	3.78	3.97
Urban/Suburban	7.66	7.21	7.48	6.80
Rural	5.79	5.36	5.80	5.31
Average needs	31.21	29.52	30.87	29.10
Low needs	14.65	16.09	14.95	16.18
Charter	1.46	1.80	1.16	1.40
ETHNICITY				
Asian	7.88	8.05	7.91	8.39
Black	19.18	19.19	18.80	19.26
Hispanics	20.99	22.34	20.96	22.14
American Indian	0.48	0.46	0.45	0.44
Multi-Racial	0.36	0.54	0.30	0.48
White	51.08	49.27	51.53	49.16
Unknown	0.04	0.15	0.05	0.13
ELL STATUS				
No	95.59	95.37	95.84	95.48
Yes	4.41	4.63	4.16	4.52
DISABILITY				
No	85.00	84.85	85.21	85.32
Yes	15.00	15.15	14.79	14.68
ACCOMMODATIONS				
No	78.57	77.71	78.90	78.37
Yes	21.43	22.29	21.10	21.63

The NRC distributions of the FT anchor form samples are compared with those of the OP samples in Tables 13 through 15. It is apparent that the FT anchor samples represent the OP student population well.

Table 13. Grades 3 and 4 Demographic Statistics for Field Test Anchor Forms

Demographics	2011 Grade 3 FT Anchor Form 1	2011 Grade 3 FT Anchor Form 2	2011 Grade 3 OP Sample	2011 Grade 4 FT Anchor Form 1	2011 Grade 4 FT Anchor Form 2	2011 Grade 4 OP Sample
	%	%	%	%	%	%
NRC SUBGROUPS						
NYC	39.88	34.31	37.24	40.79	33.14	37.37
Big 4 Cities	4.35	3.75	4.24	3.77	4.25	4.38
Urban/Suburban	10.64	7.47	7.82	11.31	8.24	7.67
Rural	4.04	5.59	5.34	3.02	5.20	4.99
Average needs	25.92	30.83	28.67	25.11	29.59	29.11
Low needs	13.72	16.05	14.21	14.39	17.55	14.51
Charter	1.45	2.00	2.49	1.62	2.03	1.96

Table 14. Grades 5 and 6 Demographic Statistics for Field Test Anchor Forms

Demographics	2011 Grade 5 FT Anchor Form 1	2011 Grade 5 FT Anchor Form 2	2011 Grade 5 OP Sample	2011 Grade 6 FT Anchor Form 1	2011 Grade 6 FT Anchor Form 2	2011 Grade 6 OP Sample
	%	%	%	%	%	%
NRC SUBGROUPS						
NYC	38.34	32.94	36.36	38.46	33.02	36.14
Big 4 Cities	3.73	3.55	4.11	3.35	3.81	4.09
Urban/Suburban	7.62	8.71	7.53	9.75	7.38	7.32
Rural	4.18	4.98	5.03	5.76	5.76	5.56
Average needs	28.65	31.12	29.06	27.96	32.49	29.26
Low needs	15.75	16.30	15.26	13.40	15.93	15.21
Charter	1.74	2.38	2.65	1.32	1.61	2.43

Table 15. Grades 7 and 8 Demographic Statistics for Field Test Anchor Forms

Demographics	2011 Grade 7 OP Anchor Form 1	2011 Grade 7 OP Anchor Form 2	2011 Grade 7 OP Sample	2011 Grade 8 FT Anchor Form 1	2011 Grade 8 FT Anchor Form 2	2011 Grade 8 OP Sample
	%	%	%	%	%	%
NRC SUBGROUPS						
NYC	36.19	32.71	36.09	38.59	33.65	37.25
Big 4 Cities	4.03	3.41	3.92	3.62	3.15	3.97
Urban/Suburban	7.21	7.69	7.21	8.63	5.96	6.80
Rural	5.39	5.81	5.36	5.46	5.89	5.31
Average needs	29.24	32.24	29.52	27.18	31.32	29.10
Low needs	16.11	16.74	16.09	15.58	18.51	16.18
Charter	1.82	1.41	1.80	0.94	1.51	1.40

Calibration Process

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

The NYSTP 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 NYSTP 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 16 presents a summary of calibration results. For the Grades 3–8 Mathematics Tests, all the calibration estimation results are reasonable.

Table 16. NYSTP Mathematics 2011 Calibration Results

Grade	Largest *a*-parameter	Lowest and highest *b*-parameters		# Items with Default *c*-parameters	Theta Mean	Theta Standard Deviation	# Students
3	2.146	-3.377	1.692	13	-0.10	1.186	194724
4	2.576	-7.283	1.370	18	0.00	1.100	191960
5	2.839	-3.840	3.840	10	-0.01	1.119	195310
6	3.165	-3.332	2.097	7	0.00	1.133	191555
7	2.576	-3.718	2.015	9	-0.03	1.170	194653
8	3.067	-2.004	1.478	8	-0.08	1.178	192536

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

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

The value of Z will increase with sample size, when all else is 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 2011 Grades 3–8 Mathematics Tests demonstrated a good model fit, further supporting use of the chosen models. No items in Grades 4, 5, 7, and 8 exhibited poor item-model fit statistics. The following items exhibited misfit: Grade 3 item 17, Grade 6 items 36 and 44, and Grade 8 item 53. 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 17–22.

Table 17. Mathematics Grade 3 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	Z_{Q1}	Z_{Q1} critical	Fit OK?
1	3PL	243.04	7	193605	63.08	516.28	Y
2	3PL	103.10	7	193605	25.68	516.28	Y
3	3PL	94.94	7	193605	23.50	516.28	Y
4	3PL	12.90	7	193605	1.58	516.28	Y
5	3PL	53.15	7	193605	12.33	516.28	Y
6	3PL	118.22	7	193605	29.72	516.28	Y
7	3PL	153.31	7	193605	39.10	516.28	Y
8	3PL	62.15	7	193605	14.74	516.28	Y
9	3PL	708.33	7	193605	187.44	516.28	Y
10	3PL	93.25	7	193605	23.05	516.28	Y
11	3PL	54.25	7	193605	12.63	516.28	Y
12	3PL	79.25	7	193605	19.31	516.28	Y
13	3PL	57.81	7	193605	13.58	516.28	Y
14	3PL	123.55	7	193605	31.15	516.28	Y
15	3PL	307.24	7	193605	80.24	516.28	Y
16	3PL	116.66	7	193605	29.31	516.28	Y
17	3PL	1948.63	7	193605	518.92	516.28	N
18	3PL	79.87	7	193605	19.48	516.28	Y
19	3PL	30.16	7	193605	6.19	516.28	Y
20	3PL	169.74	7	193605	43.50	516.28	Y
21	3PL	28.69	7	193605	5.80	516.28	Y
22	3PL	328.67	7	193605	85.97	516.28	Y
23	3PL	80.60	7	193605	19.67	516.28	Y
24	3PL	134.19	7	193605	33.99	516.28	Y
25	3PL	117.80	7	193605	29.61	516.28	Y

(Continued on next page)

Table 17. Mathematics Grade 3 Item Fit Statistics (cont.)

Item	Model	Chi Square	DF	Total N	Z_{Q1}	Z_{Q1} critical	Fit OK?
26	3PL	53.23	7	193605	12.35	516.28	Y
27	3PL	86.30	7	193605	21.19	516.28	Y
28	3PL	78.16	7	193605	19.02	516.28	Y
29	3PL	187.92	7	193605	48.35	516.28	Y
30	3PL	621.02	7	193605	164.10	516.28	Y
31	3PL	115.08	7	193605	28.89	516.28	Y
32	3PL	40.97	7	193605	9.08	516.28	Y
33	3PL	212.34	7	193605	54.88	516.28	Y
34	3PL	14.20	7	193605	1.92	516.28	Y
35	3PL	279.38	7	193605	72.80	516.28	Y
36	3PL	753.84	7	193605	199.60	516.28	Y
37	3PL	31.35	7	193605	6.51	516.28	Y
38	3PL	267.83	7	193605	69.71	516.28	Y
39	3PL	55.12	7	193605	12.86	516.28	Y
40	3PL	164.63	7	193605	42.13	516.28	Y
41	2PPC	333.43	17	193605	54.27	516.28	Y
42	2PPC	1193.38	17	193605	201.75	516.28	Y
43	2PPC	774.96	17	193605	129.99	516.28	Y
44	2PPC	1904.15	17	193605	323.64	516.28	Y
45	2PPC	2753.79	26	193605	378.28	516.28	Y
46	2PPC	873.55	26	193605	117.53	516.28	Y

Table 18. Mathematics Grade 4 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	Z_{Q1}	Z_{Q1} critical	Fit OK?
1	3PL	390.07	7	191463	102.38	510.57	Y
2	3PL	31.00	7	191463	6.41	510.57	Y
3	3PL	182.52	7	191463	46.91	510.57	Y
4	3PL	181.47	7	191463	46.63	510.57	Y
5	3PL	123.24	7	191463	31.07	510.57	Y
6	3PL	87.05	7	191463	21.39	510.57	Y
7	3PL	39.24	7	191463	8.62	510.57	Y
8	3PL	143.57	7	191463	36.50	510.57	Y
9	3PL	55.07	7	191463	12.85	510.57	Y
10	3PL	37.44	7	191463	8.13	510.57	Y
11	3PL	59.45	7	191463	14.02	510.57	Y
12	3PL	47.94	7	191463	10.94	510.57	Y
13	3PL	120.05	7	191463	30.21	510.57	Y
14	3PL	57.14	7	191463	13.40	510.57	Y
15	3PL	16.24	7	191463	2.47	510.57	Y
16	3PL	158.86	7	191463	40.59	510.57	Y
17	3PL	722.49	7	191463	191.22	510.57	Y
18	3PL	254.25	7	191463	66.08	510.57	Y

(Continued on next page)

Table 18. Mathematics Grade 4 Item Fit Statistics (cont.)

Item	Model	Chi Square	DF	Total N	Z_{Q1}	Z_{Q1} critical	Fit OK?
19	3PL	128.51	7	191463	32.47	510.57	Y
20	3PL	218.11	7	191463	56.42	510.57	Y
21	3PL	297.64	7	191463	77.68	510.57	Y
22	3PL	96.95	7	191463	24.04	510.57	Y
23	3PL	62.19	7	191463	14.75	510.57	Y
24	3PL	268.48	7	191463	69.88	510.57	Y
25	3PL	133.51	7	191463	33.81	510.57	Y
26	3PL	52.98	7	191463	12.29	510.57	Y
27	3PL	133.94	7	191463	33.93	510.57	Y
28	3PL	97.20	7	191463	24.11	510.57	Y
29	3PL	382.82	7	191463	100.44	510.57	Y
30	3PL	16.20	7	191463	2.46	510.57	Y
31	3PL	151.52	7	191463	38.63	510.57	Y
32	3PL	47.91	7	191463	10.93	510.57	Y
33	3PL	18.08	7	191463	2.96	510.57	Y
34	3PL	48.99	7	191463	11.22	510.57	Y
35	3PL	156.01	7	191463	39.82	510.57	Y
36	3PL	183.34	7	191463	47.13	510.57	Y
37	3PL	92.47	7	191463	22.84	510.57	Y
38	3PL	43.97	7	191463	9.88	510.57	Y
39	3PL	512.04	7	191463	134.98	510.57	Y
40	3PL	230.83	7	191463	59.82	510.57	Y
41	3PL	62.08	7	191463	14.72	510.57	Y
42	3PL	36.89	7	191463	7.99	510.57	Y
43	3PL	52.72	7	191463	12.22	510.57	Y
44	3PL	22.60	7	191463	4.17	510.57	Y
45	3PL	236.69	7	191463	61.39	510.57	Y
46	2PPC	289.62	17	191463	46.75	510.57	Y
47	2PPC	1726.51	17	191463	293.18	510.57	Y
48	2PPC	194.41	17	191463	30.43	510.57	Y
49	2PPC	560.48	17	191463	93.21	510.57	Y
50	2PPC	607.71	17	191463	101.31	510.57	Y
51	2PPC	941.64	17	191463	158.57	510.57	Y
52	2PPC	605.74	17	191463	100.97	510.57	Y
53	2PPC	345.47	17	191463	56.33	510.57	Y
54	2PPC	1089.40	26	191463	147.47	510.57	Y
55	2PPC	1619.83	26	191463	221.02	510.57	Y
56	2PPC	245.38	26	191463	30.42	510.57	Y
57	2PPC	1408.49	26	191463	191.72	510.57	Y

Table 19. Mathematics Grade 5 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	Z_{Q1}	Z_{Q1} critical	Fit OK?
1	3PL	54.33	7	195172	12.65	520.46	Y
2	3PL	829.39	7	195172	219.79	520.46	Y
3	3PL	284.34	7	195172	74.12	520.46	Y
4	3PL	39.78	7	195172	8.76	520.46	Y
5	3PL	105.41	7	195172	26.30	520.46	Y
6	3PL	210.95	7	195172	54.51	520.46	Y
7	3PL	354.51	7	195172	92.88	520.46	Y
8	3PL	16.10	7	195172	2.43	520.46	Y
9	3PL	100.12	7	195172	24.89	520.46	Y
10	3PL	166.02	7	195172	42.50	520.46	Y
11	3PL	295.11	7	195172	77.00	520.46	Y
12	3PL	66.65	7	195172	15.94	520.46	Y
13	3PL	27.12	7	195172	5.38	520.46	Y
14	3PL	1949.27	7	195172	519.09	520.46	Y
15	3PL	134.41	7	195172	34.05	520.46	Y
16	3PL	31.11	7	195172	6.44	520.46	Y
17	3PL	129.21	7	195172	32.66	520.46	Y
18	3PL	104.94	7	195172	26.18	520.46	Y
19	3PL	130.33	7	195172	32.96	520.46	Y
20	3PL	47.80	7	195172	10.90	520.46	Y
21	3PL	147.22	7	195172	37.47	520.46	Y
22	3PL	339.92	7	195172	88.98	520.46	Y
23	3PL	232.86	7	195172	60.36	520.46	Y
24	3PL	95.65	7	195172	23.69	520.46	Y
25	3PL	74.47	7	195172	18.03	520.46	Y
26	3PL	56.08	7	195172	13.12	520.46	Y
27	3PL	42.76	7	195172	9.56	520.46	Y
28	3PL	204.40	7	195172	52.76	520.46	Y
29	3PL	22.29	7	195172	4.09	520.46	Y
30	3PL	44.69	7	195172	10.07	520.46	Y
31	3PL	221.60	7	195172	57.36	520.46	Y
32	3PL	49.44	7	195172	11.34	520.46	Y
33	3PL	58.98	7	195172	13.89	520.46	Y
34	3PL	62.00	7	195172	14.70	520.46	Y
35	3PL	12.17	7	195172	1.38	520.46	Y
36	3PL	55.07	7	195172	12.85	520.46	Y
37	3PL	103.54	7	195172	25.80	520.46	Y
38	3PL	372.89	7	195172	97.79	520.46	Y
39	3PL	86.17	7	195172	21.16	520.46	Y
40	3PL	110.96	7	195172	27.78	520.46	Y
41	3PL	82.94	7	195172	20.30	520.46	Y
42	2PPC	325.00	17	195172	52.82	520.46	Y
43	2PPC	974.02	17	195172	164.13	520.46	Y

(Continued on next page)

Table 19. Mathematics Grade 5 Item Fit Statistics (cont.)

Item	Model	Chi Square	DF	Total N	Z_{Q1}	Z_{Q1} critical	Fit OK?
44	2PPC	1509.25	17	195172	255.92	520.46	Y
45	2PPC	795.38	17	195172	133.49	520.46	Y
46	2PPC	676.64	26	195172	90.23	520.46	Y
47	2PPC	1516.45	26	195172	206.69	520.46	Y
48	2PPC	1514.19	26	195172	206.37	520.46	Y
49	2PPC	572.97	26	195172	75.85	520.46	Y

Table 20. Mathematics Grade 6 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	Z_{Q1}	Z_{Q1} critical	Fit OK?
1	3PL	65.65	7	190284	15.68	507.42	Y
2	3PL	145.48	7	190284	37.01	507.42	Y
3	3PL	143.16	7	190284	36.39	507.42	Y
4	3PL	221.82	7	190284	57.41	507.42	Y
5	3PL	52.38	7	190284	12.13	507.42	Y
6	3PL	115.96	7	190284	29.12	507.42	Y
7	3PL	133.84	7	190284	33.90	507.42	Y
8	3PL	109.29	7	190284	27.34	507.42	Y
9	3PL	530.46	7	190284	139.90	507.42	Y
10	3PL	32.08	7	190284	6.70	507.42	Y
11	3PL	17.30	7	190284	2.75	507.42	Y
12	3PL	150.28	7	190284	38.29	507.42	Y
13	3PL	126.09	7	190284	31.83	507.42	Y
14	3PL	23.22	7	190284	4.33	507.42	Y
15	3PL	4.88	7	190284	-0.57	507.42	Y
16	3PL	77.81	7	190284	18.93	507.42	Y
17	3PL	92.69	7	190284	22.90	507.42	Y
18	3PL	43.72	7	190284	9.81	507.42	Y
19	3PL	40.44	7	190284	8.94	507.42	Y
20	3PL	199.93	7	190284	51.56	507.42	Y
21	3PL	42.65	7	190284	9.53	507.42	Y
22	3PL	126.41	7	190284	31.91	507.42	Y
23	3PL	39.96	7	190284	8.81	507.42	Y
24	3PL	88.74	7	190284	21.85	507.42	Y
25	3PL	147.81	7	190284	37.63	507.42	Y
26	3PL	255.24	7	190284	66.35	507.42	Y
27	3PL	107.35	7	190284	26.82	507.42	Y
28	3PL	171.44	7	190284	43.95	507.42	Y
29	3PL	83.24	7	190284	20.38	507.42	Y
30	3PL	16.85	7	190284	2.63	507.42	Y
31	3PL	68.52	7	190284	16.44	507.42	Y
32	3PL	152.93	7	190284	39.00	507.42	Y
33	3PL	66.53	7	190284	15.91	507.42	Y
34	3PL	136.24	7	190284	34.54	507.42	Y

(Continued on next page)

Table 20. Mathematics Grade 6 Item Fit Statistics (cont.)

Item	Model	Chi Square	DF	Total N	Z_{Q1}	Z_{Q1} critical	Fit OK?
35	3PL	87.46	7	190284	21.50	507.42	Y
36	3PL	3485.11	7	190284	929.56	507.42	N
37	3PL	240.33	7	190284	62.36	507.42	Y
38	3PL	914.56	7	190284	242.55	507.42	Y
39	3PL	151.89	7	190284	38.72	507.42	Y
40	3PL	764.12	7	190284	202.35	507.42	Y
41	2PPC	1639.86	17	190284	278.32	507.42	Y
42	2PPC	1165.93	17	190284	197.04	507.42	Y
43	2PPC	1403.86	17	190284	237.84	507.42	Y
44	2PPC	3262.42	17	190284	556.58	507.42	N
45	2PPC	576.61	17	190284	95.97	507.42	Y
46	2PPC	1985.21	17	190284	337.54	507.42	Y
47	2PPC	952.94	26	190284	128.54	507.42	Y
48	2PPC	879.90	26	190284	118.41	507.42	Y
49	2PPC	668.23	26	190284	89.06	507.42	Y
50	2PPC	938.36	26	190284	126.52	507.42	Y

Table 21. Mathematics Grade 7 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	Z_{Q1}	Z_{Q1} critical	Fit OK?
1	3PL	246.44	7	194058	63.99	517.49	Y
2	3PL	42.28	7	194058	9.43	517.49	Y
3	3PL	63.88	7	194058	15.20	517.49	Y
4	3PL	75.32	7	194058	18.26	517.49	Y
5	3PL	1452.61	7	194058	386.36	517.49	Y
6	3PL	50.12	7	194058	11.52	517.49	Y
7	3PL	218.36	7	194058	56.49	517.49	Y
8	3PL	109.12	7	194058	27.29	517.49	Y
9	3PL	71.73	7	194058	17.30	517.49	Y
10	3PL	140.13	7	194058	35.58	517.49	Y
11	3PL	27.52	7	194058	5.48	517.49	Y
12	3PL	1789.83	7	194058	476.48	517.49	Y
13	3PL	136.21	7	194058	34.53	517.49	Y
14	3PL	114.19	7	194058	28.65	517.49	Y
15	3PL	14.34	7	194058	1.96	517.49	Y
16	3PL	41.15	7	194058	9.13	517.49	Y
17	3PL	49.44	7	194058	11.34	517.49	Y
18	3PL	63.04	7	194058	14.98	517.49	Y
19	3PL	41.44	7	194058	9.20	517.49	Y
20	3PL	36.68	7	194058	7.93	517.49	Y
21	3PL	21.03	7	194058	3.75	517.49	Y
22	3PL	217.72	7	194058	56.32	517.49	Y
23	3PL	46.05	7	194058	10.44	517.49	Y
24	3PL	1092.27	7	194058	290.05	517.49	Y

(Continued on next page)

Table 21. Mathematics Grade 7 Item Fit Statistics (cont.)

