New York State Testing Program 2012: Mathematics, Grades 3–8

Technical Report

Pearson
2012

Table of Contents

List of Tables

List of Figures

Section I: Introduction and Overview

Introduction

This technical report provides an overview of the New York State Testing Program (NYSTP) Grades 3–8 Mathematics 2012 Operational (OP) Tests. 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 performance.

Target Population

Students in New York State public schools in Grades 3, 4, 5, 6, 7, and 8 (and ungraded students of equivalent ages) are the target population for the Grades 3–8 Mathematics Tests. Nonpublic schools may participate in the testing program, but participation is not mandatory for them. In 2012, some nonpublic schools participated in the testing program across all grade levels. However, the statewide nonpublic-school student population was not well represented. The New York State Education Department (NYSED) decided 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* (SAM), available online at http://www.p12.nysed.gov/apda/sam/ei/ei-sam-12w.pdf.

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 they 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, 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 New York State Regents Mathematics 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 *New York State 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 the 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 the section, 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 for multiple-choice (MC) questions onto scannable answer sheets; and the teachers transcribe the responses to the constructed-response (CR) questions onto the regular test books. The large type

editions are created by Pearson and printed by NYSED and the braille editions are produced by Braille Publishers, Inc. The lead transcribers are members of the National Braille Association, California Transcribers and Educators of the Visually Handicapped, and the Contra Costa Braille Transcribers, and they have Library of Congress and Nemeth Code [Braille] Certifications.

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

Test Translations

Since these are 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)

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 items have a maximum score of 2 (CR2), and extended-response items have a maximum score of 3 (CR3). The tests were administered in New York State classrooms during April 2012 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 Book Design and Testing Times

The OP test books were administered, in order, over the course of three consecutive days across all grades. The Grades 3–8 Mathematics Tests consist of three books. Book 1 and Book 2 contain multiple-choice questions. Books 3 contains short and extended constructed-response items.

To allow students sufficient time to demonstrate what they have learned, schools were instructed to schedule 90 minutes for each session, on each day, at each grade. This did not include approximately 10 minutes of prep time at the beginning of each session for handing out materials and reading directions.

Embedded Field Test Questions

In 2010, the Department announced its commitment to embed multiple-choice questions for field testing within the Spring 2012 Grades 3–8 Mathematics Test. Embedding field test questions allows for a better representation of the student population and more reliable field test data on which to build future operational tests.

It was not apparent to students whether a question is a field test question that did not count towards their score or an operational test question that did count towards their score. The specific locations of the embedded items on a test form are not disclosed. These data are free from the effects of differential student motivation that may characterize stand-alone field test designs because the items are answered by students taking actual tests under standard administration procedures. The embedded field test questions reduced the amount of stand-alone field testing during the spring of 2012 but did not eliminate the need for them.

Table 1 provides information on the number and type of items in each book, as well as testing times. The 2012 *Teacher's Directions* (http://www.p12.nysed.gov/apda/ei/directions/2012/math3-5-td-12w.pdf and http://www.p12.nysed.gov/apda/ei/directions/2012/math6-8-td-12w.pdf) as well as the 2012 *SAM* (http://www.p12.nysed.gov/apda/sam/ei/ei-sam-12w.pdf) provide details on security, scheduling, classroom organization and preparation, test materials, and administration.

Table 1. NYSTP Mathematics 2012 Test Configuration

Grade	Day	Book	Number of Items				
			Multiple-Choice		Constructed-Response		Total**
			Operational	Embedded	Operational	Embedded	
3	1	1	23	6	0	0	29
	2	2	22	7	0	0	29
	3	3	0	0	7	0	7
	Totals		45	13	7	0	65
4	1	1	24	7*	0	0	31
	2	2	24	7*	0	0	31
	3	3	0	0	9	0	9
	Totals		48	14	9	0	71
5	1	1	24	6	0	0	30
	2	2	24	6	0	0	30
	3	3	0	0	8	0	8
	Totals		48	12	8	0	68
6	1	1	24	6	0	0	30
	2	2	24	6	0	0	30
	3	3	0	0	9	0	9
	Totals		48	12	9	0	69
7	1	1	24	7*	0	0	31
	2	2	24	7*	0	0	31
	3	3	0	0	9	0	9
	Totals		48	14	9	0	71
8	1	1	24	7*	0	0	31
	2	2	24	7*	0	0	31
	3	3	0	0	9	0	9
	Totals		48	14	9	0	71

*There were two research questions included in the seven embedded field-test item positions.
**Reflects actual items in the test books.