Item	Model	Chi Square	DF	Total N	Z_{Q1}	Z_{Q1} critical	Fit OK?
25	3PL	19.30	7	194058	3.29	517.49	Y
26	The item was suppressed						
27	3PL	120.77	7	194058	30.41	517.49	Y
28	3PL	29.42	7	194058	5.99	517.49	Y
29	3PL	104.25	7	194058	25.99	517.49	Y
30	3PL	296.90	7	194058	77.48	517.49	Y
31	3PL	74.53	7	194058	18.05	517.49	Y
32	3PL	28.51	7	194058	5.75	517.49	Y
33	3PL	65.05	7	194058	15.51	517.49	Y
34	3PL	131.67	7	194058	33.32	517.49	Y
35	3PL	76.01	7	194058	18.44	517.49	Y
36	3PL	74.48	7	194058	18.03	517.49	Y
37	3PL	54.04	7	194058	12.57	517.49	Y
38	3PL	62.83	7	194058	14.92	517.49	Y
39	3PL	37.32	7	194058	8.10	517.49	Y
40	3PL	108.11	7	194058	27.02	517.49	Y
41	3PL	76.13	7	194058	18.48	517.49	Y
42	3PL	15.43	7	194058	2.25	517.49	Y
43	3PL	895.67	7	194058	237.51	517.49	Y
44	3PL	34.39	7	194058	7.32	517.49	Y
45	3PL	273.62	7	194058	71.26	517.49	Y
46	2PPC	324.54	17	194058	52.74	517.49	Y
47	2PPC	560.60	17	194058	93.23	517.49	Y
48	2PPC	1459.71	17	194058	247.42	517.49	Y
49	2PPC	314.76	17	194058	51.06	517.49	Y
50	2PPC	1698.65	26	194058	231.95	517.49	Y
51	2PPC	2109.52	26	194058	288.93	517.49	Y
52	2PPC	874.54	26	194058	117.67	517.49	Y
53	2PPC	1417.02	26	194058	192.90	517.49	Y

Note: Item 26 was suppressed.

Table 22. Mathematics Grade 8 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	Z_{Q1}	Z_{Q1} critical	Fit OK?
1	3PL	54.07	7	192149	12.58	512.40	Y
2	3PL	169.74	7	192149	43.49	512.40	Y
3	3PL	64.29	7	192149	15.31	512.40	Y
4	3PL	58.30	7	192149	13.71	512.40	Y
5	3PL	48.23	7	192149	11.02	512.40	Y
6	3PL	50.66	7	192149	11.67	512.40	Y
7	3PL	363.51	7	192149	95.28	512.40	Y
8	3PL	120.98	7	192149	30.46	512.40	Y
9	3PL	129.35	7	192149	32.70	512.40	Y
10	3PL	109.09	7	192149	27.28	512.40	Y
11	3PL	49.37	7	192149	11.32	512.40	Y

(Continued on next page)

Table 22. Mathematics Grade 8 Item Fit Statistics (cont.)

Item	Model	Chi Square	DF	Total N	Z_{Q1}	Z_{Q1} critical	Fit OK?
12	3PL	46.73	7	192149	10.62	512.40	Y
13	3PL	92.58	7	192149	22.87	512.40	Y
14	3PL	29.49	7	192149	6.01	512.40	Y
15	3PL	132.60	7	192149	33.57	512.40	Y
16	3PL	261.46	7	192149	68.01	512.40	Y
17	3PL	37.59	7	192149	8.18	512.40	Y
18	3PL	127.63	7	192149	32.24	512.40	Y
19	3PL	236.37	7	192149	61.30	512.40	Y
20	3PL	143.47	7	192149	36.47	512.40	Y
21	3PL	58.13	7	192149	13.67	512.40	Y
22	3PL	115.67	7	192149	29.04	512.40	Y
23	3PL	204.52	7	192149	52.79	512.40	Y
24	3PL	177.49	7	192149	45.57	512.40	Y
25	3PL	109.96	7	192149	27.52	512.40	Y
26	3PL	125.61	7	192149	31.70	512.40	Y
27	3PL	38.64	7	192149	8.46	512.40	Y
28	3PL	74.59	7	192149	18.06	512.40	Y
29	3PL	69.09	7	192149	16.59	512.40	Y
30	3PL	113.92	7	192149	28.57	512.40	Y
31	3PL	83.38	7	192149	20.41	512.40	Y
32	3PL	125.00	7	192149	31.54	512.40	Y
33	3PL	77.46	7	192149	18.83	512.40	Y
34	3PL	637.02	7	192149	168.38	512.40	Y
35	3PL	137.32	7	192149	34.83	512.40	Y
36	3PL	184.78	7	192149	47.51	512.40	Y
37	3PL	137.48	7	192149	34.87	512.40	Y
38	3PL	196.66	7	192149	50.69	512.40	Y
39	3PL	310.97	7	192149	81.24	512.40	Y
40	3PL	85.01	7	192149	20.85	512.40	Y
41	3PL	128.03	7	192149	32.35	512.40	Y
42	3PL	193.19	7	192149	49.76	512.40	Y
43	2PPC	1852.34	17	192149	314.76	512.40	Y
44	2PPC	557.96	17	192149	92.77	512.40	Y
45	2PPC	348.85	17	192149	56.91	512.40	Y
46	2PPC	316.13	17	192149	51.30	512.40	Y
47	2PPC	961.85	17	192149	162.04	512.40	Y
48	2PPC	689.95	17	192149	115.41	512.40	Y
49	2PPC	218.69	17	192149	34.59	512.40	Y
50	2PPC	258.55	17	192149	41.42	512.40	Y
51	2PPC	2368.92	26	192149	324.90	512.40	Y
52	2PPC	1077.74	26	192149	145.85	512.40	Y
53	2PPC	4675.54	26	192149	644.78	512.40	N
54	2PPC	301.07	26	192149	38.15	512.40	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 his or her response on another item. Statistically speaking, when a student's ability is accounted for, his or her 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 5, 7, and 8. In Grade 3, one pair of items was found to be locally dependent: items 2 and 20 (Q_3 = 0.319). In Grade 4, one pair of items was found to be locally dependent: items 49 and 52 (Q_3 = 0.204). In Grade 6, one pair of items was found to be locally dependent: items 13 and 41 (Q_3 = 0.309). The magnitudes of these statistics were not sufficient to warrant any concern. Anchor items were excluded from Q_3 computation.

Scaling and Equating

The 2011 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 2011 OP forms were pre-equated to the corresponding 2010 assessments. Prior to pre-equating, the FT items administered in 2010 were placed onto the OP scales in each grade. The equating of 2010 FT items to the 2010 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 NYSTP item pool were items field tested in 2009, 2008, 2006, and 2005. All items field-tested between 2005 and 2009 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*. The pool also included items owned by CTB/McGraw-Hill. These items consisted mostly of *TerraNova* items but also included items field-tested in New York State in 2010. *TerraNova* items were also equated to NYSTP OP scales.

At the pre-equating stage, the pool of FT items administered in 2005, 2006, 2007, 2008, 2009, and 2010 and the *TerraNova* items equated to NYSTP OP scale were used to select the 2011 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 2011 forms with the target 2010 OP forms (note that the 2010 OP TCC and SE curves were based on OP parameters, and the 2011 TCC and SE curves were based on FT parameters transformed to the NYS OP scale).

In the second equating procedure, the 2011 Mathematics OP data were re-calibrated after the 2011 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 ten characters of last name match (including blank spaces)
b) at least five 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) two of three parts of date of birth (MM or DD or YY) must be the same or one must be blank

In the second OP test equating step, the year 2010 item parameters for items contained in FT anchor forms were used as anchors to transform the 2011 OP item parameters onto the OP scale

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 2010 and that relationship defined by the FT anchor item parameter estimates in new administration year 2011. This places the transformed parameters for the OP test items onto

the New York State OP scale. In this procedure, new 2011 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 2010 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 (2010 FT anchor parameter estimates) and estimate (2011 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 23 presents the 2011 OP transformation parameters for New York State Grades 3–8 Mathematics Tests.

Table 23. NYSTP Mathematics 2011 Final Transformation Constants

Grade	*M1*	*M2*
3	15.90	688.50
4	29.27	688.09
5	26.39	686.81
6	27.46	682.47
7	24.67	679.50
8	25.02	680.03

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 four consecutive years—2008, 2009, 2010, and 2011—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 2010 and 2011 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 2010 FT anchor estimates and the other on transformed estimates from the 2011 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 0.96 for Grade 6 to 0.99 for Grades 3, 5, and 7. Correlations for *a*-parameter input and estimate ranged from 0.84 for Grades 3 and 8 to 0.96 for Grade 5. All correlations were above the NYS criterion.

Overall TCC alignment for anchor set input and estimate was very good. Therefore, despite the fact that a few items were flagged by the Stocking and Lord's method, 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 24–29. 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 24. Grade 3 2011 Operational Item Parameter Estimates

Item	Max Pts	*a*-par/ alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3
1	1	0.024	654.180	0.200	
2	1	0.047	653.039	0.208	
3	1	0.069	670.000	0.296	
4	1	0.046	675.667	0.097	
5	1	0.049	647.391	0.208	
6	1	0.064	670.854	0.180	
7	1	0.052	647.368	0.208	
8	1	0.057	661.420	0.157	
9	1	0.076	703.706	0.171	
10	1	0.055	651.883	0.071	
11	1	0.052	672.463	0.243	
12	1	0.044	645.068	0.200	
13	1	0.062	656.318	0.179	
14	1	0.053	652.721	0.208	
15	1	0.079	694.938	0.268	
16	1	0.057	697.899	0.153	
17	1	0.049	660.199	0.208	
18	1	0.055	673.996	0.090	
19	1	0.037	689.401	0.364	
20	1	0.047	648.882	0.208	
21	1	0.050	674.589	0.136	
22	1	0.072	683.183	0.058	
23	1	0.066	668.070	0.111	
24	1	0.066	685.043	0.199	
25	1	0.062	691.422	0.229	
26	1	0.047	671.506	0.239	
27	1	0.053	641.098	0.200	
28	1	0.056	657.689	0.140	
29	1	0.071	689.548	0.184	
30	1	0.053	670.416	0.094	
31	1	0.074	676.711	0.185	
32	1	0.035	686.671	0.139	
33	1	0.065	704.451	0.140	
34	1	0.031	670.239	0.074	
35	1	0.040	694.131	0.341	
36	1	0.072	715.405	0.138	
37	1	0.046	682.993	0.188	
38	1	0.042	664.399	0.192	
39	1	0.066	676.051	0.201	

(Continued on next page)

Table 24. Grade 3 2011 Operational Item Parameter Estimates (cont.)

Item	Max Pts	*a*-par/ alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3
40	1	0.031	705.061	0.145	
41	2	0.090	60.097	60.637	
42	2	0.080	53.523	52.934	
43	2	0.052	34.503	35.602	
44	2	0.038	27.476	24.707	
45	3	0.059	39.091	41.510	40.173
46	3	0.051	32.419	32.673	34.365

Table 25. Grade 4 2011 Operational Item Parameter Estimates

Item	Max Pts	a-par/ alpha	b-par/ gamma1	c-par/ gamma2	gamma3
1	1	0.028	626.537	0.200	
2	1	0.027	666.277	0.252	
3	1	0.035	612.216	0.045	
4	1	0.026	622.649	0.200	
5	1	0.042	657.091	0.079	
6	1	0.025	666.128	0.280	
7	1	0.038	677.437	0.161	
8	1	0.029	603.699	0.045	
9	1	0.044	657.670	0.072	
10	1	0.037	669.714	0.136	
11	1	0.035	667.998	0.194	
12	1	0.031	623.728	0.123	
13	1	0.039	642.698	0.101	
14	1	0.030	668.785	0.105	
15	1	0.033	675.994	0.305	
16	1	0.030	701.659	0.233	
17	1	0.014	571.991	0.200	
18	1	0.022	668.205	0.110	
19	1	0.040	709.775	0.102	
20	1	0.034	715.908	0.102	
21	1	0.022	612.215	0.200	
22	1	0.038	684.953	0.110	
23	1	0.025	605.138	0.200	
24	1	0.022	617.617	0.045	
25	1	0.034	688.305	0.139	
26	1	0.036	677.956	0.113	
27	1	0.044	668.000	0.115	
28	1	0.021	616.474	0.082	
29	1	0.044	711.591	0.182	
30	1	0.025	672.619	0.186	
31	1	0.025	666.893	0.121	

(Continued on next page)

Table 25. Grade 4 2011 Operational Item Parameter Estimates (cont.)

Item	Max Pts	a-par/ alpha	b-par/ gamma1	c-par/ gamma2	gamma3
32	1	0.039	683.521	0.160	
33	1	0.028	687.531	0.315	
34	1	0.032	690.325	0.221	
35	1	0.044	698.886	0.251	
36	1	0.039	679.809	0.169	
37	1	0.047	677.806	0.191	
38	1	0.026	632.115	0.105	
39	1	0.037	728.176	0.078	
40	1	0.012	670.846	0.045	
41	1	0.023	661.955	0.200	
42	1	0.038	683.709	0.148	
43	1	0.034	701.098	0.196	
44	1	0.029	686.349	0.212	
45	1	0.026	676.248	0.149	
46	2	0.042	27.360	27.424	
47	2	0.043	29.012	28.978	
48	2	0.052	32.127	34.555	
49	2	0.048	32.626	31.487	
50	2	0.035	22.209	24.635	
51	2	0.041	28.310	28.730	
52	2	0.043	29.731	28.089	
53	2	0.055	36.830	37.679	
54	3	0.024	16.792	17.461	17.715
55	3	0.029	18.159	19.392	18.474
56	3	0.029	18.518	19.049	19.647
57	3	0.029	18.456	18.730	20.076

Table 26. Grade 5 2011 Operational Item Parameter Estimates

Item	Max Pts	a-par/ alpha	b-par/ gamma1	c-par/ gamma2	gamma3
1	1	0.037	665.186	0.370	
2	1	0.039	681.124	0.102	
3	1	0.049	741.900	0.332	
4	1	0.032	672.910	0.276	
5	1	0.052	667.312	0.187	
6	1	0.028	609.909	0.200	
7	1	0.056	697.682	0.327	
8	1	0.016	669.236	0.069	
9	1	0.039	710.880	0.158	
10	1	0.051	682.627	0.192	
11	1	0.053	667.505	0.164	
12	1	0.039	687.593	0.190	

(Continued on next page)

Table 26. Grade 5 2011 Operational Item Parameter Estimates (cont.)

Item	Max Pts	a-par/ alpha	b-par/ gamma1	c-par/ gamma2	gamma3
13	1	0.038	671.181	0.207	
14	1	0.006	788.149	0.200	
15	1	0.043	668.803	0.205	
16	1	0.033	664.437	0.369	
17	1	0.054	694.721	0.383	
18	1	0.033	631.601	0.200	
19	1	0.047	687.871	0.190	
20	1	0.035	645.529	0.155	
21	1	0.021	603.224	0.200	
22	1	0.039	658.133	0.037	
23	1	0.052	705.103	0.190	
24	1	0.034	650.831	0.156	
25	1	0.051	669.228	0.163	
26	1	0.030	670.252	0.377	
27	1	0.032	695.715	0.178	
28	1	0.049	659.217	0.179	
29	1	0.022	701.257	0.253	
30	1	0.036	649.468	0.184	
31	1	0.045	683.280	0.375	
32	1	0.023	676.492	0.228	
33	1	0.036	693.499	0.129	
34	1	0.034	707.334	0.183	
35	1	0.021	686.248	0.120	
36	1	0.027	675.378	0.218	
37	1	0.029	685.731	0.198	
38	1	0.051	710.301	0.204	
39	1	0.049	679.603	0.264	
40	1	0.029	629.928	0.103	
41	1	0.029	649.134	0.186	
42	2	0.033	21.100	23.463	
43	2	0.043	28.250	29.763	
44	2	0.037	24.152	25.443	
45	2	0.029	19.028	19.315	
46	3	0.020	12.173	13.627	14.346
47	3	0.024	15.963	15.570	15.942
48	3	0.049	31.344	33.733	34.615
49	3	0.038	24.378	25.306	26.097

Table 27. Grade 6 2011 Operational Item Parameter Estimates

Item	Max Pts	*a*-par/ alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3
1	1	0.021	647.841	0.115	
2	1	0.035	629.055	0.200	
3	1	0.035	680.488	0.318	
4	1	0.034	616.727	0.200	
5	1	0.036	665.576	0.074	
6	1	0.042	681.642	0.025	
7	1	0.045	670.677	0.288	
8	1	0.051	685.005	0.188	
9	1	0.014	590.972	0.200	
10	1	0.022	653.728	0.069	
11	1	0.030	679.133	0.212	
12	1	0.026	617.137	0.200	
13	1	0.035	657.502	0.072	
14	1	0.037	660.768	0.124	
15	1	0.030	677.202	0.170	
16	1	0.042	702.107	0.112	
17	1	0.047	685.864	0.063	
18	1	0.034	679.697	0.196	
19	1	0.024	692.384	0.263	
20	1	0.044	664.002	0.075	
21	1	0.035	690.596	0.172	
22	1	0.032	639.220	0.297	
23	1	0.018	686.885	0.077	
24	1	0.045	702.697	0.144	
25	1	0.044	682.609	0.100	
26	1	0.046	673.002	0.209	
27	1	0.046	674.150	0.322	
28	1	0.031	620.179	0.200	
29	1	0.040	665.943	0.139	
30	1	0.036	682.990	0.451	
31	1	0.051	697.605	0.225	
32	1	0.052	670.733	0.076	
33	1	0.046	681.716	0.162	
34	1	0.056	673.500	0.322	
35	1	0.034	636.272	0.057	
36	1	0.068	720.160	0.241	
37	1	0.023	662.833	0.319	
38	1	0.073	740.046	0.424	
39	1	0.040	665.895	0.078	
40	1	0.059	714.041	0.088	
41	2	0.032	22.434	19.826	
42	2	0.034	22.714	21.857	

(Continued on next page)

Table 27. Grade 6 2011 Operational Item Parameter Estimates (cont.)

Item	Max Pts	*a*-par/ alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3
43	2	0.043	28.333	27.591	
44	2	0.033	21.899	22.399	
45	2	0.047	32.313	30.048	
46	2	0.035	25.612	23.834	
47	3	0.038	25.724	25.113	26.031
48	3	0.024	14.403	15.729	14.668
49	3	0.035	24.336	22.586	23.955
50	3	0.055	35.485	37.412	39.016

Table 28. Grade 7 2011 Operational Item Parameter Estimates

Item	Max Pts	a-par/ alpha	b-par/ gamma1	c-par/ gamma2	gamma3
1	1	0.032	607.914	0.200	
2	1	0.040	664.880	0.221	
3	1	0.043	702.544	0.082	
4	1	0.028	632.635	0.158	
5	1	0.010	690.866	0.200	
6	1	0.046	674.324	0.190	
7	1	0.055	700.650	0.216	
8	1	0.028	715.468	0.238	
9	1	0.028	631.862	0.158	
10	1	0.032	669.696	0.500	
11	1	0.040	659.207	0.054	
12	1	0.018	683.829	0.500	
13	1	0.045	729.201	0.308	
14	1	0.042	680.021	0.485	
15	1	0.029	673.773	0.173	
16	1	0.042	692.575	0.208	
17	1	0.039	652.873	0.041	
18	1	0.054	687.159	0.418	
19	1	0.039	674.474	0.139	
20	1	0.030	662.082	0.243	
21	1	0.033	689.378	0.188	
22	1	0.029	587.773	0.200	
23	1	0.040	661.666	0.180	
24	1	0.033	643.535	0.158	
25	1	0.029	660.464	0.160	
26	1	The item was suppressed			
27	1	0.056	690.917	0.163	
28	1	0.036	664.375	0.185	
29	1	0.036	650.767	0.126	
30	1	0.053	692.581	0.100	

(Continued on next page)

Table 28. Grade 7 2011 Operational Item Parameter Estimates (cont.)