Test Blueprint

The NYSTP Mathematics Tests assess students on the content and process strands of New York State Mathematics Learning Standards. 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 2012 OP tests.

Table 2. NYSTP Mathematics 2012 Test Blueprint[1]

Grade	Total Points on OP Test	Content Strand	Target Points	Selected Points	Target % of Test	Selected % of Test
3	60	Number Sense and Operations	29	29	48.4	48.3
		Algebra	8	9	12.9	15.0
		Geometry	8	7	12.9	11.7
		Measurement	8	7	12.9	11.7
		Statistics and Probability	8	8	12.9	13.3
4	69	Number Sense and Operations	32	31	45.7	44.9
		Algebra	10	9	14.3	13.0
		Geometry	8	7	11.4	10.1
		Measurement	12	12	17.1	17.4
		Statistics and Probability	8	10	11.4	14.5
5*	61	Number Sense and Operations	24	25	38.0	41.0
		Algebra	7	8	11.3	13.1
		Geometry	16	13	25.4	21.3
		Measurement	9	9	14.1	14.8
		Statistics and Probability	7	6	11.3	9.8
6	70	Number Sense and Operations	26	25	37	35.7
		Algebra	13	14	19	20.0
		Geometry	12	13	17	18.6
		Measurement	8	8	11	11.4
		Statistics and Probability	11	10	16	14.3

[1] There were 0–4 multiple choice items excluded from the operational test at each grade level to take into account the learning standards typically taught in May and June of each year.

Grade	Total Points on OP Test	Content Strand	Target Points	Selected Points	Target % of Test	Selected % of Test
7	68	Number Sense and Operations	21	21	30.7	30.9
		Algebra	8	7	12.0	10.3
		Geometry	9	11	13.3	16.2
		Measurement	9	8	13.3	11.8
		Statistics and Probability	21	21	30.7	30.9
8	68	Number Sense and Operations	8	7	11.4	10.3
		Algebra	30	29	44.3	42.6
		Geometry	23	24	34.2	35.3
		Measurement	7	8	10.1	11.8
		Statistics and Probability	0	0	0.0	0.0

*For grade 5, one item was exposed to the public during test administration. Therefore, this item was removed from the analyses, scoring, and reporting.

New York State Educators' Involvement in Test Development

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

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

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

Content Rationale

In June 2004, test specifications meetings were held in Albany, New York, during which committees of state educators, along with NYSED staff, reviewed the standards and performance indicators and made 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 who were selected for their grade-level expertise, were grouped by grade band (i.e., Grades 3/4, 5/6, 7/8), and met for four days. The committees were composed of approximately ten participants per grade band. Upon completion of the committee meetings, NYSED reviewed the committees' determinations and approved them, with minor adjustments when necessary to maintain consistency across the grades. In January 2005, a second specifications meeting was held 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 and Review

The first step in the process of item development for the NYSED-owned items appearing in the 2012 Grades 3–8 Mathematics Tests was selection of items to be used. The Pearson content specialists were provided with specifications based on the test design (see Appendix A).

The content specialists at Pearson then selected items that would best elicit the types of items outlined during the test specifications meetings and distributed writing assignments to experienced item writers. The writers' assignments outlined the number and type of items (including depth-of-knowledge or thinking skill level) to write for each 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 to help them in their writing process.

Pearson content specialists 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 Pearson staff had been incorporated, the items were prepared for field testing.

Materials Development

Pearson staff assembled the items into field test (FT) forms and submitted the FT forms to NYSED for their review and approval. NYSED verified that the items met the specifications. Pearson staff incorporated the SED revisions and the forms were finalized for field testing. The FT forms were administered to students across New York State, using a census sample (in 2011) to ensure appropriate sampling of students. In addition, Pearson, in conjunction with NYSED test specialists, developed a combined *Teacher's Directions and School Administrator's Manual* to help ensure 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.

Item Selection and Test Creation (Criteria and Process)

The NYSTP Grades 3–8 Mathematics OP Tests were administered in April 2012. The test items were selected from the pool of items primarily field tested in 2011, using the data from those FT forms.

The OP test constructions were iterative processes at fine-tuning the item selection. Using the item pool, Pearson made preliminary selections for each grade. The selections were reviewed for alignment with the test design, blueprint, and research guidelines for item selection (see Appendix C). Item selection for the Grades 3–8 ELA Tests was based on the classical and item response theory (IRT) statistics of the test items. Selection was conducted by content experts from Pearson and NYSED and reviewed by psychometricians at Pearson and at NYSED. Final approval of the selected items was given by NYSED. Two criteria governed the item selection process. The first of these was to meet the content specifications provided by NYSED. Second, within the limits set by these requirements, developers selected items with the best psychometric characteristics from the FT item pool.

Pearson content specialists traveled to Albany, New York, in September 2011 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 April 2012.

In addition to the test books, Pearson 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

The original proficiency cut score recommendations and the drafting of performance standards occurred at the NYSTP ELA standard-setting review held in Albany in June 2006. In 2010, change in the test administration window between the 2008–2009 and 2009–2010 school years, and a decision to align the proficiency standards with Grade 8 student performance on the New

York State Regents Mathematics examinations, led to changes in the proficiency cut scores. The results were reviewed by the New York State 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) addressed the concept of validity in testing. Validity is the most important consideration in test evaluation. The concept refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores. Test validation is the process for accumulating evidence to support any particular inference. Validity, however, is a unitary concept. Although evidence may be accumulated in many ways, validity refers to the degree to which evidence supports the inferences made from test scores.

Content Validity

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

Logical analysis of test content indicates the degree to which the content of a test covers the domain of content the test is intended to measure. In the case of 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 NYSTP 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 0.92–0.94, 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 3 and 8 displayed a good item-model fit. One item in Grade 4, one item in Grade 5, two items in Grade 6, and one item in Grade 7 were flagged for poor fit. The fact that only a few items were deemed to have an 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)

	Initial Eigenvalues			
Grade	Component	Total	% of Variance	Cumulative %
3	**1**	**10.85**	**21.70**	**21.70**
	2	1.49	2.98	24.68
	3	1.14	2.28	26.96
	4	1.10	2.21	29.17
	5	1.01	2.02	31.19
	6	1.00	2.01	33.20
4	**1**	**12.49**	**22.30**	**22.30**
	2	1.70	3.03	25.33
	3	1.11	1.99	27.32
	4	1.07	1.91	29.23
	5	1.01	1.80	31.03
5	**1**	**12.71**	**24.93**	**24.93**
	2	1.40	2.75	27.68

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

		Initial Eigenvalues		
Grade	Component	Total	% of Variance	Cumulative %
5	3	1.11	2.17	29.85
	4	1.04	2.03	31.88
6	**1**	**14.54**	**25.51**	**25.51**
	2	1.89	3.32	28.83
	3	1.20	2.11	30.94
	4	1.09	1.91	32.85
	5	1.03	1.80	34.65
7	**1**	**12.57**	**22.86**	**22.86**
	2	1.85	3.36	26.22
	3	1.31	2.38	28.60
	4	1.06	1.92	30.52
	5	1.02	1.86	32.38
8	**1**	**13.89**	**25.25**	**25.25**
	2	2.07	3.76	29.01
	3	1.49	2.72	31.73
	4	1.23	2.24	33.97
	5	1.02	1.86	35.83

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 that are classified with disabilities and use one or more disability-related accommodation.

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 Pearson's editorial policies and guidelines for an equitable assessment, as well as NYSED's guidelines for item development. At the same time, all materials were written to NYSED's specifications and carefully checked by groups of trained New York State educators during the item review process.

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

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

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

In the third procedure, New York State educators reviewed all FT materials. These professionals were asked to consider and comment on the appropriateness of language, content, 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 and Mantel-Haenszel (see Section V, "Operational Test Data Collection and Classical Analysis"). 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 the *Scoring Site Operations Manual* (2012) located at http://www.p12.nysed.gov/apda/ei/ssom/ssom-12w.pdf.

Test Administration

NYSTP Grades 3–8 Mathematics Tests were administered at the classroom level during April 2012. The testing window for Grades 3–8 was April 17–19. The makeup test administration window for Grades 3–8 was April 20–24. The makeup test administration windows allowed students who were ill or otherwise unable to test during the assigned window to take the test.