Item	Max Pts	a-par/ alpha	b-par/ gamma1	c-par/ gamma2	gamma3
31	1	0.042	683.156	0.234	
32	1	0.040	651.987	0.222	
33	1	0.049	704.003	0.188	
34	1	0.041	681.898	0.212	
35	1	0.018	653.954	0.042	
36	1	0.041	664.270	0.160	
37	1	0.018	669.462	0.035	
38	1	0.045	693.394	0.146	
39	1	0.035	668.892	0.212	
40	1	0.048	685.373	0.246	
41	1	0.038	659.811	0.182	
42	1	0.018	665.163	0.101	
43	1	0.027	678.898	0.158	
44	1	0.041	679.704	0.266	
45	1	0.061	696.359	0.122	
46	2	0.035	23.059	23.569	
47	2	0.040	28.023	26.873	
48	2	0.058	38.291	40.319	
49	2	0.029	21.124	18.386	
50	3	0.035	23.705	22.655	22.996
51	3	0.033	21.731	21.712	22.220
52	3	0.044	30.715	30.262	30.062
53	3	0.029	18.672	19.729	19.791

Table 29. Grade 8 2011 Operational Item Parameter Estimates

Item	Max Pts	a-par/ alpha	b-par/ gamma1	c-par/ gamma2	gamma3
1	1	0.047	694.281	0.096	
2	1	0.030	659.411	0.285	
3	1	0.045	699.758	0.119	
4	1	0.030	660.004	0.087	
5	1	0.048	668.016	0.256	
6	1	0.044	671.287	0.134	
7	1	0.018	661.601	0.076	
8	1	0.047	667.879	0.165	
9	1	0.040	669.294	0.195	
10	1	0.042	665.669	0.369	
11	1	0.044	674.676	0.306	
12	1	0.041	672.711	0.239	
13	1	0.042	684.035	0.113	
14	1	0.031	678.325	0.191	
15	1	0.018	667.965	0.118	

(Continued on next page)

Table 29. Grade 8 2011 Operational Item Parameter Estimates (cont.)

Item	Max Pts	a-par/ alpha	b-par/ gamma1	c-par/ gamma2	gamma3
16	1	0.020	637.427	0.200	
17	1	0.029	702.646	0.259	
18	1	0.058	682.695	0.224	
19	1	0.046	699.560	0.326	
20	1	0.025	644.796	0.245	
21	1	0.042	708.465	0.242	
22	1	0.044	700.375	0.121	
23	1	0.040	665.036	0.276	
24	1	0.051	686.618	0.289	
25	1	0.072	680.269	0.333	
26	1	0.032	653.235	0.153	
27	1	0.043	690.461	0.199	
28	1	0.053	676.546	0.273	
29	1	0.049	696.451	0.219	
30	1	0.031	661.759	0.125	
31	1	0.043	683.265	0.275	
32	1	0.068	704.981	0.323	
33	1	0.040	683.877	0.329	
34	1	0.016	637.352	0.200	
35	1	0.060	716.995	0.245	
36	1	0.024	658.063	0.081	
37	1	0.040	675.063	0.089	
38	1	0.057	683.436	0.195	
39	1	0.032	692.223	0.237	
40	1	0.059	664.815	0.214	
41	1	0.028	645.437	0.076	
42	1	0.049	669.573	0.196	
43	2	0.030	21.727	18.591	
44	2	0.060	39.418	41.823	
45	2	0.056	36.598	39.825	
46	2	0.055	37.301	37.186	
47	2	0.046	31.693	30.284	
48	2	0.033	20.880	23.285	
49	2	0.059	40.693	39.430	
50	2	0.061	41.836	42.595	
51	3	0.013	10.703	9.407	7.671
52	3	0.036	24.189	25.346	24.584
53	3	0.044	30.509	30.543	30.246
54	3	0.041	26.782	27.083	28.176

Test Characteristic Curves

Test Characteristic Curves (TCCs) provide an overview of the test in the IRT scale score metric. The 2010 and 2011 TCCs were generated using final OP item parameters. TCCs are the summation of all the Item Characteristic Curves (ICCs) for items that contribute to the OP scale score. Standard Error (SE) curves graphically show the amount of measurement error at different ability levels. The 2010 and 2011 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. It should be noted that although the 2010 OP curves are considered to be target curves for the 2011 OP test TCCs, NYSED requested that the 2011 forms be more difficult than the 2010 forms, which was taken into consideration during new form selection. Note that in all figures the blue TCCs and SE curves represent the 2010 OP test and pink TCCs and SE curves represent the 2011 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 2010 and 2011 OP TCCs and SE

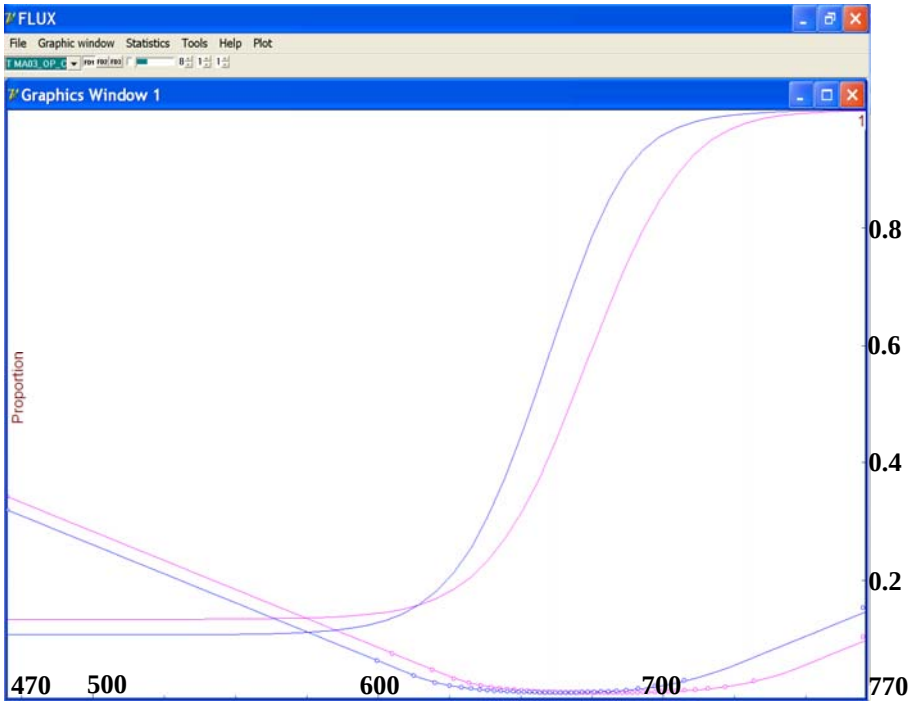


Figure 2. Grade 4 Mathematics 2010 and 2011 OP TCCs and SE

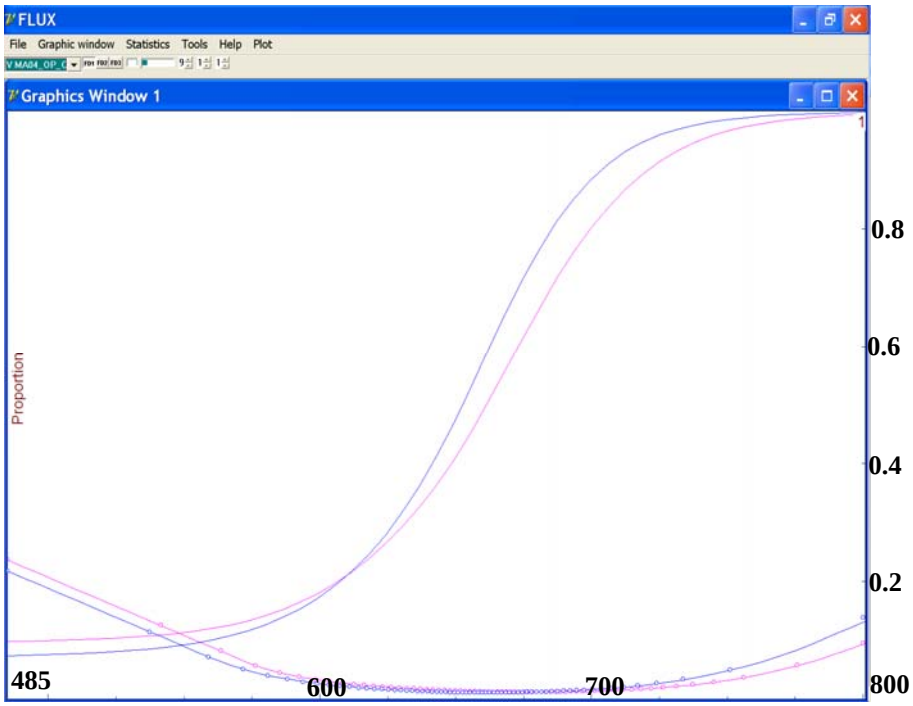


Figure 3. Grade 5 Mathematics 2010 and 2011 OP TCCs and SE

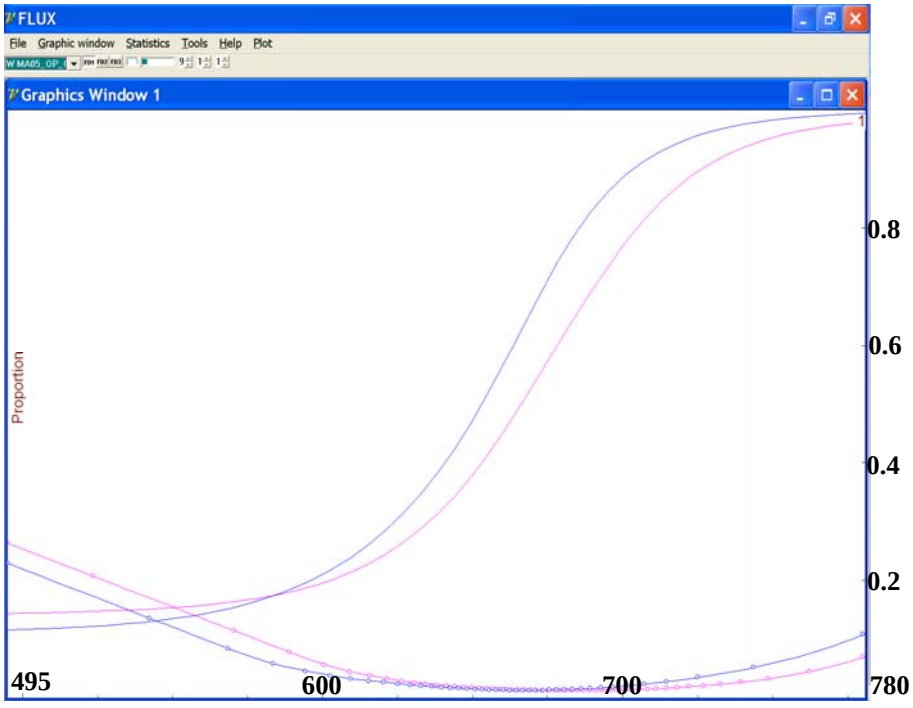


Figure 4. Grade 6 Mathematics 2010 and 2011 OP TCCs and SE

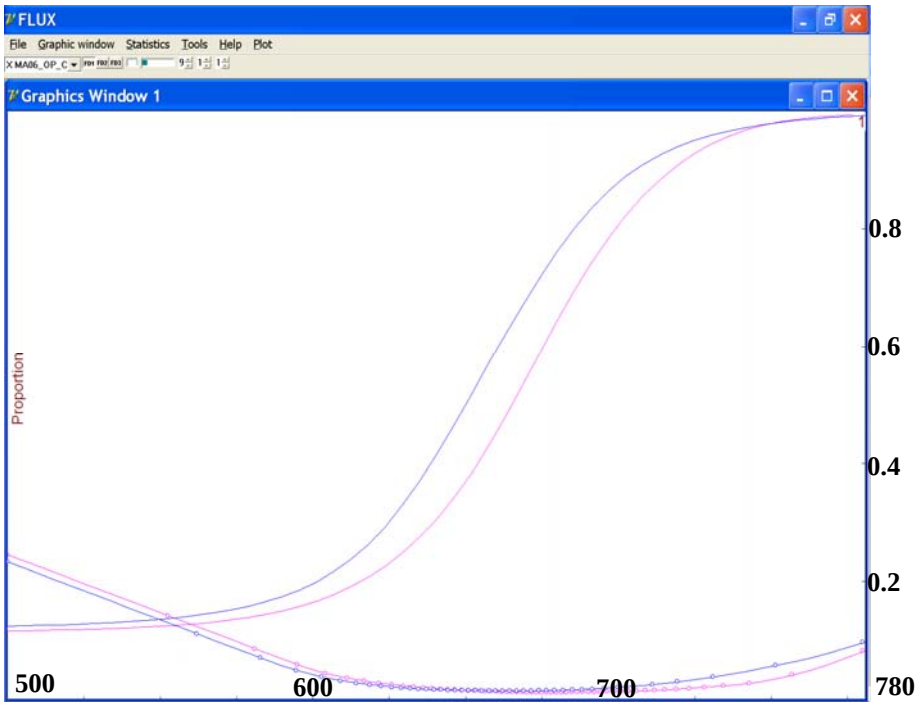


Figure 5. Grade 7 Mathematics 2010 and 2011 OP TCCs and SE

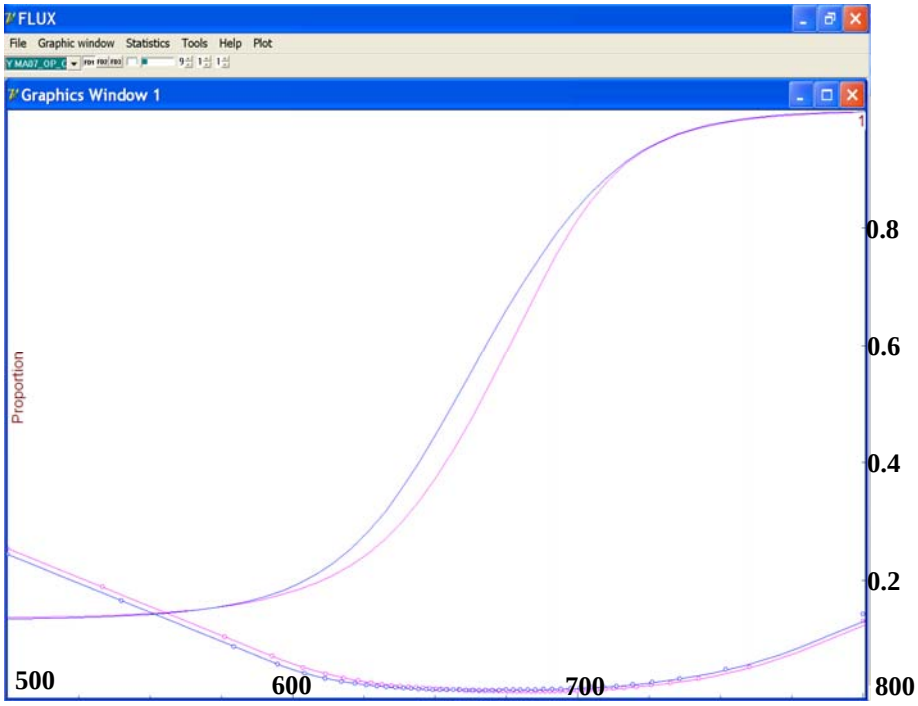
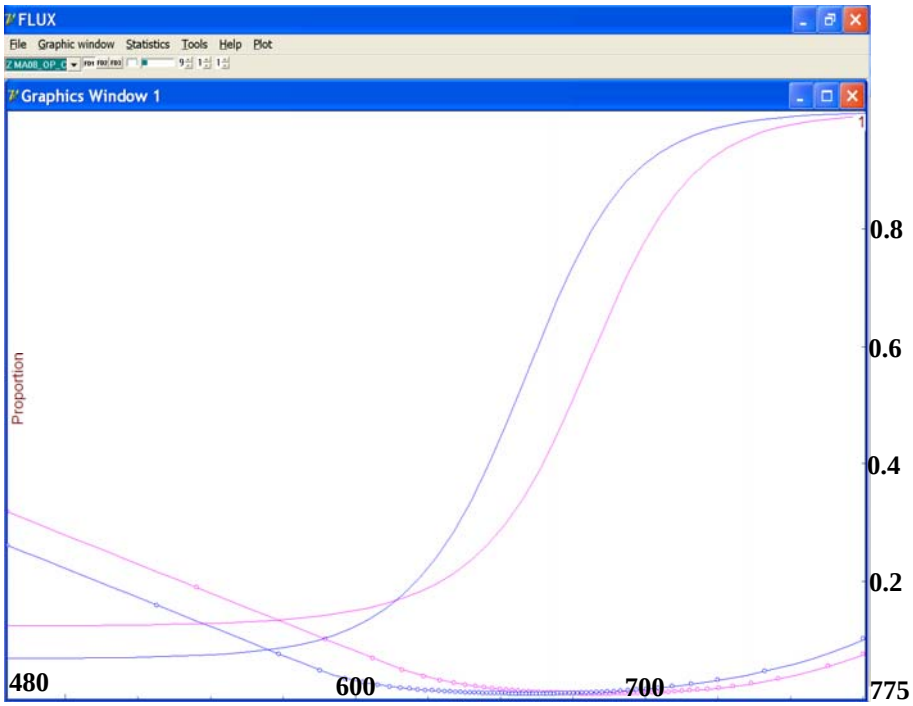


Figure 6. Grade 8 Mathematics 2010 and 2011 OP TCCs and SE



As seen in Figures 1–6, the 2011 TCCs for all grades were found to be to the right of the 2010 TCCs indicating that the 2011 forms tended to be more difficult than the 2010 forms for most of the students. The SE curves were well aligned for all grades. It should be noted that potential differences in test form difficulty at different ability levels are accounted for in the equating and in the resulting raw score-to-scale score conversion tables, so that students of the same ability are expected to obtain the same scale score regardless of which form they took.

Scoring Procedure

New York State students were scored using the number correct (NC) scoring method. This method considers how many score points a student obtained on a test in determining his or her 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 54 points (Grade 3) to 73 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:

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

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), and

$\tilde{\theta}$ 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 Mathematics Tests. It is used to derive other scores that describe test performance, such as the four performance levels and the standard-based performance index scores (SPIs). Scores on the NYSTP examinations are determined using number-correct scoring. Raw score-to-scale score conversion tables are presented in this section. The lowest and highest obtainable scores for each grade were the same as in 2006 (baseline year).