Scoring Procedures of Operational Tests

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

Scoring Models

For the 2011–2012 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, district-wide, or individually. Schools were required to enter one of the following scoring model codes on student answer sheets:

1. Regional scoring—The scorers for the school's test papers included either staff from three or more school districts or staff from all nonpublic schools in an affiliation group

(nonpublic or charter schools may participate in regional scoring with public school districts and may be counted as one district);

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

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

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

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

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

Scoring of Constructed-Response Items

The scoring of CR items was based primarily on the scoring guides, which were created by Pearson from responses consensus scored by NYSED and New York State teachers during Rangefinding sessions. In 2012, the Pearson Mathematics hand-scoring team was composed of six team leaders, each representing one grade. Team leaders were selected on the basis of their hand-scoring 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. Trainers used these materials to train scoring committee members on the criteria for scoring CR items. Scoring Leader Handbooks were distributed to outline the responsibilities of the scoring roles. Pearson staff also conducted training sessions 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. 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 Mathematics Helpline (see the subsection "Quality Control Process").

Scorer Qualifications and Training

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

Quality Control Process

Test books were randomly distributed throughout each scoring room so that books from each region, district, school, or class were evenly dispersed. Teams were divided into groups of three to ensure that a variety of scorers graded each book. If a scorer and facilitator could not reach a decision on a paper after reviewing the scoring guides and audio files, they called the New York State Mathematics Helpline. This call center was established to help teachers and administrators during OP scoring. The helpline staff consisted of trained Pearson 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 twice daily during the scoring window. To affirm that all schools across the state adhered to scoring guidelines and policies, approximately 5% of the schools' results are audited each year by an outside vendor.

Section V: Operational Test Data Collection and Classical Analysis

Data Collection

OP test data were collected in two phases. During phase 1, a sample of approximately 98% of the student test records were received from the data warehouse and delivered to Pearson at the beginning of June 2012. These data were used for all data analyses. Phase 2 involved submitting "straggler files" to Pearson in late June 2012. 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. NYSED and the data repository were provided the results of the checking. Pearson 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 students from nonpublic schools and students with incorrect or incomplete grade information. Other deleted cases included cases with mismatched form language indicators for translated versions, 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 Number of Cases	0	203,999
Wrong Subject	0	203,999
No Grade	44	203,955
Wrong Grade	134	203,821
Nonpublic School	6,397	197,424
No Response	2	197,422
Invalid Score	0	197,422
Out of Range CR Scores	0	197,422
Duplicated Record	12	197,410
Language Mismatched Form	66	197,344

Table 4B. NYSTP Mathematics Data Cleaning, Grade 4

Exclusion Rule	# Deleted	# Cases Remain
Initial Number of Cases	0	210,087
Wrong Subject	0	210,087
No Grade	45	210,042
Wrong Grade	134	209,908
Nonpublic School	16,700	193,208
No Response	3	193,205
Invalid Score	0	193,205
Out of Range CR Scores	0	193,205
Duplicated Record	6	193,199
Language Mismatched Form	76	193,123

Table 4C. NYSTP Mathematics Data Cleaning, Grade 5

Exclusion Rule	# Deleted	# Cases Remain
Initial Number of Cases	0	202,243
Wrong Subject	0	202,243
No Grade	279	201,964
Wrong Grade	49	201,915
Nonpublic School	6,372	195,543
No Response	1	195,542
Invalid Score	0	195,542
Out of Range CR Scores	0	195,542
Duplicated Record	6	195,536
Language Mismatched Form	115	195,421

Table 4D. NYSTP Mathematics Data Cleaning, Grade 6

Exclusion Rule	# Deleted	# Cases Remain
Initial Number of Cases	0	213,249
Wrong Subject	0	213,249
No Grade	77	213,172
Wrong Grade	158	213,014
Nonpublic School	14,536	198,478
No Response	3	198,475
Invalid Score	0	198,475
Out of Range CR Scores	0	198,475
Duplicated Record	0	198,475
Language Mismatched Form	133	198,342

Table 4E. NYSTP Mathematics Data Cleaning, Grade 7

Exclusion Rule	# Deleted	# Cases Remain
Initial Number of Cases	0	202,460
Wrong Subject	0	202,460
No Grade	53	202,407
Wrong Grade	206	202,201
Nonpublic School	5,872	196,329
No Response	0	196,329
Invalid Score	0	196,329
Out of Range CR Scores	0	196,329
Duplicated Record	2	196,327
Language Mismatched Form	99	196,228