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

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

where

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

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

It should be noted that the information is estimated based on thetas in 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 30. Grade 3 Raw Score-to-Scale Score (with Standard Error)

Raw Score	Scale Score	Standard Error
0	470	174
1	470	174
2	470	174
3	470	174
4	470	174
5	470	174
6	470	174
7	470	174
8	605	39
9	619	25
10	627	17
11	632	13
12	636	11
13	640	10
14	643	9
15	645	8
16	648	7
17	650	7
18	652	7
19	654	6
20	656	6
21	658	6
22	660	6
23	661	5
24	663	5
25	664	5
26	666	5
27	668	5
28	669	5
29	670	5
30	672	5
31	673	5
32	675	5
33	676	5
34	678	5
35	679	5
36	681	5
37	682	5
38	684	5
39	686	5
40	687	5
41	689	5
42	691	5

(Continued on next page)

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

Raw Score	Scale Score	Standard Error
43	693	5
44	695	5
45	697	6
46	699	6
47	702	6
48	705	6
49	708	7
50	711	7
51	716	8
52	722	10
53	732	15
54	770	53

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

Raw Score	Scale Score	Standard Error
0	485	120
1	485	120
2	485	120
3	485	120
4	485	120
5	485	120
6	485	120
7	485	120
8	542	64
9	564	42
10	576	29
11	585	23
12	593	19
13	598	17
14	604	15
15	608	14
16	612	13
17	616	12
18	620	12
19	623	11
20	626	11
21	629	10
22	632	10
23	635	9
24	637	9

(Continued on next page)

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

Raw Score	Scale Score	Standard Error
25	639	9
26	642	8
27	644	8
28	646	8
29	648	8
30	650	7
31	652	7
32	654	7
33	655	7
34	657	7
35	659	7
36	661	7
37	662	7
38	664	7
39	665	6
40	667	6
41	669	6
42	670	6
43	672	6
44	673	6
45	675	6
46	677	6
47	678	6
48	680	6
49	682	6
50	683	6
51	685	6
52	687	7
53	689	7
54	691	7
55	692	7
56	694	7
57	697	7
58	699	7
59	701	7
60	703	8
61	706	8
62	709	8
63	712	8
64	715	9
65	718	9
66	722	10
67	726	10

(Continued on next page)

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

Raw Score	Scale Score	Standard Error
68	731	11
69	737	13
70	745	15
71	756	19
72	776	30
73	800	49

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

Raw Score	Scale Score	Standard Error
0	495	134
1	495	134
2	495	134
3	495	134
4	495	134
5	495	134
6	495	134
7	495	134
8	495	134
9	523	105
10	570	59
11	589	40
12	600	29
13	609	23
14	616	19
15	621	17
16	626	15
17	631	14
18	634	13
19	638	12
20	641	11
21	644	10
22	647	10
23	650	9
24	652	9
25	655	9
26	657	8
27	659	8
28	661	8
29	663	8
30	665	7
31	667	7
32	669	7
33	671	7

(Continued on next page)

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

Raw Score	Scale Score	Standard Error
34	673	7
35	675	7
36	677	7
37	679	7
38	681	7
39	683	7
40	685	7
41	687	7
42	689	7
43	691	7
44	693	7
45	696	7
46	698	7
47	700	8
48	703	8
49	705	8
50	708	8
51	711	9
52	714	9
53	718	10
54	722	10
55	727	11
56	732	13
57	739	14
58	748	17
59	762	23
60	780	36
61	780	36

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

Raw Score	Scale Score	Standard Error
0	500	125
1	500	125
2	500	125
3	500	125
4	500	125
5	500	125
6	500	125
7	500	125
8	553	72
9	581	44
10	595	30
11	604	22

(Continued on next page)

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

Raw Score	Scale Score	Standard Error
12	611	18
13	617	16
14	622	14
15	626	13
16	630	12
17	633	11
18	636	10
19	639	10
20	642	9
21	644	9
22	646	8
23	649	8
24	651	8
25	653	7
26	655	7
27	657	7
28	658	7
29	660	7
30	662	7
31	663	7
32	665	6
33	667	6
34	668	6
35	670	6
36	672	6
37	673	6
38	675	6
39	676	6
40	678	6
41	680	6
42	681	6
43	683	6
44	685	6
45	687	6
46	688	6
47	690	6
48	692	7
49	694	7
50	696	7
51	699	7
52	701	7
53	703	7
54	706	8

(Continued on next page)

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

Raw Score	Scale Score	Standard Error
55	709	8
56	712	8
57	715	9
58	719	9
59	723	10
60	728	11
61	734	12
62	743	14
63	757	21
64	780	42

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

Raw Score	Scale Score	Standard Error
0	500	129
1	500	129
2	500	129
3	500	129
4	500	129
5	500	129
6	500	129
7	500	129
8	500	129
9	533	96
10	576	53
11	593	36
12	604	27
13	612	21
14	618	18
15	623	15
16	628	14
17	631	12
18	635	11
19	638	10
20	641	10
21	643	9
22	646	9
23	648	8
24	650	8
25	652	8
26	654	8
27	656	7
28	658	7
29	660	7

(Continued on next page)

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

Raw Score	Scale Score	Standard Error
30	662	7
31	664	7
32	666	7
33	667	7
34	669	7
35	671	6
36	672	6
37	674	6
38	676	6
39	677	6
40	679	6
41	680	6
42	682	6
43	684	6
44	685	6
45	687	6
46	689	6
47	690	6
48	692	6
49	694	6
50	696	6
51	698	6
52	700	7
53	702	7
54	704	7
55	707	7
56	710	8
57	713	8
58	716	9
59	721	10
60	726	11
61	733	13
62	742	17
63	760	27
64	800	66

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

Raw Score	Scale Score	Standard Error
0	480	162
1	480	162
2	480	162
3	480	162
4	480	162

(Continued on next page)

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

Raw Score	Scale Score	Standard Error
5	480	162
6	480	162
7	480	162
8	480	162
9	545	96
10	590	52
11	606	36
12	616	26
13	623	20
14	629	17
15	634	14
16	638	13
17	641	11
18	644	10
19	647	10
20	650	9
21	652	8
22	654	8
23	656	8
24	658	7
25	660	7
26	661	7
27	663	6
28	665	6
29	666	6
30	668	6
31	669	6
32	670	6
33	672	6
34	673	5
35	674	5
36	676	5
37	677	5
38	678	5
39	679	5
40	680	5
41	682	5
42	683	5
43	684	5
44	685	5
45	687	5
46	688	5
47	689	5

(Continued on next page)

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

Raw Score	Scale Score	Standard Error
48	690	5
49	692	5
50	693	5
51	694	6
52	696	6
53	697	6
54	699	6
55	701	6
56	702	6
57	704	6
58	706	6
59	708	7
60	710	7
61	713	7
62	715	8
63	718	8
64	721	9
65	725	10
66	730	11
67	737	14
68	746	18
69	763	29
70	775	39

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 E.

For the 2011 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 36 presents SPI target ranges. The objectives in this table are denoted as follows: 1—Number Sense and Operations, 2—Algebra, 3—Geometry, 4—Measurement, and 5—Statistics and Probability.

Table 36. SPI Target Ranges

Grade	Objective	# Items	Total Points	Level III Cut SPI Target Range
3	1	23	25	66–76
	2	5	7	49–64
	3	6	7	67–76
	4	7	7	61–75
	5	5	8	75–85
4	1	29	34	55–67
	2	6	9	68–77
	3	7	10	72–80
	4	9	10	50–62
	5	6	10	48–56
5	1	17	20	49–62
	2	6	9	48–58
	3	13	15	53–64
	4	9	11	60–70
	5	4	6	53–68
6	1	21	25	46–58
	2	7	10	52–67
	3	9	12	56–67
	4	5	6	57–67
	5	8	11	63–72
7	1	16	17	42–53
	2	7	8	41–53
	3	6	8	42–56
	4	6	8	48–59
	5	17	23	59–69
8	1	7	10	33–43
	2	24	32	44–57
	3	19	22	54–62
	4	4	6	33–43

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 and Harnisch, 1981). In this procedure, item parameters (discrimination, location, and guessing) and the scale score (θ) for each examinee were estimated for the 3PL model, or the 2PPC model in the case of CR items. The item parameters were based on data from the total population of examinees. Then the population was divided into NRC, gender, or ethnic groups, and the members in each group are sorted into ten equal score categories (deciles) based upon their location on the scale score (θ) scale. The expected proportion correct for each group, based on the model prediction, is compared to the observed (actual) proportion correct obtained by the group.

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

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

where

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

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

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

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

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

where

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

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

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

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

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

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

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

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

The following groups were analyzed using the IRT-based DIF analysis: Female, Male, Asian, Black, Hispanic, White, High Needs districts (by NRC code), Low Needs districts (by NRC code), Spanish language test version, and 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 one item was flagged for DIF on the Grade 3 test; four items were flagged on the Grade 4 test; one item was flagged on the Grade 5 test; two items were flagged on the Grade 6 test; three items were flagged on the Grade 7 test; and eight items were flagged on the Grade 8 test, as is shown in Table 37.

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

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

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

Section VII: Reliability and Standard Error of Measurement

This section presents specific information on various test reliability statistics (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 38 are the case counts (N-count), number of test items (# Items), Cronbach's alpha and associated SEM, and Feldt-Raju coefficient and associated SEM obtained for the total mathematics tests.

Table 38. Reliability and Standard Error of Measurement

Grade	N-count	# Items	# RS Points	Cronbach's Alpha	SEM of Cronbach	Feldt-Raju	SEM of Feldt-Raju
3	198574	46	54	0.90	2.99	0.91	2.86
4	199134	57	73	0.94	3.60	0.94	3.41
5	202346	49	61	0.91	3.40	0.92	3.27
6	200076	50	64	0.93	3.62	0.94	3.40
7	202109	52	64	0.92	3.72	0.93	3.51
8	203235	54	70	0.93	4.05	0.94	3.81

All the coefficients for total test reliability were in the range 0.90–0.94, which indicated high internal consistency. As expected, the lowest reliabilities were found for the shortest tests (Grades 3 and 5) 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 39 presents reliabilities for the MC subsets.

Table 39. Reliability and Standard Error of Measurement—MC Items Only

Grade	N-count	# Items	Cronbach's Alpha	SEM of Cronbach	Feldt-Raju	SEM of Feldt-Raju
3	198574	40	0.88	2.36	0.89	2.34
4	199134	45	0.91	2.55	0.91	2.51
5	202346	41	0.89	2.56	0.89	2.53
6	200076	40	0.90	2.52	0.90	2.48
7	202109	44	0.90	2.73	0.90	2.71
8	203235	42	0.91	2.73	0.91	2.72

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 6–12 CR items depending on grade level. The results are presented in Table 40.

Table 40. Reliability and Standard Error of Measurement—CR Items Only

Grade	N-count	# Items	# RS Points	Cronbach's Alpha	SEM of Cronbach	Feldt-Raju	SEM of Feldt-Raju
3	198574	6	14	0.74	1.67	0.75	1.64
4	199134	12	28	0.87	2.30	0.88	2.27
5	202346	8	20	0.78	2.07	0.78	2.05
6	200076	10	24	0.84	2.36	0.85	2.31
7	202109	8	20	0.82	2.25	0.83	2.21
8	203235	12	28	0.87	2.72	0.88	2.66

Note: Results should be interpreted with caution for Grades 3, 5, 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 others. Accommodations available to English language learners are: Time Extension, Separate Location, Bilingual Dictionaries and Glossaries, Translated Edition, Oral Translation, and Responses Written in Native Language. In addition, reliability coefficients were computed for the following subgroups of English

language learners: students taking the English version of the mathematics test and students taking the mathematics tests in each of the five translated languages (Chinese, Haitian Creole, Korean, Russian, and Spanish). As shown in Tables 41a–41f, the estimated reliabilities for subgroups were close in magnitude to the test reliability estimates of the population. Cronbach's alpha reliability coefficients across subgroups were equal to or greater than 0.81. 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 41a. Grade 3 Test Reliability by Subgroup

Group	Subgroup	N-count	Cronbach's Alpha	SEM of Cronbach	Feldt-Raju	SEM of Feldt-Raju
State	All Students	198574	0.90	2.99	0.91	2.86
Gender	Female	97071	0.90	2.99	0.91	2.86
	Male	101503	0.91	2.99	0.92	2.86
Ethnicity	Asian	16508	0.90	2.69	0.91	2.58
	Black	36722	0.90	3.15	0.91	3.02
	Hispanic	46778	0.90	3.10	0.91	2.98
	American Indian	1102	0.91	3.08	0.91	2.95
	Multi-Racial	1568	0.91	3.00	0.92	2.87
	Unknown	287	0.91	2.92	0.92	2.78
	White	95609	0.89	2.89	0.90	2.77
NRC	New York City	72337	0.91	3.03	0.92	2.89
	Big 4 Cites	8258	0.91	3.25	0.91	3.12
	High Needs Urban/Suburban	15541	0.90	3.10	0.91	2.98
	High Needs Rural	11168	0.89	3.05	0.90	2.93
	Average Needs	57719	0.89	2.93	0.90	2.82
	Low Needs	28059	0.87	2.74	0.88	2.63
	Charter	4975	0.87	2.93	0.88	2.83
SWD	All Codes	28117	0.91	3.25	0.92	3.12
SUA	All Codes	50021	0.91	3.21	0.92	3.07
SWD/ SUA	SUA=504 Plan Codes	24535	0.90	3.26	0.91	3.14
ELL/ SUA	SUA=ELL Codes	17932	0.90	3.21	0.91	3.09
ELL	English	16520	0.90	3.20	0.91	3.08
	Chinese	588	0.90	2.94	0.90	2.82
	Haitian Creole	85	0.90	3.36	0.91	3.23
	Korean	48	0.83	2.58	0.84	2.45
	Russian	80	0.92	3.19	0.93	3.03
	Spanish	3238	0.90	3.29	0.91	3.15
	All Translations	4039	0.91	3.25	0.92	3.11

Table 41b. Grade 4 Test Reliability by Subgroup

Group	Subgroup	N-count	Cronbach's Alpha	SEM of Cronbach	Feldt-Raju	SEM of Feldt-Raju
State	All Students	199134	0.94	3.60	0.94	3.41
Gender	Female	97365	0.93	3.59	0.94	3.41
	Male	101769	0.94	3.59	0.95	3.39
Ethnicity	Asian	16046	0.93	3.16	0.94	3.00
	Black	37063	0.93	3.80	0.94	3.60
	Hispanic	45981	0.93	3.75	0.94	3.56
	American Indian	964	0.93	3.76	0.94	3.56
	Multi-Racial	1385	0.94	3.59	0.94	3.40
	Unknown	271	0.94	3.59	0.95	3.39
	White	97424	0.93	3.46	0.94	3.30
NRC	New York City	71483	0.94	3.63	0.95	3.43
	Big 4 Cites	8425	0.93	3.90	0.94	3.68
	High Needs Urban/Suburban	15519	0.93	3.76	0.94	3.56
	High Needs Rural	11554	0.93	3.72	0.93	3.54
	Average Needs	59129	0.93	3.54	0.93	3.37
	Low Needs	28554	0.92	3.25	0.93	3.11
	Charter	3909	0.92	3.58	0.92	3.43
SWD	All Codes	29918	0.93	3.89	0.94	3.68
SUA	All Codes	50961	0.94	3.87	0.94	3.65
SWD/ SUA	SUA=504 Plan Codes	27044	0.93	3.90	0.94	3.69
ELL/ SUA	SUA=ELL Codes	16311	0.93	3.90	0.94	3.69
ELL	English	14848	0.93	3.89	0.94	3.68
	Chinese	578	0.92	3.51	0.93	3.32
	Haitian Creole	88	0.94	3.95	0.94	3.69
	Korean	52	0.92	3.11	0.93	2.93
	Russian	83	0.95	3.87	0.96	3.58
	Spanish	3049	0.93	3.95	0.94	3.71
	All Translations	3850	0.94	3.92	0.95	3.67

Table 41c. Grade 5 Test Reliability by Subgroup

Group	Subgroup	N-count	Cronbach's Alpha	SEM of Cronbach	Feldt-Raju	SEM of Feldt-Raju
State	All Students	202346	0.91	3.40	0.92	3.27
Gender	Female	98841	0.91	3.41	0.91	3.28
	Male	103505	0.92	3.39	0.92	3.25
Ethnicity	Asian	17190	0.91	3.07	0.92	2.93
	Black	37491	0.90	3.55	0.91	3.42
	Hispanic	45177	0.90	3.52	0.91	3.38
	American Indian	978	0.90	3.55	0.91	3.41
	Multi-Racial	1319	0.91	3.40	0.92	3.26
	Unknown	262	0.93	3.39	0.94	3.19
	White	99929	0.90	3.31	0.91	3.20
NRC	New York City	70764	0.92	3.43	0.92	3.27
	Big 4 Cites	8053	0.90	3.62	0.91	3.49
	High Needs Urban/Suburban	15177	0.90	3.52	0.91	3.39
	High Needs Rural	11553	0.90	3.50	0.90	3.38
	Average Needs	60794	0.90	3.36	0.91	3.25
	Low Needs	30178	0.89	3.15	0.90	3.05
	Charter	5223	0.89	3.44	0.90	3.33
SWD	All Codes	30737	0.90	3.61	0.90	3.48
SUA	All Codes	50430	0.91	3.59	0.91	3.45
SWD/ SUA	SUA=504 Plan Codes	27974	0.89	3.61	0.90	3.48
ELL/ SUA	SUA=ELL Codes	13904	0.90	3.61	0.91	3.47
ELL	English	12450	0.90	3.62	0.91	3.48
	Chinese	619	0.89	3.21	0.90	3.10
	Haitian Creole	94	0.89	3.64	0.90	3.47
	Korean	48	0.81	2.95	0.82	2.88
	Russian	77	0.93	3.54	0.94	3.35
	Spanish	3147	0.89	3.62	0.90	3.48
	All Translations	3985	0.92	3.60	0.92	3.44

Table 41d. Grade 6 Test Reliability by Subgroup

Group	Subgroup	N-count	Cronbach's Alpha	SEM of Cronbach	Feldt-Raju	SEM of Feldt-Raju
State	All Students	200076	0.93	3.62	0.94	3.40
Gender	Female	97366	0.92	3.59	0.93	3.38
	Male	102710	0.93	3.63	0.94	3.40
Ethnicity	Asian	15718	0.93	3.18	0.94	2.98
	Black	37827	0.92	3.78	0.92	3.57
	Hispanic	44309	0.92	3.76	0.93	3.54
	American Indian	937	0.92	3.75	0.93	3.55
	Multi-Racial	1259	0.93	3.61	0.94	3.38
	Unknown	273	0.93	3.59	0.94	3.35
	White	99753	0.92	3.49	0.93	3.30
NRC	New York City	69038	0.93	3.68	0.94	3.43
	Big 4 Cites	7869	0.91	3.82	0.92	3.62
	High Needs Urban/Suburban	14787	0.91	3.76	0.92	3.56
	High Needs Rural	11668	0.91	3.69	0.92	3.50
	Average Needs	61101	0.92	3.56	0.92	3.37
	Low Needs	30113	0.91	3.28	0.92	3.11
	Charter	4856	0.91	3.61	0.92	3.42
SWD	All Codes	30353	0.91	3.81	0.92	3.60
SUA	All Codes	46024	0.92	3.82	0.93	3.61
SWD/ SUA	SUA=504 Plan Codes	27442	0.90	3.81	0.91	3.61
ELL/ SUA	SUA=ELL Codes	11747	0.91	3.84	0.92	3.63
ELL	English	10281	0.91	3.84	0.92	3.63
	Chinese	731	0.91	3.48	0.92	3.30
	Haitian Creole	165	0.92	3.79	0.93	3.56
	Korean	58	0.92	3.09	0.93	2.90
	Russian	86	0.93	3.86	0.94	3.59
	Spanish	3378	0.91	3.84	0.92	3.62
	All Translations	4418	0.93	3.84	0.94	3.59

Table 41e. Grade 7 Test Reliability by Subgroup

Group	Subgroup	N-count	Cronbach's Alpha	SEM of Cronbach	Feldt-Raju	SEM of Feldt-Raju
State	All Students	202109	0.92	3.72	0.93	3.51
Gender	Female	98985	0.92	3.72	0.93	3.51
	Male	103124	0.92	3.72	0.93	3.50
Ethnicity	Asian	15874	0.93	3.36	0.94	3.15
	Black	38113	0.90	3.84	0.91	3.66
	Hispanic	44029	0.90	3.85	0.91	3.66
	American Indian	984	0.91	3.80	0.92	3.60
	Multi-Racial	1113	0.92	3.69	0.93	3.47
	Unknown	285	0.93	3.77	0.94	3.53
	White	101711	0.91	3.60	0.92	3.41
NRC	New York City	70022	0.92	3.78	0.93	3.55
	Big 4 Cites	7633	0.90	3.83	0.91	3.64
	High Needs Urban/Suburban	14675	0.90	3.82	0.91	3.65
	High Needs Rural	11778	0.90	3.77	0.91	3.60
	Average Needs	61250	0.91	3.64	0.92	3.47
	Low Needs	32151	0.91	3.44	0.91	3.27
	Charter	3734	0.90	3.77	0.91	3.60
SWD	All Codes	30697	0.89	3.83	0.90	3.65
SUA	All Codes	44768	0.90	3.86	0.91	3.67
SWD/ SUA	SUA=504 Plan Codes	27601	0.88	3.83	0.89	3.65
ELL/ SUA	SUA=ELL Codes	10788	0.90	3.87	0.91	3.68
ELL	English	9102	0.88	3.86	0.90	3.68
	Chinese	852	0.92	3.59	0.93	3.38
	Haitian Creole	130	0.89	3.80	0.91	3.55
	Korean	76	0.92	3.46	0.93	3.23
	Russian	102	0.90	3.90	0.91	3.67
	Spanish	3290	0.87	3.87	0.88	3.70
	All Translations	4450	0.91	3.87	0.92	3.66

Table 41f. Grade 8 Test Reliability by Subgroup

Group	Subgroup	N-count	Cronbach's Alpha	SEM of Cronbach	Feldt-Raju	SEM of Feldt-Raju
State	All Students	203235	0.93	4.05	0.94	3.81
Gender	Female	99215	0.93	4.05	0.94	3.82
	Male	104020	0.94	4.04	0.94	3.79
Ethnicity	Asian	16436	0.94	3.69	0.95	3.45
	Black	38118	0.92	4.04	0.93	3.86
	Hispanic	43715	0.92	4.07	0.93	3.87
	American Indian	988	0.93	4.05	0.94	3.84
	Multi-Racial	980	0.93	4.02	0.94	3.78
	Unknown	265	0.94	4.00	0.95	3.71
	White	102733	0.93	4.01	0.93	3.80
NRC	New York City	71338	0.94	4.05	0.95	3.79
	Big 4 Cites	7651	0.91	3.89	0.92	3.74
	High Needs Urban/Suburban	14374	0.92	4.06	0.92	3.88
	High Needs Rural	11674	0.91	4.10	0.92	3.92
	Average Needs	61742	0.92	4.04	0.93	3.84
	Low Needs	32482	0.92	3.86	0.93	3.68
	Charter	2854	0.92	4.01	0.93	3.82
SWD	All Codes	29949	0.90	3.87	0.91	3.74
SUA	All Codes	43618	0.92	3.93	0.93	3.78
SWD/ SUA	SUA=504 Plan Codes	26931	0.90	3.86	0.91	3.74
ELL/ SUA	SUA=ELL Codes	11080	0.93	3.92	0.94	3.76
ELL	English	8818	0.92	3.88	0.93	3.73
	Chinese	1070	0.93	3.79	0.94	3.58
	Haitian Creole	173	0.90	3.96	0.90	3.83
	Korean	51	0.94	3.73	0.94	3.54
	Russian	85	0.92	3.98	0.93	3.77
	Spanish	3351	0.91	3.89	0.91	3.77
	All Translations	4730	0.94	3.96	0.95	3.75

Standard Error of Measurement

SEMs, as computed from Cronbach's alpha and the Feldt-Raju reliability statistics, are presented in Table 38. SEMs based on Cronbach's alpha ranged 2.99–4.05, 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 ± 3 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.

SEMs for subpopulations, as computed from Cronbach's alpha and the Feldt-Raju reliability statistics, are presented in Tables 41a–41f. SEMs associated with all reliability estimates for all subpopulations are in the range 2.45–4.10, 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 and two independent administrations of the same test (or two parallel forms of the test). Because obtaining test scores from two independent administrations of New York State tests was not feasible due to item release after each administration, a psychometric model was used to obtain the estimated classification consistency indices using test scores from a single administration. Classification accuracy can be defined as the agreement between the actual classifications using observed cut scores and true classifications based on known true cut scores (Livingston and Lewis, 1995).

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

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

Classification consistency and accuracy were estimated using the IRT procedure suggested by Lee, Hanson, and Brennan (2002) and Wang, Kolen, and Harris (2000) and implemented by CTB/McGraw-Hill proprietary software WLCLASS (Kim, 2004). Appendix F includes a description of the calculations and procedure based on the paper by Lee et al. (2002).

Consistency

The results for classifying students into four performance levels are separated from results based solely on the Level III cut. Included in Tables 42 and 43 are case counts (N-count), classification consistency (Agreement), classification inconsistency (Inconsistency), and Cohen's kappa (Kappa). Consistency indicates the rate that a second administration would yield the same performance category designation (or a different designation for the inconsistency rate). The agreement index is a sum of the diagonal element in the contingency table. The inconsistency index is equal to the "1 – agreement index." Kappa is a measure of agreement corrected for chance.

Table 42 depicts the consistency study results based on the range of performance levels for all grades. Overall, between 74% 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 0.62–0.70.

Table 42. Decision Consistency (All Cuts)

Grade	N-count	Agreement	Inconsistency	Kappa
3	198549	0.74	0.26	0.62
4	201418	0.79	0.21	0.70
5	199254	0.76	0.24	0.65
6	200415	0.77	0.23	0.68
7	202359	0.76	0.24	0.66
8	206346	0.78	0.22	0.69

Table 43 depicts the consistency study results based on two performance levels (passing and not passing) as defined by the Level III cut. Overall, about 88%–92% 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 0.75–0.82.

Table 43. Decision Consistency (Level III Cut)

Grade	N-count	Agreement	Inconsistency	Kappa
3	198549	0.88	0.12	0.75
4	201418	0.92	0.08	0.82
5	199254	0.90	0.10	0.77
6	200415	0.91	0.09	0.80
7	202359	0.90	0.10	0.78
8	206346	0.91	0.09	0.82

Accuracy

The results of classification accuracy are presented in Table 44. Included in the table are case counts (N-count), classification accuracy (Accuracy) for all performance levels (All Cuts) and for the Level III cut score, 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 81%–85% of the time across all performance levels and approximately 91%–94% of the time in regards to the Level III cut score.

Table 44. Decision Agreement (Accuracy)

Grade	N-count	Accuracy					
		All Cuts	False Positive (All Cuts)	False Negative (All Cuts)	**Level III Cut**	False Positive (Level III Cut)	False Negative (Level III Cut)
3	198549	**0.81**	0.12	0.07	**0.91**	0.06	0.03
4	201418	**0.85**	0.11	0.05	**0.94**	0.04	0.02
5	199254	**0.82**	0.11	0.06	**0.93**	0.04	0.03
6	200415	**0.84**	0.10	0.06	**0.93**	0.04	0.02
7	202359	**0.83**	0.10	0.07	**0.93**	0.04	0.03
8	206346	**0.85**	0.08	0.07	**0.94**	0.03	0.03

Section VIII: Summary of Operational Test Results

This section summarizes the distribution of OP scale score results on the New York State 2011 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 statuses are true and who use one or more ELL-related accommodations. 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 G.

Scale Score Distribution Summary

Scale score distribution summaries are presented and discussed in Table 45. First, scale score statistics for total populations of students from public and charter schools are presented. Next, scale score statistics are presented for selected subgroups in each grade level. The statistics for groups with small number counts should be interpreted with caution. Some general observations: Females and Males had very similar achievement patterns; Asian and White students outperformed their peers from other ethnic groups; Low Needs and Average Needs schools (as identified by NRC) outperformed other school types (New York City, Big 4 Cities, Urban/Suburban, Rural, and Charter); students taking the Chinese and Korean translations met or exceeded the population at every reported percentile, whereas the other translation subgroups (Haitian Creole, Spanish, and Russian) were below the population scale score at each percentile; 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 G.

Table 45. 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	198574	686.66	20.98	663	675	687	699	708
4	199134	687.96	32.21	648	669	689	709	726
5	202346	686.12	30.49	650	669	687	705	722
6	200076	682.16	31.60	644	665	683	701	719
7	202109	678.66	30.59	643	662	680	698	713
8	203235	677.25	33.66	641	661	680	697	713

Grade 3

Scale score statistics and N-counts of demographic groups for Grade 3 are presented in Table 46. The population scale score mean was 686.66 with a standard deviation of 20.98. The gender subgroups performed the same, with a mean difference of 0.42 scale score points. Asian, Multi-Racial, and White ethnic subgroups had scale score means that exceeded the State mean scale score on the test, as did students from Low Needs and Average Needs districts and the Charter schools. The lowest performing NRC subgroup was the Big 4 Cities,

with a mean of 672.70, and the lowest performing ethnic subgroup was Black (mean scale score of 678.94). 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 669 (Haitian Creole subgroup) to 699 (Korean subgroup), a difference that exceeds a standard deviation. The subgroup that used the Haitian Creole translation had a scale score mean of 23 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), Average Needs (689), Low Needs (695), Charter schools (689), and students who used the Chinese (689) and Korean (699) translations.

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

Demographic Category (Subgroup)		N-count	SS Mean	SS Std Dev	10th %tile	25th %tile	50th %tile	75th %tile	90th %tile
State	All Students	198574	686.66	20.98	663	675	687	699	708
Gender	Female	97071	686.45	20.02	664	675	687	697	708
	Male	101503	686.87	21.84	663	675	687	699	711
Ethnicity	Asian	16508	696.93	21.56	675	686	697	708	722
	Black	36722	678.94	20.99	656	669	681	691	702
	Hispanic	46778	680.85	19.91	658	670	682	693	702
	American Indian	1102	682.07	22.50	660	672	684	695	705
	Multi-Racial	1568	686.83	21.14	663	675	687	699	711
	Unknown	287	688.65	24.66	664	676	689	702	711
	White	95609	690.75	19.51	669	681	691	702	711
NRC	New York City	72337	684.77	21.55	661	673	686	697	708
	Big 4 Cites	8258	672.70	24.64	648	661	675	687	697
	High Needs Urban/Suburban	15541	681.33	19.16	660	670	682	693	702
	High Needs Rural	11168	683.16	18.04	661	673	684	695	705
	Average Needs	57719	688.82	19.12	668	678	689	699	711
	Low Needs	28059	696.00	18.79	675	686	695	705	716
	Charter	4975	688.40	16.39	669	678	689	699	708
SWD	All Codes	28117	670.78	25.30	645	660	673	686	695
SUA	All Codes	50021	674.83	23.62	650	664	676	689	699
SWD/SUA	SUA=504 Plan Codes	24535	669.63	24.76	645	658	672	684	695
ELL/SUA	SUA=ELL Codes	17932	674.65	22.13	652	664	676	687	697
ELL	English	16520	674.94	21.54	652	664	676	687	697
	Chinese	588	688.59	17.59	664	678	689	699	711
	Haitian Creole	85	663.09	36.14	640	656	669	684	691
	Korean	48	698.71	13.96	679	691	699	708	722
	Russian	80	673.89	38.41	649	664	680	693	699
	Spanish	3238	670.09	24.24	648	660	672	684	693
	All Translations	4039	673.05	25.03	648	663	675	687	697

Grade 4

Scale score statistics and N-counts of demographic groups for Grade 4 are presented in Table 47. The population scale score mean was 687.96 with a standard deviation of 32.21. 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 more than two-thirds 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 664.79, more than two-thirds 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 689: Asian (709), White (697), Average Needs (692), Low Needs (703), and students who used the Chinese (699) and Korean (714) translations.

Table 47. Scale Score Distribution Summary, by Subgroup, Grade 4

Demographic Category (Subgroup)		N-count	SS Mean	SS Std Dev	10th %tile	25th %tile	50th %tile	75th %tile	90th %tile
State	All Students	199134	687.96	32.21	648	669	689	709	726
Gender	Female	97365	688.13	31.30	650	669	689	709	726
	Male	101769	687.79	33.05	646	667	689	709	726
Ethnicity	Asian	16046	707.31	32.86	669	689	709	726	745
	Black	37063	674.13	30.64	637	655	675	694	712
	Hispanic	45981	678.58	30.47	642	661	680	699	715
	American Indian	964	680.78	30.22	642	662	681	701	718
	Multi-Racial	1385	689.18	32.38	650	667	689	712	726
	Unknown	271	688.94	32.00	648	667	689	712	726
	White	97424	694.50	30.01	657	677	697	715	731
NRC	New York City	71483	685.89	33.73	644	665	687	709	726
	Big 4 Cites	8425	664.79	32.42	623	646	665	685	703
	High Needs Urban/Suburban	15519	677.45	29.98	639	659	678	697	715
	High Needs Rural	11554	681.60	28.39	646	664	683	701	715
	Average Needs	59129	691.05	29.13	655	673	692	709	726
	Low Needs	28554	702.89	28.86	669	685	703	722	737
	Charter	3909	688.16	26.29	655	672	689	706	722

(Continued on next page)

Table 47. Scale Score Distribution Summary, by Subgroup, Grade 4 (cont.)

Demographic Category (Subgroup)		N-count	SS Mean	SS Std Dev	10th %tile	25th %tile	50th %tile	75th %tile	90th %tile
SWD	All Codes	29918	659.49	32.66	616	639	661	682	699
SUA	All Codes	50961	666.75	33.10	623	646	669	689	706
SWD/SUA	SUA=504 Plan Codes	27044	657.89	31.99	616	637	661	680	697
ELL/SUA	SUA=ELL Codes	16311	666.96	31.47	626	648	669	687	703
ELL	English	14848	666.98	30.75	626	650	669	687	703
	Chinese	578	698.41	30.85	662	680	699	715	737
	Haitian Creole	88	653.85	34.94	604	632	658	680	697
	Korean	52	710.69	29.73	672	694	714	731	745
	Russian	83	671.95	42.39	629	648	677	699	722
	Spanish	3049	657.55	32.01	616	637	661	680	694
	All Translations	3850	664.63	35.71	620	642	665	687	706

Grade 5

Grade 5 demographic group N-counts and scale score statistics are presented in Table 48. The population scale score mean was 686.12 with a standard deviation of 30.49. 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 18 scale score points. American Indian, Black, and Hispanic ethnic subgroups had mean scale scores approximately one standard deviation below the Asian subgroup. Students from Low Needs and Average Needs districts outperformed the other NRC subgroups. The lowest performing NRC subgroup was the Big 4 Cities, with a mean of 662.72, nearly one-half of a standard deviation below the second lowest performing NRC subgroup (High Needs/Urban/Suburban: 676.72) and 38 scale score units below the Low Needs subgroup mean. SWD, SUA, and ELL without testing in alternate language subgroups scored consistently below the Statewide percentile scale score rankings. The Haitian Creole translation subgroup, which had a scale score mean (652.03) of more than 34 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 708.25, about two-thirds of standard deviation above the population mean. At the 50th percentile, the following groups exceeded the population scale score of 687: Asian (705), White (693), Average Needs (689), Low Needs (700), and students who used the Chinese (698) and Korean (705) translations.

Table 48. Scale Score Distribution Summary, by Subgroup, Grade 5

Demographic Category (Subgroup)		N-count	SS Mean	SS Std Dev	10th %tile	25th %tile	50th %tile	75th %tile	90th %tile
State	All Students	202346	686.12	30.49	650	669	687	705	722
Gender	Female	98841	686.39	29.47	652	669	687	705	722
	Male	103505	685.87	31.43	650	669	687	705	722
Ethnicity	Asian	17190	704.63	31.52	669	687	705	722	739
	Black	37491	673.37	29.89	638	657	675	691	708
	Hispanic	45177	677.40	29.60	641	661	679	696	711
	American Indian	978	676.39	29.62	641	659	678	696	711
	Multi-Racial	1319	686.63	30.85	650	669	687	705	722
	Unknown	262	686.47	37.65	644	667	689	711	727
	White	99929	691.75	27.71	659	677	693	708	722
NRC	New York City	70764	684.78	32.06	647	665	685	705	722
	Big 4 Cites	8053	662.72	34.02	626	644	665	683	700
	High Needs Urban/Suburban	15177	676.72	28.77	644	661	679	696	711
	High Needs Rural	11553	678.88	27.78	647	663	681	696	711
	Average Needs	60794	688.53	27.17	657	673	689	705	722
	Low Needs	30178	699.74	26.18	669	685	700	714	732
	Charter	5223	683.68	25.04	652	669	685	700	714
SWD	All Codes	30737	659.46	33.45	621	644	663	681	696
SUA	All Codes	50430	665.51	33.29	626	650	667	687	703
SWD/SUA	SUA=504 Plan Codes	27974	658.47	33.02	621	641	661	679	696
ELL/SUA	SUA=ELL Codes	13904	663.20	33.60	626	647	665	685	700
ELL	English	12450	663.43	32.72	626	647	665	683	700
	Chinese	619	696.04	25.17	665	681	698	714	722
	Haitian Creole	94	652.03	38.18	609	638	657	677	689
	Korean	48	708.25	21.81	683	697	705	722	739
	Russian	77	673.34	35.36	634	652	673	703	714
	Spanish	3147	655.32	35.03	616	638	659	677	693
	All Translations	3985	662.55	37.10	621	641	665	687	705

Grade 6

Grade 6 scale score statistics and N-counts of demographic groups are presented in Table 49. The population scale score mean was 682.16 with a standard deviation of 31.60. 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. 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 659.92. New York City, High Needs

Urban/Suburban, High Needs Rural, and Charter subgroups had similar scale score means (ranging from approximately 672–682). SWD, SUA, and ELL without testing in alternate language subgroups scored consistently below the Statewide percentile scale score rankings. The Haitian Creole translation subgroup, which had a scale score mean (648.67) more than 33 units below the population mean, was the lowest performing group analyzed. Asian students (the highest performing subgroup with a mean of 702.92) exceeded the State mean by over 20 scale score points. At the 50th percentile, the following groups exceeded the population scale score of 683: Asian (703), White (688), Average Needs (685), Low Needs (696), and students who used the Chinese (692) and Korean (706) translations.

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

Demographic Category (Subgroup)		N-count	SS Mean	SS Std Dev	10th %tile	25th %tile	50th %tile	75th %tile	90th %tile
State	All Students	200076	682.16	31.60	644	665	683	701	719
Gender	Female	97366	683.11	30.15	649	667	683	701	719
	Male	102710	681.26	32.89	642	663	681	701	719
Ethnicity	Asian	15718	702.99	32.75	665	683	703	723	743
	Black	37827	668.06	29.94	633	653	670	687	701
	Hispanic	44309	671.65	30.19	636	655	673	690	706
	American Indian	937	672.56	31.97	639	657	675	690	709
	Multi-Racial	1259	683.97	31.73	646	663	683	701	723
	Unknown	273	685.56	33.75	649	667	687	701	728
	White	99753	688.96	28.70	657	673	688	706	723
NRC	New York City	69038	678.28	33.90	639	658	678	699	719
	Big 4 Cites	7869	659.92	30.67	626	644	663	678	694
	High Needs Urban/Suburban	14787	672.00	28.41	639	657	673	690	703
	High Needs Rural	11668	677.25	26.33	646	663	678	694	709
	Average Needs	61101	685.57	27.97	653	670	685	701	719
	Low Needs	30113	698.04	27.89	667	681	696	715	734
	Charter	4856	682.85	25.96	653	667	683	699	715
SWD	All Codes	30353	653.05	32.67	617	636	657	673	688
SUA	All Codes	46024	658.36	33.09	622	642	662	678	694
SWD/SUA	SUA=504 Plan Codes	27442	652.16	32.36	617	636	655	673	687
ELL/SUA	SUA=ELL Codes	11747	656.83	32.99	617	639	660	676	692
ELL	English	10281	655.47	32.27	617	639	658	675	690
	Chinese	731	692.12	27.24	662	676	692	709	723
	Haitian Creole	165	648.67	36.92	611	630	653	672	690
	Korean	58	704.57	28.32	673	687	706	719	743
	Russian	86	659.21	38.96	617	636	665	681	701
	Spanish	3378	653.33	32.93	617	636	658	675	688
	All Translations	4418	660.36	35.74	617	642	663	683	701

Grade 7

N-counts and scale score statistics of demographic groups for Grade 7 are presented in Table 50. The population scale score mean was 678.66 with a standard deviation of 30.59. The gender subgroups performed very similarly, with a mean difference of less than two scale score points. Asian, Multi-Racial, 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 655.40, while the Low Needs subgroup's scale score mean was 693.91. 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 680: Asian (698), Multi-Racial (682), White (687), Average Needs (684), Low Needs (694), and students who used the Chinese (689) and Korean (690) translations.