Table 4F. NYSTP Mathematics Data Cleaning, Grade 8

Exclusion Rule	# Deleted	# Cases Remain
Initial Number of Cases	0	212,278
Wrong Subject	0	212,278
No Grade	52	212,226
Wrong Grade	139	212,087
Nonpublic School	15,535	196,552
No Response	0	196,552
Invalid Score	0	196,552
Out of Range CR Scores	0	196,552
Duplicated Record	6	196,546
Language Mismatched Form	111	196,435

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 Capacity Category (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 = 197,344)

Demographic Category		N-count	% of Total N-count
NRC	NYC	72,821	36.90
	Big 4 Cities	8,108	4.11
	Urban/Suburban	15,193	7.70
	Rural	10,364	5.25
	Average Needs	57,032	28.90
	Low Needs	28,025	14.20
	Charter	5,801	2.94
Ethnicity	Asian	16,863	8.54
	Black	35,603	18.04
	Hispanic	47,869	24.26
	American Indian	1,082	0.55
	Multiracial	1,964	1.00
	Other	413	0.21
	White	93,550	47.40
ELL	No	178,709	90.56
	Yes	18,635	9.44
SWD	No	169,537	85.91
	Yes	27,807	14.09
SUA	No	155,509	78.80
	Yes	41,835	21.20

Table 5B. Grade 4 Sample Characteristics (N = 193,123)

Demographic Category		N-count	% of Total N-count
NRC	NYC	70,751	36.64
	Big 4 Cities	8,097	4.19
	Urban/Suburban	14,153	7.33
	Rural	10,409	5.39
	Average Needs	56,836	29.43
	Low Needs	28,287	14.65
	Charter	4,590	2.38
Ethnicity	Asian	16,548	8.57
	Black	34,902	18.07
	Hispanic	46,083	23.86
	American Indian	1,039	0.54
	Multiracial	1,654	0.86
	Other	325	0.17
	White	92,572	47.93
ELL	No	175,715	90.99
	Yes	17,408	9.01
SWD	No	164,021	84.93
	Yes	29,102	15.07
SUA	No	151,704	78.55
	Yes	41,419	21.45

Table 5C. Grade 5 Sample Characteristics (N = 195,421)

Demographic Category		N-count	% of Total N-count
NRC	NYC	69,656	35.64
	Big 4 Cities	8,067	4.13
	Urban/Suburban	14,587	7.46
	Rural	10,766	5.51
	Average Needs	58,132	29.75
	Low Needs	28,325	14.49
	Charter	5,888	3.01
Ethnicity	Asian	16,164	8.27
	Black	36,518	18.69
	Hispanic	45,875	23.47
	American Indian	983	0.50
	Multiracial	1,494	0.76
	Other	322	0.16
	White	94,065	48.13
ELL	No	180,386	92.31
	Yes	15,035	7.69
SWD	No	165,164	84.52
	Yes	30,257	15.48
SUA	No	153,565	78.58
	Yes	41,856	21.42

Table 5D. Grade 6 Sample Characteristics (N = 198,342)

Demographic Category		N-count	% of Total N-count
NRC	NYC	70,027	35.31
	Big 4 Cities	7,859	3.96
	Urban/Suburban	14,355	7.24
	Rural	10,738	5.41
	Average Needs	59,745	30.12
	Low Needs	30,104	15.18
	Charter	5,514	2.78
Ethnicity	Asian	16,988	8.57
	Black	37,106	18.71
	Hispanic	45,310	22.84
	American Indian	1,014	0.51
	Multiracial	1,399	0.71
	Other	96,162	48.48
	White	363	0.18
ELL	No	185,598	93.57
	Yes	12,744	6.43
SWD	No	168,349	84.88
	Yes	29,993	15.12
SUA	No	159,647	80.49
	Yes	38,695	19.51

Table 5E. Grade 7 Sample Characteristics (N = 196,228)