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

Demographic Category (Subgroup)		N-count	SS Mean	SS Std Dev	10th %tile	25th %tile	50th %tile	75th %tile	90th %tile
State	All Students	202109	678.66	30.59	643	662	680	698	713
Gender	Female	98985	679.54	29.21	646	664	680	698	713
	Male	103124	677.81	31.83	641	660	679	698	713
Ethnicity	Asian	15874	697.47	32.95	660	679	698	716	733
	Black	38113	663.80	29.17	631	648	666	682	696
	Hispanic	44029	667.82	28.78	635	652	669	685	700
	American Indian	984	670.39	31.90	635	656	672	689	704
	Multi-Racial	1113	681.89	30.44	646	664	682	700	716
	Unknown	285	676.57	35.64	638	658	677	696	716
	White	101711	686.03	27.30	654	671	687	702	716
NRC	New York City	70022	674.04	32.39	638	656	674	694	713
	Big 4 Cites	7633	655.40	33.60	618	641	658	676	690
	High Needs Urban/Suburban	14675	667.58	27.61	638	652	669	685	698
	High Needs Rural	11778	673.81	25.15	646	660	676	689	702
	Average Needs	61250	683.28	26.30	652	669	684	700	713
	Low Needs	32151	693.91	26.90	664	679	694	710	726
	Charter	3734	676.84	24.51	646	662	677	692	707
SWD	All Codes	30697	650.34	32.57	618	635	654	671	684
SUA	All Codes	44768	654.97	32.65	618	641	658	674	690
SWD/SUA	SUA=504 Plan Codes	27601	649.86	32.10	618	635	652	669	684

(Continued on next page)

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

Demographic Category (Subgroup)		N-count	SS Mean	SS Std Dev	10th %tile	25th %tile	50th %tile	75th %tile	90th %tile
ELL/SUA	SUA=ELL Codes	10788	652.27	33.46	618	638	654	672	689
ELL	English	9102	649.93	32.85	618	635	652	669	684
	Chinese	852	686.15	26.91	652	671	689	704	716
	Haitian Creole	130	640.29	36.61	604	618	646	666	682
	Korean	76	691.70	31.27	664	679	690	713	726
	Russian	102	658.54	31.40	623	643	662	679	690
	Spanish	3290	648.15	32.75	612	635	652	669	682
	All Translations	4450	656.18	35.47	618	638	658	677	696

Grade 8

Grade 8 scale score statistics and N-counts of demographic groups are presented in Table 51. The population scale score mean was 677.25 with a standard deviation of 33.66. The gender subgroups performed similarly, with a mean difference of less than 4 scale score points. Asian, Multi-Racial, and White ethnic subgroups' scale score means exceeded the State mean scale score. The Black, Hispanic, and American Indian ethnic subgroups' scale score means were all close to or more than 10 scale score points below the population mean. The lowest performing NRC subgroup, Big 4 Cities, had a scale score mean of 647.98, while the Low Needs subgroup's scale score mean was 692.31, which indicated a large performance discrepancy by school district NRC designation. SWD, SUA, and ELL without testing in 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 680: Female (682), Asian (701), Multi-Racial (684), White (685), Average Needs (684), Low Needs (693), and students who used the Chinese (696) and Korean (699) translations.

Table 51. Scale Score Distribution Summary, by Subgroup, Grade 8

Demographic Category (Subgroup)		N-count	SS Mean	SS Std Dev	10th %tile	25th %tile	50th %tile	75th %tile	90th %tile
State	All Students	203235	677.25	33.66	641	661	680	697	713
Gender	Female	99215	678.96	31.90	647	665	682	697	713
	Male	104020	675.61	35.17	641	660	679	696	713
Ethnicity	Asian	16436	698.51	32.09	665	682	701	718	737
	Black	38118	662.04	36.18	629	650	666	683	697
	Hispanic	43715	667.12	33.67	634	654	670	687	701
	American Indian	988	666.50	35.40	634	652	670	687	701
	Multi-Racial	980	679.83	33.67	644	666	684	699	713
	Unknown	265	679.85	39.17	647	663	680	704	718
	White	102733	683.87	28.88	654	670	685	701	715

(Continued on next page)

Table 51. Scale Score Distribution Summary, by Subgroup, Grade 8 (cont.)

Demographic Category (Subgroup)		N-count	SS Mean	SS Std Dev	10th %tile	25th %tile	50th %tile	75th %tile	90th %tile
SWD	All Codes	29949	646.45	42.78	606	634	654	672	684
SUA	All Codes	43618	653.04	41.45	616	641	660	677	692
SWD/SUA	SUA=504 Plan Codes	26931	645.78	42.56	606	634	654	670	683
ELL/SUA	SUA =ELL Codes	11080	656.15	40.66	616	641	661	679	696
ELL	English	8818	651.22	41.89	616	638	658	676	689
	Chinese	1070	694.20	28.87	661	680	696	713	725
	Haitian Creole	173	652.46	36.81	616	641	658	674	685
	Korean	51	691.16	40.44	674	684	699	710	721
	Russian	85	659.84	40.73	629	647	668	684	694
	Spanish	3351	650.51	40.28	616	638	658	674	687
	All Translations	4730	661.07	42.11	616	644	666	685	704

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 2010–2011 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.

Due to the re-configuration of the tests in 2011 and the lengthening of the mathematics tests, further small adjustments to the cut score values established after the 2010 administration were made after the 2011 test administration. Instead of implementing the "theoretical" cut score values established in 2010 (shown in columns 1–3 of Table 52), the minimum scale score values required to be classified in Levels II, III, and IV in 2010 (shown in columns 4–6 of Table 52) were adopted as definitive test cut scores (i.e., new "theoretical cuts") for 2011 and subsequent test administrations. These values are equal to or greater than theoretical cut scores established in 2010. This approach produced proficiency levels in 2011 consistent with the proficiency levels established and used in 2010 while preserving the impact data between the two years. This approach was endorsed by New York State Technical Advisory Group. Details and rationale for this cut score adjustment were described in the July 5, 2011, memorandum from CTB/McGraw-Hill to NYSED "NYS cut score implementation options for the 2011 ELA and Mathematics operational tests."

Table 52 shows the mathematics cut scores used for classification of students into the four performance levels in 2011.

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

Content	Grade	2010 NYS cut scores "Theoretical Cuts"			2011 NYS cut scores (Minimum scale score in each proficiency level in 2010 "Operational Cuts")		
		Level			Level		
		II	III	IV	II	III	IV
	Column	1	2	3	4	5	6
Math	3	661	684	707	**662**	**684**	**707**
	4	636	676	707	**636**	**676**	**707**
	5	640	674	702	**640**	**676**	**707**
	6	640	674	699	**640**	**674**	**700**
	7	639	670	694	**639**	**670**	**694**
	8	639	673	702	**639**	**674**	**704**

Tables 53–59 show the performance level distributions for all examinees from public and charter schools with valid scores. Table 53 presents performance level data for total populations of students in Grades 3–8. Tables 54–59 contain performance level data for selected subgroups of students. In general, these summaries reflect the same achievement trends 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 who used the Korean or Chinese translations outperformed other test translation subgroups. The Level III and above rates for SWD and SUA subgroups were low compared to the total population of examinees. Across grades, the following subgroups consistently performed above the population average: Asian, White, Average Needs, Low Needs, Chinese translation, and Korean translation. Please note that the case counts for the Haitian Creole, Korean, and Russian translation subgroups were very low, and the results might have been heavily influenced by very high and/or very low achieving individual students.

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

Grade	N-count	Percent of New York State Population in Performance Level				
		Level I	Level II	Level III	Level IV	Levels III & IV
3	198574	9.09	31.22	46.27	13.43	59.70
4	199134	5.54	27.73	39.99	26.74	66.73
5	202346	5.75	27.89	42.85	23.51	66.36
6	200076	7.91	28.99	36.72	26.38	63.10
7	202109	7.86	27.44	34.24	30.46	64.70
8	203235	8.63	31.39	42.29	17.69	59.98

Grade 3

Performance level summaries and N-counts of demographic groups for Grade 3 are presented in Table 54. Statewide, 59.70% of third-graders were in Levels III and IV. American Indian, Black, Hispanic, and Multi-Racial subgroups had a lower percentage of students in Levels III and IV than the rest of the population, but the percentage of Asian and White ethnic subgroups in Levels III and IV exceeded the overall State population. Student achievement

varied widely by NRC subgroup as well. Over 79% of students from Low Needs districts were classified in Levels III and IV, whereas only about 31% of Big 4 Cities students were in Levels III and IV. Less than 40% of SWD, SUA, or those who used translated test forms were classified in Level III or above; however, the subgroups for Korean and Chinese translations had more than 64% in Levels III and IV, with Korean students having the greatest percentage of more than 87%.

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

Demographic Category (Subgroup)		N-count	Level I %	Level II %	Level III %	Level IV %	Levels III & IV %
State	All Students	198574	9.09	31.22	46.27	13.43	59.70
Gender	Female	97071	8.62	32.07	46.90	12.41	59.30
	Male	101503	9.53	30.40	45.67	14.41	60.07
Ethnicity	Asian	16508	3.99	16.09	50.91	29.01	79.92
	Black	36722	15.71	41.66	36.87	5.76	42.63
	Hispanic	46778	13.07	39.91	40.63	6.38	47.02
	American Indian	1102	11.89	36.84	43.38	7.89	51.27
	Multi-Racial	1568	8.99	32.27	43.75	14.99	58.74
	Unknown	287	8.36	28.57	45.99	17.07	63.07
	White	95609	5.45	25.49	51.91	17.16	69.06
NRC	New York City	72337	11.11	33.98	42.80	12.10	54.91
	Big 4 Cites	8258	26.06	42.35	27.90	3.69	31.59
	High Needs Urban/Suburban	15541	12.70	39.62	40.81	6.87	47.68
	High Needs Rural	11168	10.13	38.25	43.71	7.91	51.62
	Average Needs	57719	5.97	28.94	50.98	14.11	65.09
	Low Needs	28059	2.64	18.19	54.42	24.75	79.17
	Charter	4975	4.72	31.06	52.84	11.38	64.22
SWD	All Codes	28117	28.84	43.01	24.99	3.16	28.15
SUA	All Codes	50021	22.33	42.13	31.17	4.36	35.54
SWD/SUA	SUA=504 Plan Codes	24535	30.50	43.93	23.15	2.43	25.57
ELL/SUA	SUA=ELL Codes	17932	20.98	45.18	30.66	3.17	33.83
ELL	ELL status = Y	20025	21.81	44.87	30.21	3.12	33.32
ELL Test Language	English	16520	20.63	45.21	30.97	3.20	34.16
	Chinese	588	7.82	27.21	50.85	14.12	64.97
	Haitian Creole	85	36.47	37.65	24.71	1.18	25.88
	Korean	48	0.00	12.50	60.42	27.08	87.50
	Russian	80	23.75	30.00	41.25	5.00	46.25
	Spanish	3238	28.13	45.03	25.51	1.33	26.84
	All Translations	4039	24.93	41.59	29.91	3.57	33.47

Grade 4

Performance level summaries and N-counts of demographic groups for Grade 4 are presented in Table 55. Statewide, 66.73% of the fourth-grade population was placed in Levels III and IV. Around 7%–10% of American Indian, Black, and Hispanic students were Level I, as compared to only about 2.37% of Asian students and 3.28% of White students. American

Indian, Black, and Hispanic ethnic subgroups had percentages of students in Levels III and IV ranging from 48%–57%, but the percentages of the Multi-Racial, White, and Asian subgroups meeting standards for Levels III and IV (65.92%, 75.95%, and 85.36%, respectively) exceeded the population. Student achievement also varied widely by NRC subgroup. About 85% 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 subgroups or students 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 (79.24%). The Korean translation subgroup had 86.54% of students in Levels III and IV. The following subgroups had a higher percentage of students meeting standards for Levels III and IV than the State population: Female, Asian, White, Average Needs, Low Needs, Chinese translation, and Korean translation.

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

Demographic Category (Subgroup)		N-count	Level I %	Level II %	Level III %	Level IV %	Levels III & IV %
State	All Students	199134	5.54	27.73	39.99	26.74	66.73
Gender	Female	97365	4.89	27.93	40.95	26.23	67.17
	Male	101769	6.16	27.53	39.08	27.23	66.31
Ethnicity	Asian	16046	2.37	12.26	33.53	51.83	85.36
	Black	37063	9.87	41.55	36.03	12.56	48.58
	Hispanic	45981	7.93	36.51	39.63	15.93	55.56
	American Indian	964	7.26	35.48	39.94	17.32	57.26
	Multi-Racial	1385	4.77	29.31	37.55	28.38	65.92
	Unknown	271	6.27	27.31	37.27	29.15	66.42
	White	97424	3.28	20.77	42.78	33.17	75.95
NRC	New York City	71483	6.45	31.07	37.19	25.29	62.48
	Big 4 Cites	8425	16.99	46.30	28.40	8.31	36.71
	High Needs Urban/Suburban	15519	8.18	38.17	38.77	14.88	53.65
	High Needs Rural	11554	5.87	34.03	42.99	17.11	60.10
	Average Needs	59129	3.60	23.87	44.35	28.18	72.53
	Low Needs	28554	1.76	13.47	40.47	44.30	84.77
	Charter	3909	2.69	27.86	47.17	22.28	69.46
SWD	All Codes	29918	22.04	47.13	25.08	5.76	30.83
SUA	All Codes	50961	16.38	43.63	30.56	9.43	39.99
SWD/SUA	SUA=504 Plan Codes	27044	22.99	48.18	24.23	4.59	28.83
ELL/SUA	SUA=ELL Codes	16311	14.62	45.61	31.70	8.07	39.77
ELL	ELL status = Y	18207	15.53	45.65	31.04	7.79	38.83
ELL Test Language	English	14848	14.23	46.22	31.84	7.71	39.55
	Chinese	578	2.42	18.34	44.46	34.78	79.24
	Haitian Creole	88	29.55	40.91	25.00	4.55	29.55
	Korean	52	1.92	11.54	34.62	51.92	86.54
	Russian	83	15.66	33.73	32.53	18.07	50.60
	Spanish	3049	22.83	46.54	26.63	4.00	30.63
	All Translations	3850	19.48	41.43	29.51	9.58	39.09

Grade 5

Performance level summaries and N-counts of demographic groups for Grade 5 are presented in Table 56. Statewide, 66.36% of the fifth-grade population was placed in Levels III and IV. There was little performance differentiation by gender subgroup, with less than 2% difference between each level. However, across ethnic and test translation subgroups, there were marked differences. American Indian, Black, Hispanic, and Multi-Racial ethnic subgroups were below the State average of students meeting standards for Levels III and IV (ranging 48%–66%), as compared to the percentage of Asian and White students meeting standards for Levels III and IV (85% and 75%, respectively). Over 84% of students from Low Needs districts were in Levels III or IV, but only about 34% of the Big 4 Cities students were in those levels. Only about 5%–8% of SWD or SUA subgroups were placed in Level IV, compared to the population's 23.51% 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 (20.78%) and Chinese and Korean translation subgroups that had very high percentages of students in Level IV (35.22% and 45.83%, 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 56. Performance Level Distribution Summary, by Subgroup, Grade 5

Demographic Category (Subgroup)		N-count	Level I %	Level II %	Level III %	Level IV %	Levels III & IV %
State	All Students	202346	5.75	27.89	42.85	23.51	66.36
Gender	Female	98841	5.12	28.20	43.82	22.86	66.68
	Male	103505	6.35	27.59	41.92	24.14	66.06
Ethnicity	Asian	17190	2.56	11.93	36.42	49.09	85.51
	Black	37491	10.18	40.97	38.44	10.41	48.85
	Hispanic	45177	8.57	36.74	41.06	13.62	54.69
	American Indian	978	8.49	38.34	40.39	12.78	53.17
	Multi-Racial	1319	6.22	27.90	41.47	24.41	65.88
	Unknown	262	8.02	23.66	38.55	29.77	68.32
	White	99929	3.32	21.64	46.47	28.58	75.05
NRC	New York City	70764	6.63	30.27	40.01	23.10	63.11
	Big 4 Cites	8053	19.42	46.45	27.15	6.98	34.12
	High Needs Urban/Suburban	15177	8.15	37.96	41.50	12.39	53.89
	High Needs Rural	11553	6.66	36.19	43.59	13.55	57.15
	Average Needs	60794	3.79	25.18	47.30	23.73	71.03
	Low Needs	30178	1.66	13.66	45.22	39.46	84.68
	Charter	5223	4.33	32.28	46.51	16.89	63.39
SWD	All Codes	30737	21.50	47.88	25.99	4.63	30.62
SUA	All Codes	50430	17.01	44.48	30.97	7.54	38.51
SWD/SUA	SUA=504 Plan Codes	27974	22.07	48.78	25.28	3.87	29.15
ELL/SUA	SUA=ELL Codes	13904	18.67	46.27	28.45	6.62	35.06

(Continued on next page)

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

Demographic Category (Subgroup)		N-count	Level I %	Level II %	Level III %	Level IV %	Levels III & IV %
ELL	ELL status = Y	15786	19.11	46.73	27.83	6.33	34.16
ELL Test Language	English	12450	17.44	47.61	29.03	5.93	34.96
	Chinese	619	2.75	16.32	45.72	35.22	80.94
	Haitian Creole	94	26.60	44.68	25.53	3.19	28.72
	Korean	48	0.00	4.17	50.00	45.83	95.83
	Russian	77	18.18	33.77	27.27	20.78	48.05
	Spanish	3147	26.88	47.22	22.53	3.37	25.90
	All Translations	3985	22.63	41.58	26.62	9.16	35.78

Grade 6

Performance level summaries and N-counts of demographic groups for Grade 6 are presented in Table 57. Statewide, 63.10% of the sixth-grade population was placed in Levels III and IV. There was a slight performance differentiation by gender subgroup with less than 3% difference between each level. There were marked differences across ethnic and test translation subgroups. About 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 43%–52%), as compared to the percentage of Asian and White students meeting standards for Levels III and IV (84.62% and 73.39%, respectively). About 84% of students from Low Needs districts were in Levels III or IV, but only about 32% 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 26.38% in Level IV. 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 who had very high percentages of students in Level IV (37.07% and 58.62%, 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 57. Performance Level Distribution Summary, by Subgroup, Grade 6

Demographic Category (Subgroup)		N-count	Level I %	Level II %	Level III %	Level IV %	Levels III & IV %
State	All Students	200076	7.91	28.99	36.72	26.38	63.10
Gender	Female	97366	6.74	28.66	38.19	26.41	64.60
	Male	102710	9.02	29.31	35.32	26.36	61.68
Ethnicity	Asian	15718	3.02	12.37	30.83	53.79	84.62
	Black	37827	14.44	41.97	32.25	11.35	43.60
	Hispanic	44309	12.36	38.49	34.69	14.46	49.15
	American Indian	937	10.78	37.46	35.86	15.90	51.76
	Multi-Racial	1259	6.59	30.34	36.14	26.93	63.07
	Unknown	273	7.33	25.27	38.83	28.57	67.40
	White	99753	4.22	22.39	40.25	33.14	73.39

(Continued on next page)

Table 57. Performance Level Distribution Summary, by Subgroup, Grade 6 (cont.)

Demographic Category (Subgroup)		N-count	Level I %	Level II %	Level III %	Level IV %	Levels III & IV %
NRC	New York City	69038	10.82	33.01	32.54	23.64	56.17
	Big 4 Cites	7869	21.40	46.21	25.96	6.43	32.39
	High Needs Urban/Suburban	14787	11.22	39.62	35.66	13.51	49.16
	High Needs Rural	11668	7.26	34.87	40.81	17.06	57.87
	Average Needs	61101	4.73	25.94	41.48	27.84	69.33
	Low Needs	30113	2.15	13.76	38.25	45.84	84.09
	Charter	4856	4.06	30.68	41.66	23.60	65.26
SWD	All Codes	30353	29.16	46.63	19.88	4.33	24.21
SUA	All Codes	46024	24.17	44.64	24.12	7.08	31.19
SWD/SUA	SUA=504 Plan Codes	27442	29.94	47.01	19.35	3.71	23.05
ELL/SUA	SUA=ELL Codes	11747	25.78	44.92	22.79	6.51	29.30

Grade 7

Performance level summaries and N-counts of demographic groups for Grade 7 are presented in Table 58. Statewide, 64.70% 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 3% difference between each level. However, there were marked differences across ethnic and test translation subgroups. Black, Hispanic, and American Indian ethnic subgroups had around 42%–56% of students meeting standards for Levels III and IV, with less than 20% of those students in Level IV, whereas over 83% of Asian students were meeting standards for Levels III and IV (and over 57% were in Level IV.) About 33% of Big 4 Cities students were meeting standards for Levels III and IV, with less than 9% in Level IV, yet over 85% of students from Low Needs districts were meeting standards for Levels III and IV (about 52% in Level IV). Less than 9% of SWD and SUA subgroups were placed in Level IV and over 24% 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 who had very high rates (42.49% and 44.74%, respectively). Across all subgroups, the Haitian Creole translation subgroup had the largest percentage of students placed in Level I (42.31%), and the Korean translation subgroup had the largest percentage of students (82.89%) 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, Multi-Racial, White, Average Needs, Low Needs, Chinese translation, and Korean translation.