Demographic Category		N-count	% of Total N-count
NRC	NYC	68,506	34.91
	Big 4 Cities	7,695	3.92
	Urban/Suburban	14,194	7.23
	Rural	11,046	5.63
	Average Needs	58,731	29.93
	Low Needs	31,387	16.00
	Charter	4,669	2.38
Ethnicity	Asian	15,883	8.09
	Black	37,058	18.89
	Hispanic	44,229	22.54
	American Indian	1,010	0.51
	Multiracial	1,346	0.69
	Other	328	0.17
	White	96,374	49.11
ELL	No	183,928	93.73
	Yes	12,300	6.27
SWD	No	166,880	85.04
	Yes	29,348	14.96
SUA	No	159,635	81.35
	Yes	36,593	18.65

Table 5F. Grade 8 Sample Characteristics (N = 196,435)

Demographic Category		N-count	% of Total N-count
NRC	NYC	69,870	35.57
	Big 4 Cities	7,395	3.76
	Urban/Suburban	13,183	6.71
	Rural	10,950	5.57
	Average Needs	59,454	30.27
	Low Needs	32,180	16.38
	Charter	3,403	1.73
Ethnicity	Asian	16,082	8.19
	Black	36,621	18.64
	Hispanic	43,676	22.23
	American Indian	1,038	0.53
	Multiracial	1,123	0.57
	Other	351	0.18
	White	97,544	49.66
ELL	No	184,246	93.79
	Yes	12,189	6.21
SWD	No	167,392	85.21
	Yes	29,043	14.79
SUA	No	184,246	93.79
	Yes	12,189	6.21

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 a correct distractor), item analysis is the stage in which errors should be flagged and evaluated for rectification (suppression, credit, or another 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 (see Section III, "Validity," and Section VII, "Reliability and Standard Error of Measurement").

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 0.21 to 0.96. For Grade 3, the item p-values were between 0.27 and 0.94 with a mean of 0.76. For Grade 4, the item p-values were between 0.38 and 0.96 with a mean of 0.73. For Grade 5, the item p-values were between 0.37 and 0.93 with a mean of 0.68. For Grade 6, the item p-values were between 0.32 and 0.94 with a mean of 0.66. For Grade 7, the item p-values were between 0.21 and 0.91 with a mean of 0.63. For Grade 8, the item p-values were between 0.25 and 0.90 with a mean of 0.65. 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	PbisKey
01	MC	197,283	0.91	0.03	0.42
02	MC	197,273	0.94	0.04	0.34
03	MC	197,229	0.91	0.06	0.41

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

Item	Item Type	N-count	P-value	% Omit	PbisKey
04	MC	197,138	0.68	0.10	0.47
05	MC	197,196	0.59	0.07	0.45
06	MC	197,211	0.81	0.07	0.33
07	MC	197,008	0.63	0.17	0.53
08	MC	197,194	0.83	0.08	0.35
09	MC	197,258	0.85	0.04	0.35
10	MC	197,233	0.84	0.06	0.45
11	MC	197,118	0.66	0.11	0.57
12	MC	197,070	0.82	0.14	0.47
13	MC	197,144	0.89	0.10	0.36
14	MC	196,987	0.68	0.18	0.46
15	MC	197,091	0.80	0.13	0.58
16	MC	197,067	0.78	0.14	0.31
17	MC	197,085	0.78	0.13	0.47
18	MC	197,058	0.90	0.14	0.43
19	MC	197,104	0.87	0.12	0.48
20	MC	197,069	0.67	0.14	0.40
21	MC	197,058	0.87	0.14	0.55
22	MC	196,174	0.74	0.59	0.43
23	MC	197,323	0.93	0.01	0.39
24	MC	197,254	0.87	0.05	0.30
25	MC	197,226	0.83	0.06	0.45
26	MC	197,239	0.81	0.05	0.37
27	MC	197,232	0.62	0.06	0.20
28	MC	197,167	0.57	0.09	0.48
29	MC	196,481	0.82	0.06	0.36
30	MC	197,253	0.27	0.05	0.33
31	MC	197,203	0.60	0.07	0.49
32	MC	197,182	0.79	0.08	0.48

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

Item	Item Type	N-count	P-value	% Omit	PbisKey
33	MC	197,109	0.74	0.12	0.52
34	MC	197,065	0.46	0.14	0.36
35	MC	197,154	0.88	0.09	0.40
36	MC	197,140	0.74	0.10	0.58
37	MC	197,162	0.81	0.09	0.56
38	MC	197,153	0.88	0.10	0.29
39	MC	197,231	0.88	0.06	0.50
40	MC	197,174	0.81	0.09	0.48
41	MC	197,032	0.75	0.16	0.59
42	MC	196,878	0.67	0.23	0.46
43	MC	196,347	0.69	0.51	0.43
44	CR	197,126	0.78	0.11	
45	CR	197,152	0.75	0.10	
46	CR	197,092	0.54	0.13	
47	CR	197,039	0.84	0.15	
48	CR	197,189	0.69	0.08	
49	CR	197,174	0.76	0.09	
50	CR	196,991	0.64	0.18	