Table 58. Performance Level Distribution Summary, by Subgroup, Grade 7

Demographic Category (Subgroup)		N-count	Level I %	Level II %	Level III %	Level IV %	Levels III & IV %
State	All Students	202109	7.86	27.44	34.24	30.46	64.70
Gender	Female	98985	6.67	27.25	35.37	30.71	66.07
	Male	103124	9.00	27.61	33.15	30.23	63.38
Ethnicity	Asian	15874	3.36	12.92	26.36	57.36	83.72
	Black	38113	15.17	41.99	30.26	12.59	42.85
	Hispanic	44029	12.17	38.04	33.73	16.06	49.79
	American Indian	984	12.30	31.91	36.48	19.31	55.79
	Multi-Racial	1113	6.92	25.61	33.69	33.78	67.48
	Unknown	285	10.53	28.42	33.68	27.37	61.05
	White	101711	3.93	19.63	37.16	39.28	76.44
NRC	New York City	70022	10.68	33.61	30.47	25.24	55.71
	Big 4 Cites	7633	23.75	43.43	24.21	8.61	32.82
	High Needs Urban/Suburban	14675	11.57	39.39	34.09	14.95	49.04
	High Needs Rural	11778	6.89	34.11	39.12	19.88	59.00
	Average Needs	61250	4.25	21.97	39.13	34.65	73.78
	Low Needs	32151	2.33	12.14	33.97	51.55	85.52
	Charter	3734	5.65	31.41	39.18	23.75	62.94
SWD	All Codes	30697	28.84	45.61	20.54	5.01	25.55
SUA	All Codes	44768	24.50	43.82	23.66	8.02	31.67
SWD/SUA 3	SUA=504 Plan Codes	27601	29.25	46.14	20.03	4.57	24.6
ELL/SUA 2	SUA=ELL Codes	10788	27.48	45.08	19.91	7.53	27.44
ELL	ELL status = Y	12718	28.98	45.03	19.19	6.80	25.99
ELL Test Language	English	9102	28.96	46.89	18.85	5.30	24.15
	Chinese	852	4.23	20.54	32.75	42.49	75.23
	Haitian Creole	130	42.31	37.69	17.69	2.31	20.00
	Korean	76	2.63	14.47	38.16	44.74	82.89
	Russian	102	20.59	40.20	32.35	6.86	39.22
	Spanish	3290	30.88	45.81	19.97	3.34	23.31
	All Translations	4450	25.39	40.07	22.94	11.60	34.54

Grade 8

Performance level summaries and N-counts of demographic groups for Grade 8 are presented in Table 59. Statewide, 59.98% 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. Across ethnic and test translation subgroups, there were marked differences in performance. Around 12%–17% 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 38%–45% of students meeting standards for Levels III and IV, respectively, whereas about 83% of Asian students were meeting Levels III and IV standards. About 23% of Big 4 Cities students were in Levels III and IV, yet over 82% of students from Low Needs districts were classified in these proficiency levels. Approximately 25%–30% of SWD, SUA, and ELL students were

placed in Level I. Less than 11% 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 who had a very high percentage of students in Level IV (38.41% and 35.29%, respectively). Across all subgroups, the Spanish translation subgroup had the largest percentage of students placed in Level I (26.83%), and the Chinese translation subgroup had the largest percentage of students placed in Level IV (38.41%). 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, Charter, Chinese translation, and Korean translation.

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

Demographic Category (Subgroup)		N-count	Level I %	Level II %	Level III %	Level IV %	Levels III & IV %
State	All Students	203235	8.63	31.39	42.29	17.69	59.98
Gender	Female	99215	7.29	30.42	44.17	18.12	62.29
	Male	104020	9.92	32.31	40.50	17.28	57.77
Ethnicity	Asian	16436	3.18	13.33	37.89	45.60	83.49
	Black	38118	16.64	44.98	31.93	6.45	38.38
	Hispanic	43715	12.97	41.77	36.85	8.40	45.25
	American Indian	988	13.66	42.61	34.62	9.11	43.72
	Multi-Racial	980	7.65	26.63	47.55	18.16	65.71
	Unknown	265	7.17	32.08	35.09	25.66	60.75
	White	102733	4.65	24.76	49.19	21.40	70.59
NRC	New York City	71338	11.15	36.18	35.78	16.90	52.68
	Big 4 Cites	7651	29.26	47.47	20.44	2.82	23.26
	High Needs Urban/Suburban	14374	12.81	45.08	35.31	6.80	42.11
	High Needs Rural	11674	7.66	39.70	44.03	8.61	52.64
	Average Needs	61742	4.72	27.57	49.92	17.79	67.71
	Low Needs	32482	2.45	15.29	50.58	31.68	82.26
	Charter	2854	6.13	31.81	47.65	14.40	62.05
SWD	All Codes	29949	30.22	48.09	19.90	1.78	21.69
SUA	All Codes	43618	24.98	45.97	24.81	4.25	29.06
SWD/SUA 3	SUA=504 Plan Codes	26931	30.82	48.45	19.21	1.52	20.73
ELL/SUA 2	SUA=ELL Codes	11080	23.29	44.31	26.17	6.22	32.39
ELL	ELL status = Y	12932	25.34	44.18	24.86	5.61	30.47
ELL Test Language	English	8818	26.64	46.25	23.35	3.77	27.11
	Chinese	1070	2.90	14.39	44.30	38.41	82.71
	Haitian Creole	173	24.28	49.13	24.86	1.73	26.59
	Korean	51	7.84	1.96	54.90	35.29	90.20
	Russian	85	17.65	42.35	34.12	5.88	40.00
	Spanish	3351	26.83	47.09	24.47	1.61	26.08
	All Translations	4730	20.95	39.20	29.47	10.38	39.85

Section IX: Longitudinal Comparison of Results

This section provides a longitudinal comparison of OP scale score results on the NYSTP 2006–2011 Grades 3–8 Mathematics Tests. These include the scale score means, standard deviations, and performance level distributions for each grade's public and charter school population. The longitudinal results are presented in Table 60.

Table 60. Mathematics Grades 3–8 Tests 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	2011	198574	686.66	20.98	9.09	31.22	46.27	13.43	59.70
	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	2011	199134	687.96	32.21	5.54	27.73	39.99	26.74	66.73
	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	2011	202346	686.12	30.49	5.75	27.89	42.85	23.51	66.36
	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	2011	200076	682.16	31.60	7.91	28.99	36.72	26.38	63.10
	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	2011	202109	678.66	30.59	7.86	27.44	34.24	30.46	64.70
	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

(Continued on next page)

Table 60. Mathematics Grades 3–8 Tests Longitudinal Results (cont.)

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
8	2011	203235	677.25	33.66	8.63	31.39	42.29	17.69	59.98
	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 2006 and 2011 administrations and the scale scores from the 2006 and 2011 administrations can be directly compared, the performance level results between 2006–2009 and 2010–2011 OP tests are *not* directly comparable because of re-setting the proficiency level cut score values after the 2010 OP test administration.

As seen in Table 60, an increase in scale score means was observed for all mathematics grades between 2006 and 2011. The least gain was observed for Grades 3 and 4, for which the total gain was about 9 and 11 scale score points, respectively, between the 2006 and 2011 test administrations. The greatest gain in scale score points between the 2006 and 2011 test administrations was noted for Grades 6, 7, and 8 (26, 27, and 25 scale score points, respectively).

The variability of scale score distribution decreased steadily across years for mathematics Grades 5, 6, and 7. The scale score standard deviation was around 40 scale score points for those grades in the first test administration year and decreased to around 30–32 scale score points in 2011. For Grade 8, the variability of scale score distribution decreased steadily from 41 scale score points in 2006 to 32 scale score points in 2010 and then increased to 33 scale score points in 2011.

The scale score standard deviation for Grades 3 decreased slightly between 2006 and 2009 (less than 3 scale score points), then decreased about 4 scale score points between 2009 and 2010, and then dropped 12 scale score points between 2010 and 2011. The Grade 4 standard deviation had a similar trend as Grade 3 between 2006 and 2010, and only dropped 2 scale score points between 2010 and 2011.

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 2011 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 2011 item selection will include 2005, 2006, 2008, 2009, and 2010 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 2011 OP forms and the 2010 OP forms. Select the 2011 forms to be more difficult than the 2010 forms.
- Try to get the best scale coverage—make sure that items cover a wide range of the scale.
- Provide Research with the following item selection information:
    - Percentage of score points per learning standard (target, 2011 full selection, 2011 MC items only)
    - Item number in 2011 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 (ELL), students with disabilities (SWD), 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**	**9.12**	**19.83**	**19.83**
		2	1.90	4.13	23.95
		3	1.18	2.56	26.51
		4	1.07	2.34	28.85
		5	1.05	2.28	31.12
		6	1.03	2.23	33.35
	SWD	**1**	**9.81**	**21.32**	**21.32**
		2	1.86	4.05	25.37
		3	1.14	2.48	27.85
		4	1.12	2.42	30.27
		5	1.05	2.27	32.55
		6	1.02	2.21	34.76
	SUA	**1**	**9.90**	**21.51**	**21.51**
		2	1.90	4.13	25.64
		3	1.14	2.47	28.11
		4	1.09	2.38	30.49
		5	1.01	2.20	32.69
		6	1.01	2.19	34.88
	SWD/SUA	**1**	**9.48**	**20.61**	**20.61**
		2	1.83	3.97	24.58
		3	1.15	2.50	27.08
		4	1.13	2.45	29.52
		5	1.05	2.29	31.81
		6	1.03	2.23	34.05
	ELL/SUA	**1**	**9.12**	**19.83**	**19.83**
		2	1.92	4.16	23.99
		3	1.18	2.56	26.55
		4	1.10	2.38	28.93
		5	1.03	2.24	31.17
		6	1.03	2.23	33.40

(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	ELL	**1**	**12.10**	**21.23**	**21.23**
		2	1.99	3.49	24.72
		3	1.22	2.14	26.86
		4	1.18	2.06	28.92
		5	1.07	1.88	30.80
		6	1.02	1.78	32.58
		7	1.01	1.76	34.35
	SWD	**1**	**12.59**	**22.08**	**22.08**
		2	2.06	3.61	25.69
		3	1.25	2.20	27.89
		4	1.17	2.06	29.95
		5	1.08	1.90	31.84
		6	1.03	1.81	33.65
	SUA	**1**	**13.17**	**23.10**	**23.10**
		2	2.05	3.59	26.69
		3	1.22	2.14	28.83
		4	1.17	2.06	30.89
		5	1.06	1.85	32.74
		6	1.01	1.77	34.51
	SWD/SUA	**1**	**12.15**	**21.31**	**21.31**
		2	2.03	3.57	24.88
		3	1.27	2.22	27.10
		4	1.18	2.06	29.16
		5	1.09	1.91	31.07
		6	1.04	1.83	32.90
	ELL/SUA	**1**	**12.29**	**21.56**	**21.56**
		2	2.01	3.53	25.09
		3	1.21	2.13	27.21
		4	1.16	2.04	29.25
		5	1.07	1.88	31.13
		6	1.02	1.79	32.92
		7	1.02	1.78	34.71

(Continued on next page)

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

Grade	Subgroup	Initial Eigenvalues			
		Component	Total	% of Variance	Cumulative %
5	ELL	**1**	**9.00**	**18.37**	**18.37**
		2	1.56	3.19	21.56
		3	1.23	2.52	24.07
		4	1.11	2.27	26.34
		5	1.11	2.26	28.60
		6	1.04	2.11	30.71
		7	1.01	2.07	32.78
	SWD	**1**	**8.95**	**18.27**	**18.27**
		2	1.53	3.13	21.40
		3	1.21	2.48	23.88
		4	1.10	2.24	26.12
		5	1.07	2.19	28.30
		6	1.03	2.11	30.41
		7	1.01	2.06	32.47
	SUA	**1**	**9.64**	**19.67**	**19.67**
		2	1.55	3.15	22.82
		3	1.21	2.46	25.28
		4	1.08	2.21	27.49
		5	1.05	2.14	29.63
		6	1.02	2.09	31.72
	SWD/SUA	**1**	**8.66**	**17.67**	**17.67**
		2	1.51	3.09	20.76
		3	1.22	2.50	23.25
		4	1.10	2.25	25.51
		5	1.08	2.21	27.71
		6	1.04	2.12	29.83
		7	1.01	2.07	31.90
		8	1.00	2.04	33.94
	ELL/SUA	**1**	**9.34**	**19.06**	**19.06**
		2	1.58	3.22	22.29
		3	1.20	2.45	24.73
		4	1.11	2.26	27.00
		5	1.08	2.20	29.20
		6	1.03	2.11	31.31
		7	1.01	2.05	33.36
		8	1.00	2.04	35.40

(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	**1**	**9.91**	**19.82**	**19.82**
		2	1.86	3.73	23.55
		3	1.47	2.94	26.49
		4	1.15	2.30	28.79
		5	1.06	2.12	30.91
		6	1.03	2.07	32.98
		7	1.00	2.00	34.98
	SWD	**1**	**9.86**	**19.72**	**19.72**
		2	1.86	3.73	23.44
		3	1.58	3.15	26.59
		4	1.12	2.24	28.84
		5	1.05	2.10	30.94
	SUA	1	10.67	21.34	21.34
		2	1.91	3.81	25.15
		3	1.52	3.05	28.20
		4	1.12	2.25	30.45
		5	1.04	2.09	32.54
	SWD/SUA	1	9.61	19.22	19.22
		2	1.84	3.67	22.89
		3	1.58	3.16	26.05
		4	1.12	2.24	28.29
		5	1.06	2.12	30.42
		6	1.00	2.00	32.42
	ELL/SUA	1	10.46	20.91	20.91
		2	1.95	3.90	24.81
		3	1.46	2.91	27.72
		4	1.14	2.28	30.01
		5	1.06	2.12	32.13
		6	1.01	2.02	34.15

(Continued on next page)

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

Grade	Subgroup	Initial Eigenvalues			
		Component	Total	% of Variance	Cumulative %
7	ELL	**1**	**8.21**	**15.79**	**15.79**
		2	1.65	3.18	18.96
		3	1.51	2.90	21.86
		4	1.17	2.24	24.10
		5	1.07	2.07	26.17
		6	1.04	2.00	28.17
		7	1.04	1.99	30.16
		8	1.03	1.97	32.13
		9	1.01	1.95	34.08
	SWD	**1**	**8.30**	**15.97**	**15.97**
		2	1.65	3.17	19.14
		3	1.49	2.86	22.00
		4	1.14	2.19	24.18
		5	1.05	2.02	26.20
		6	1.02	1.96	28.16
		7	1.01	1.95	30.11
	SUA	**1**	**9.13**	**17.55**	**17.55**
		2	1.71	3.29	20.84
		3	1.49	2.86	23.70
		4	1.13	2.17	25.88
		5	1.04	2.00	27.88
		6	1.01	1.94	29.82
	SWD/SUA	**1**	**8.07**	**15.51**	**15.51**
		2	1.64	3.15	18.67
		3	1.49	2.86	21.52
		4	1.14	2.18	23.70
		5	1.06	2.03	25.73
		6	1.03	1.97	27.70
		7	1.02	1.96	29.66
		8	1.00	1.92	31.59
	ELL/SUA	**1**	**9.00**	**17.31**	**17.31**
		2	1.66	3.19	20.50
		3	1.50	2.88	23.38
		4	1.16	2.23	25.61
		5	1.07	2.06	27.67
		6	1.03	1.99	29.66
		7	1.02	1.96	31.62
		8	1.01	1.94	33.55

(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	ELL	**1**	**11.02**	**20.40**	**20.40**
		2	1.44	2.67	23.07
		3	1.33	2.47	25.54
		4	1.13	2.10	27.64
		5	1.04	1.93	29.57
		6	1.03	1.91	31.48
		7	1.01	1.87	33.35
	SWD	**1**	**9.63**	**17.83**	**17.83**
		2	1.47	2.73	20.56
		3	1.32	2.45	23.01
		4	1.08	2.00	25.01
		5	1.06	1.96	26.97
		6	1.04	1.92	28.89
		7	1.02	1.89	30.78
		8	1.00	1.85	32.64
	SUA	**1**	**11.16**	**20.66**	**20.66**
		2	1.45	2.68	23.35
		3	1.34	2.48	25.82
		4	1.05	1.94	27.77
		5	1.04	1.92	29.69
		6	1.01	1.87	31.56
	SWD/SUA	**1**	**9.37**	**17.35**	**17.35**
		2	1.48	2.73	20.08
		3	1.31	2.43	22.51
		4	1.09	2.03	24.54
		5	1.07	1.98	26.52
		6	1.04	1.92	28.44
		7	1.02	1.90	30.34
		8	1.01	1.86	32.20
	ELL/SUA	**1**	**12.12**	**22.45**	**22.45**
		2	1.49	2.77	25.21
		3	1.27	2.35	27.56
		4	1.11	2.06	29.62
		5	1.04	1.92	31.54
		6	1.00	1.86	33.39

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 2011 Classical DIF Item Flags

Grade	Item #	Subgroup	DIF	SMD	Mantel-Haenszel	Delta
3	37	Black	Against	-0.133	2231.680	-1.572
3	37	Hispanic	Against	-0.113	No Flag	No Flag
3	44	Black	In Favor	0.161	No Flag	No Flag
3	44	Hispanic	In Favor	0.108	No Flag	No Flag
3	44	High Need	In Favor	0.109	No Flag	No Flag
3	44	ELL	In Favor	0.118	No Flag	No Flag
3	44	Spanish	In Favor	0.136	No Flag	No Flag
4	1	Spanish	Against	No Flag	217.147	-1.677
4	11	Spanish	Against	-0.114	No Flag	No Flag
4	12	Spanish	Against	No Flag	160.651	-1.504
4	39	Female	Against	No Flag	3200.530	-1.642
4	50	Black	In Favor	0.122	No Flag	No Flag
4	50	Hispanic	In Favor	0.114	No Flag	No Flag
4	54	Female	In Favor	0.145	No Flag	No Flag
4	55	Female	In Favor	0.112	No Flag	No Flag
4	56	Female	In Favor	0.172	No Flag	No Flag
4	57	Spanish	Against	-0.128	No Flag	No Flag
5	35	Asian	Against	-0.108	No Flag	No Flag
5	46	Female	In Favor	0.149	No Flag	No Flag
5	46	Black	In Favor	0.170	No Flag	No Flag
5	46	Hispanic	In Favor	0.126	No Flag	No Flag
5	46	High Need	In Favor	0.103	No Flag	No Flag
5	49	Spanish	Against	-0.194	No Flag	No Flag
6	23	Female	Against	-0.139	4158.582	-1.511
6	28	Asian	Against	No Flag	238.089	-1.739
6	28	ELL	Against	No Flag	644.910	-1.512
6	42	ELL	Against	-0.167	No Flag	No Flag
6	42	Spanish	Against	-0.133	No Flag	No Flag
6	44	Female	In Favor	0.105	No Flag	No Flag
6	48	ELL	Against	-0.202	No Flag	No Flag
6	48	Spanish	Against	-0.213	No Flag	No Flag

(Continued on next page)

Table D1. NYSTP Mathematics 2011 Classical DIF Item Flags (cont.)

Grade	Item #	Subgroup	DIF	SMD	Mantel-Haenszel	Delta
7	2	ELL	In Favor	0.101	No Flag	No Flag
7	3	Asian	In Favor	0.112	987.698	1.555
7	4	Asian	Against	No Flag	517.526	-1.687
7	9	Asian	Against	No Flag	602.059	-1.789
7	9	ELL	Against	-0.105	No Flag	No Flag
7	32	ELL	Against	-0.116	No Flag	No Flag
7	32	Spanish	Against	-0.158	420.349	-1.891
7	45	Asian	In Favor	No Flag	860.917	1.556
7	46	Asian	In Favor	0.118	No Flag	No Flag
7	46	Black	In Favor	0.149	No Flag	No Flag
7	46	Hispanic	In Favor	0.120	No Flag	No Flag
7	46	High Need	In Favor	0.177	No Flag	No Flag
7	47	Female	In Favor	0.106	No Flag	No Flag
7	48	High Need	In Favor	-0.128	No Flag	No Flag
7	49	Female	In Favor	-0.105	No Flag	No Flag
7	50	Spanish	In Favor	-0.124	No Flag	No Flag
7	51	Female	In Favor	0.107	No Flag	No Flag
7	51	ELL	Against	-0.143	No Flag	No Flag
7	51	Spanish	In Favor	-0.103	No Flag	No Flag
7	52	ELL	In Favor	0.214	No Flag	No Flag
7	52	Spanish	In Favor	0.236	No Flag	No Flag
7	53	Female	In Favor	0.104	No Flag	No Flag
7	53	Asian	In Favor	-0.155	No Flag	No Flag
7	53	Black	In Favor	-0.142	No Flag	No Flag
7	53	Hispanic	In Favor	-0.139	No Flag	No Flag
7	53	High Need	In Favor	-0.153	No Flag	No Flag
8	9	Spanish	Against	-0.125	No Flag	No Flag
8	10	Spanish	Against	-0.120	No Flag	No Flag
8	43	Asian	In Favor	0.105	No Flag	No Flag
8	43	Black	In Favor	0.136	No Flag	No Flag
8	43	Hispanic	In Favor	0.147	No Flag	No Flag
8	43	High Need	In Favor	0.123	No Flag	No Flag
8	46	Black	Against	-0.125	No Flag	No Flag
8	46	Hispanic	Against	-0.109	No Flag	No Flag
8	46	High Need	Against	-0.111	No Flag	No Flag
8	46	ELL	Against	-0.120	No Flag	No Flag
8	46	Spanish	Against	-0.136	No Flag	No Flag
8	47	Female	In Favor	0.113	No Flag	No Flag
8	50	ELL	In Favor	0.116	No Flag	No Flag
8	50	Spanish	In Favor	0.103	No Flag	No Flag
8	51	ELL	Against	-0.103	No Flag	No Flag

(Continued on next page)

Table D1. NYSTP Mathematics 2011 Classical DIF Item Flags (cont.)

Grade	Item #	Subgroup	DIF	SMD	Mantel-Haenszel	Delta
8	51	Spanish	Against	-0.129	No Flag	No Flag
8	52	Female	In Favor	0.161	No Flag	No Flag
8	52	ELL	Against	-0.173	No Flag	No Flag
8	52	Spanish	Against	-0.209	No Flag	No Flag
8	53	Asian	Against	-0.115	No Flag	No Flag
8	53	Black	Against	-0.161	No Flag	No Flag
8	53	Hispanic	Against	-0.158	No Flag	No Flag
8	53	High Need	Against	-0.145	No Flag	No Flag
8	53	ELL	Against	-0.151	No Flag	No Flag
8	53	Spanish	Against	-0.196	No Flag	No Flag
8	54	Asian	Against	-0.111	No Flag	No Flag
8	54	Black	Against	-0.118	No Flag	No Flag
8	54	Hispanic	Against	-0.111	No Flag	No Flag
8	54	High Need	Against	-0.115	No Flag	No Flag
8	54	ELL	Against	-0.150	No Flag	No Flag
8	54	Spanish	Against	-0.107	No Flag	No Flag

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

Grade	Item	Focal Group	Direction	Magnitude
3	44	Spanish	In Favor	0.165
4	11	Spanish	Against	-0.101
4	46	Spanish	In Favor	0.119
4	49	Spanish	In Favor	0.108
4	57	Spanish	Against	-0.100
5	49	Spanish	Against	-0.126
6	42	Spanish	Against	-0.127
6	42	LEP	Against	-0.165
6	48	Spanish	Against	-0.191
6	48	LEP	Against	-0.175
7	32	Spanish	Against	-0.137
7	51	LEP	Against	-0.115
7	52	LEP	In Favor	0.194
7	52	Spanish	In Favor	0.229
8	9	Spanish	Against	-0.114
8	10	Spanish	Against	-0.110
8	46	LEP	Against	-0.109
8	46	Spanish	Against	-0.116
8	50	LEP	In Favor	0.114
8	50	Spanish	In Favor	0.122
8	51	Spanish	Against	-0.127

(Continued on next page)

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

Grade	Item	Focal Group	Direction	Magnitude
8	52	LEP	Against	-0.151
8	52	Spanish	Against	-0.180
8	53	LEP	Against	-0.159
8	53	Black	Against	-0.109
8	53	Spanish	Against	-0.185
8	54	LEP	Against	-0.123

Appendix E—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:

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

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

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

with

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

and

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

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

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

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

Estimation of the Prior Distribution of T_j

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

The 3PL model is

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

where

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

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

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

where

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

and

$$\gamma_{i0} = 0\,.$$

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

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

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

where

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

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

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

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

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

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

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

and

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

Solving these equations for r_j and s_j produces

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

and

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

where

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

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

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

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

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

where

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

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

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

and

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

Thus,

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

and the parameters of the prior beta distribution for T_j can be expressed in terms of the parameters of the 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:

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

and

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

The OPI is

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

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

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

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

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

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

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

Check on Consistency and Adjustment of Weight Given to Prior Estimate

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

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

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

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

Possible Violations of the Assumptions

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

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

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

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

Appendix F—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

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

and

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

Because obtaining test scores from two independent administrations of New York State tests was not feasible due to item release after each OP administration, a psychometric model was used to obtain the estimated classification consistency indices using test scores from a single administration. Based on the psychometric model, a symmetric H*H contingency table can be constructed. The elements of 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 G—Scale Score Frequency Distributions

Tables G1–G6 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 G1. Grade 3 Mathematics 2011 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
470	278	0.14	278	0.14
605	205	0.10	483	0.24
619	293	0.15	776	0.39
627	381	0.19	1157	0.58
632	525	0.26	1682	0.85
636	663	0.33	2345	1.18
640	753	0.38	3098	1.56
643	864	0.44	3962	2.00
645	999	0.50	4961	2.50
648	1123	0.57	6084	3.06
650	1224	0.62	7308	3.68
652	1301	0.66	8609	4.34
654	1513	0.76	10122	5.10
656	1773	0.89	11895	5.99
658	1888	0.95	13783	6.94
660	1996	1.01	15779	7.95
661	2263	1.14	18042	9.09
663	2467	1.24	20509	10.33
664	2698	1.36	23207	11.69
666	3040	1.53	26247	13.22
668	3256	1.64	29503	14.86
669	3419	1.72	32922	16.58
670	3801	1.91	36723	18.49
672	4056	2.04	40779	20.54
673	4433	2.23	45212	22.77
675	4778	2.41	49990	25.17
676	5122	2.58	55112	27.75

(Continued on next page)

Table G1. Grade 3 Mathematics 2011 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
678	5626	2.83	60738	30.59
679	6079	3.06	66817	33.65
681	6422	3.23	73239	36.88
682	6795	3.42	80034	40.30
684	7346	3.70	87380	44.00
686	7509	3.78	94889	47.79
687	8258	4.16	103147	51.94
689	8395	4.23	111542	56.17
691	8541	4.30	120083	60.47
693	8738	4.40	128821	64.87
695	8841	4.45	137662	69.33
697	8907	4.49	146569	73.81
699	8904	4.48	155473	78.29
702	8451	4.26	163924	82.55
705	7985	4.02	171909	86.57
708	7281	3.67	179190	90.24
711	6386	3.22	185576	93.45
716	5316	2.68	190892	96.13
722	4014	2.02	194906	98.15
732	2540	1.28	197446	99.43
770	1128	0.57	198574	100.00

Table G2. Grade 4 Mathematics 2011 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
485	77	0.04	77	0.04
542	77	0.04	154	0.08
564	120	0.06	274	0.14
576	195	0.10	469	0.24
585	278	0.14	747	0.38
593	376	0.19	1123	0.56
598	471	0.24	1594	0.80
604	568	0.29	2162	1.09
608	677	0.34	2839	1.43
612	730	0.37	3569	1.79
616	820	0.41	4389	2.20

(Continued on next page)

Table G2. Grade 4 Mathematics 2011 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
620	942	0.47	5331	2.68
623	972	0.49	6303	3.17
626	1052	0.53	7355	3.69
629	1168	0.59	8523	4.28
632	1233	0.62	9756	4.90
635	1280	0.64	11036	5.54
637	1420	0.71	12456	6.26
639	1469	0.74	13925	6.99
642	1691	0.85	15616	7.84
644	1621	0.81	17237	8.66
646	1808	0.91	19045	9.56
648	1851	0.93	20896	10.49
650	1936	0.97	22832	11.47
652	2132	1.07	24964	12.54
654	2209	1.11	27173	13.65
655	2386	1.20	29559	14.84
657	2438	1.22	31997	16.07
659	2609	1.31	34606	17.38
661	2603	1.31	37209	18.69
662	2828	1.42	40037	20.11
664	2867	1.44	42904	21.55
665	2940	1.48	45844	23.02
667	3079	1.55	48923	24.57
669	3187	1.60	52110	26.17
670	3350	1.68	55460	27.85
672	3528	1.77	58988	29.62
673	3610	1.81	62598	31.44
675	3649	1.83	66247	33.27
677	3776	1.90	70023	35.16
678	3979	2.00	74002	37.16
680	4013	2.02	78015	39.18
682	4279	2.15	82294	41.33
683	4458	2.24	86752	43.56
685	4542	2.28	91294	45.85
687	4834	2.43	96128	48.27
689	4835	2.43	100963	50.70
691	5050	2.54	106013	53.24
692	5226	2.62	111239	55.86

(Continued on next page)

Table G2. Grade 4 Mathematics 2011 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
694	5335	2.68	116574	58.54
697	5573	2.80	122147	61.34
699	5745	2.88	127892	64.22
701	5975	3.00	133867	67.22
703	5885	2.96	139752	70.18
706	6133	3.08	145885	73.26
709	6230	3.13	152115	76.39
712	6108	3.07	158223	79.46
715	6117	3.07	164340	82.53
718	5964	2.99	170304	85.52
722	5941	2.98	176245	88.51
726	5538	2.78	181783	91.29
731	4968	2.49	186751	93.78
737	4473	2.25	191224	96.03
745	3534	1.77	194758	97.80
756	2470	1.24	197228	99.04
776	1404	0.71	198632	99.75
800	502	0.25	199134	100.00

Table G3. Grade 5 Mathematics 2011 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
495	322	0.16	322	0.16
523	251	0.12	573	0.28
570	392	0.19	965	0.48
589	545	0.27	1510	0.75
600	680	0.34	2190	1.08
609	860	0.43	3050	1.51
616	992	0.49	4042	2.00
621	1198	0.59	5240	2.59
626	1280	0.63	6520	3.22
631	1547	0.76	8067	3.99
634	1703	0.84	9770	4.83
638	1857	0.92	11627	5.75
641	2063	1.02	13690	6.77
644	2293	1.13	15983	7.90
647	2411	1.19	18394	9.09

(Continued on next page)

Table G3. Grade 5 Mathematics 2011 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
650	2581	1.28	20975	10.37
652	2830	1.40	23805	11.76
655	3041	1.50	26846	13.27
657	3111	1.54	29957	14.80
659	3454	1.71	33411	16.51
661	3704	1.83	37115	18.34
663	3864	1.91	40979	20.25
665	3958	1.96	44937	22.21
667	4304	2.13	49241	24.34
669	4445	2.20	53686	26.53
671	4463	2.21	58149	28.74
673	4855	2.40	63004	31.14
675	5061	2.50	68065	33.64
677	5218	2.58	73283	36.22
679	5535	2.74	78818	38.95
681	5622	2.78	84440	41.73
683	5726	2.83	90166	44.56
685	5961	2.95	96127	47.51
687	6119	3.02	102246	50.53
689	6281	3.10	108527	53.63
691	6373	3.15	114900	56.78
693	6580	3.25	121480	60.04
696	6521	3.22	128001	63.26
698	6662	3.29	134663	66.55
700	6767	3.34	141430	69.90
703	6790	3.36	148220	73.25
705	6548	3.24	154768	76.49
708	6625	3.27	161393	79.76
711	6497	3.21	167890	82.97
714	6209	3.07	174099	86.04
718	6023	2.98	180122	89.02
722	5625	2.78	185747	91.80
727	4964	2.45	190711	94.25
732	4119	2.04	194830	96.29
739	3257	1.61	198087	97.90
748	2203	1.09	200290	98.98
762	1326	0.66	201616	99.64
780	730	0.36	202346	100.00

Table G4. Grade 6 Mathematics 2011 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
500	449	0.22	449	0.22
553	359	0.18	808	0.40
581	545	0.27	1353	0.68
595	734	0.37	2087	1.04
604	890	0.44	2977	1.49
611	1073	0.54	4050	2.02
617	1320	0.66	5370	2.68
622	1469	0.73	6839	3.42
626	1543	0.77	8382	4.19
630	1733	0.87	10115	5.06
633	1814	0.91	11929	5.96
636	1943	0.97	13872	6.93
639	1952	0.98	15824	7.91
642	2171	1.09	17995	8.99
644	2263	1.13	20258	10.13
646	2356	1.18	22614	11.30
649	2497	1.25	25111	12.55
651	2568	1.28	27679	13.83
653	2658	1.33	30337	15.16
655	2893	1.45	33230	16.61
657	2929	1.46	36159	18.07
658	3011	1.50	39170	19.58
660	3297	1.65	42467	21.23
662	3386	1.69	45853	22.92
663	3631	1.81	49484	24.73
665	3670	1.83	53154	26.57
667	3809	1.90	56963	28.47
668	4055	2.03	61018	30.50
670	4160	2.08	65178	32.58
672	4214	2.11	69392	34.68
673	4443	2.22	73835	36.90
675	4483	2.24	78318	39.14
676	4681	2.34	82999	41.48
678	4911	2.45	87910	43.94
680	5036	2.52	92946	46.46
681	5179	2.59	98125	49.04
683	5182	2.59	103307	51.63

(Continued on next page)

Table G4. Grade 6 Mathematics 2011 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
685	5306	2.65	108613	54.29
687	5449	2.72	114062	57.01
688	5444	2.72	119506	59.73
690	5623	2.81	125129	62.54
692	5555	2.78	130684	65.32
694	5653	2.83	136337	68.14
696	5565	2.78	141902	70.92
699	5392	2.69	147294	73.62
701	5380	2.69	152674	76.31
703	5297	2.65	157971	78.96
706	5019	2.51	162990	81.46
709	4907	2.45	167897	83.92
712	4809	2.40	172706	86.32
715	4691	2.34	177397	88.66
719	4406	2.20	181803	90.87
723	4163	2.08	185966	92.95
728	3931	1.96	189897	94.91
734	3531	1.76	193428	96.68
743	2984	1.49	196412	98.17
757	2346	1.17	198758	99.34
780	1318	0.66	200076	100.00

Table G5. Grade 7 Mathematics 2011 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
500	527	0.26	527	0.26
533	386	0.19	913	0.45
576	557	0.28	1470	0.73
593	762	0.38	2232	1.10
604	1029	0.51	3261	1.61
612	1214	0.60	4475	2.21
618	1470	0.73	5945	2.94
623	1683	0.83	7628	3.77
628	1749	0.87	9377	4.64
631	1960	0.97	11337	5.61
635	2197	1.09	13534	6.70

(Continued on next page)

Table G5. Grade 7 Mathematics 2011 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
638	2359	1.17	15893	7.86
641	2458	1.22	18351	9.08
643	2726	1.35	21077	10.43
646	2885	1.43	23962	11.86
648	3190	1.58	27152	13.43
650	3224	1.60	30376	15.03
652	3427	1.70	33803	16.73
654	3566	1.76	37369	18.49
656	3611	1.79	40980	20.28
658	4001	1.98	44981	22.26
660	4081	2.02	49062	24.28
662	4268	2.11	53330	26.39
664	4278	2.12	57608	28.50
666	4483	2.22	62091	30.72
667	4586	2.27	66677	32.99
669	4666	2.31	71343	35.30
671	4778	2.36	76121	37.66
672	4719	2.33	80840	40.00
674	4737	2.34	85577	42.34
676	4903	2.43	90480	44.77
677	4973	2.46	95453	47.23
679	5077	2.51	100530	49.74
680	5086	2.52	105616	52.26
682	5069	2.51	110685	54.77
684	5007	2.48	115692	57.24
685	4982	2.47	120674	59.71
687	4919	2.43	125593	62.14
689	4966	2.46	130559	64.60
690	5041	2.49	135600	67.09
692	4939	2.44	140539	69.54
694	4917	2.43	145456	71.97
696	4882	2.42	150338	74.38
698	4957	2.45	155295	76.84
700	4927	2.44	160222	79.28
702	4919	2.43	165141	81.71
704	4917	2.43	170058	84.14
707	4789	2.37	174847	86.51

(Continued on next page)

Table G5. Grade 7 Mathematics 2011 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
710	4635	2.29	179482	88.80
713	4302	2.13	183784	90.93
716	4176	2.07	187960	93.00
721	3850	1.90	191810	94.90
726	3301	1.63	195111	96.54
733	2906	1.44	198017	97.98
742	2099	1.04	200116	99.01
760	1386	0.69	201502	99.70
800	607	0.30	202109	100.00

Table G6. Grade 8 Mathematics 2011 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
480	1554	0.76	1554	0.76
545	960	0.47	2514	1.24
590	1316	0.65	3830	1.88
606	1706	0.84	5536	2.72
616	2041	1.00	7577	3.73
623	2222	1.09	9799	4.82
629	2425	1.19	12224	6.01
634	2564	1.26	14788	7.28
638	2757	1.36	17545	8.63
641	2950	1.45	20495	10.08
644	3029	1.49	23524	11.57
647	3188	1.57	26712	13.14
650	3281	1.61	29993	14.76
652	3320	1.63	33313	16.39
654	3274	1.61	36587	18.00
656	3422	1.68	40009	19.69
658	3497	1.72	43506	21.41
660	3728	1.83	47234	23.24
661	3588	1.77	50822	25.01
663	3612	1.78	54434	26.78
665	3857	1.90	58291	28.68
666	3805	1.87	62096	30.55
668	3789	1.86	65885	32.42
669	3915	1.93	69800	34.34

(Continued on next page)

Table G6. Grade 8 Mathematics 2011 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
670	3843	1.89	73643	36.24
672	3873	1.91	77516	38.14
673	3828	1.88	81344	40.02
674	3794	1.87	85138	41.89
676	3954	1.95	89092	43.84
677	3911	1.92	93003	45.76
678	3990	1.96	96993	47.72
679	3966	1.95	100959	49.68
680	3904	1.92	104863	51.60
682	3882	1.91	108745	53.51
683	4022	1.98	112767	55.49
684	3823	1.88	116590	57.37
685	3914	1.93	120504	59.29
687	3945	1.94	124449	61.23
688	3873	1.91	128322	63.14
689	3994	1.97	132316	65.10
690	3923	1.93	136239	67.04
692	3957	1.95	140196	68.98
693	3927	1.93	144123	70.91
694	3968	1.95	148091	72.87
696	3864	1.90	151955	74.77
697	3943	1.94	155898	76.71
699	3846	1.89	159744	78.60
701	3783	1.86	163527	80.46
702	3762	1.85	167289	82.31
704	3906	1.92	171195	84.23
706	3711	1.83	174906	86.06
708	3574	1.76	178480	87.82
710	3546	1.74	182026	89.56
713	3321	1.63	185347	91.20
715	3200	1.57	188547	92.77
718	3020	1.49	191567	94.26
721	2660	1.31	194227	95.57
725	2462	1.21	196689	96.78
730	2161	1.06	198850	97.84
737	1791	0.88	200641	98.72
746	1321	0.65	201962	99.37
763	878	0.43	202840	99.81
775	395	0.19	203235	100.00

References

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

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

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

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

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

Cattell, R.B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research* 1:245–276.

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

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

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

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

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

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

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

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

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

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

Johnson, N.L. and S. Kotz (1970). *Distributions in statistics: Continuous univariate distributions* (Vol. 2). New York: John Wiley.

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

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

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

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

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

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

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

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

Mehrens, W.A. and I.J. Lehmann (1991). *Measurement and evaluation in education and psychology* (3rd ed.). New York: Holt, Rinehart, and Winston.

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

Muraki, E. and R.D. Bock (1991). *PARSCALE: Parameter scaling of rating data* [Computer program]. Chicago: Scientific Software, Inc.

Novick, M.R. and P.H. Jackson (1974). *Statistical methods for educational and psychological research.* New York: McGraw-Hill.

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

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

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

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

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

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

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

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

Yen, W.M. (1997). The technical quality of performance assessments: Standard errors of percents of students reaching standards. *Educational Measurement: Issues and Practice* 16:5–15.

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

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

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