Table 6B. Item Analysis, Grade 4

Item	Item Type	N-count	P-value	% Omit	PbisKey
01	MC	193,089	0.96	0.02	0.28
02	MC	193,074	0.74	0.03	0.47
03	MC	193,063	0.93	0.03	0.36
04	MC	193,040	0.86	0.04	0.44
05	MC	193,008	0.60	0.06	0.45
06	MC	192,956	0.45	0.09	0.52
07	MC	193,069	0.94	0.03	0.34
08	MC	193,004	0.91	0.06	0.47

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

Item	Item Type	N-count	P-value	% Omit	PbisKey
09	MC	193,044	0.91	0.04	0.34
10	MC	193,028	0.91	0.05	0.34
11	MC	192,875	0.77	0.13	0.49
12	MC	193,022	0.70	0.05	0.43
13	MC	192,964	0.66	0.08	0.49
14	MC	192,902	0.70	0.11	0.55
15	MC	189,697	0.46	0.18	0.40
16	MC	192,956	0.59	0.09	0.42
17	MC	193,005	0.90	0.06	0.43
18	MC	192,875	0.55	0.13	0.43
19	MC	192,951	0.83	0.09	0.50
20	MC	192,884	0.76	0.12	0.42
21	MC	192,900	0.81	0.12	0.46
22	MC	192,791	0.47	0.17	0.22
23	MC	192,459	0.73	0.34	0.49
24	MC	193,092	0.92	0.02	0.32
25	MC	192,973	0.73	0.08	0.55
26	MC	193,010	0.86	0.05	0.48
27	MC	193,033	0.75	0.05	0.44
28	MC	193,037	0.81	0.04	0.39
29	MC	193,020	0.67	0.05	0.46
30	MC	192,951	0.51	0.09	0.51
31	MC	193,028	0.88	0.05	0.50
32	MC	192,517	0.70	0.04	0.33
33	MC	193,011	0.77	0.06	0.51
34	MC	193,028	0.85	0.05	0.36
35	MC	193,000	0.89	0.06	0.39
36	MC	192,940	0.83	0.09	0.51

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

Item	Item Type	N-count	P-value	% Omit	PbisKey
37	MC	192,949	0.86	0.09	0.44
38	MC	192,993	0.55	0.07	0.32
39	MC	192,966	0.89	0.08	0.35
40	MC	192,960	0.75	0.08	0.49
41	MC	192,996	0.77	0.07	0.58
42	MC	192,958	0.78	0.09	0.36
43	MC	192,812	0.84	0.16	0.40
44	MC	192,903	0.79	0.11	0.55
45	MC	192,935	0.87	0.10	0.50
46	MC	192,881	0.76	0.13	0.48
47	MC	192,631	0.53	0.25	0.60
48	CR	193,010	0.66	0.06	
49	CR	192,922	0.77	0.10	
50	CR	192,646	0.51	0.25	
51	CR	192,969	0.65	0.08	
52	CR	192,759	0.44	0.19	
53	CR	192,792	0.40	0.17	
54	CR	192,851	0.59	0.14	
55	CR	192,736	0.67	0.20	
56	CR	192,823	0.38	0.16	

Table 6C. Item Analysis, Grade 5

Item	Item Type	N-count	P-value	% Omit	PbisKey
01	MC	195,394	0.93	0.01	0.30
02	MC	195,388	0.87	0.02	0.38
03	MC	195,347	0.63	0.04	0.51
04	MC	195,289	0.74	0.07	0.49
05	MC	195,280	0.83	0.07	0.54
06	MC	195,219	0.53	0.10	0.47
07	MC	195,333	0.81	0.05	0.47

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

Item	Item Type	N-count	P-value	% Omit	PbisKey
08	MC	195,287	0.68	0.07	0.57
09	MC	195,309	0.85	0.06	0.31
10	MC	195,261	0.63	0.08	0.57
11	MC	195,296	0.60	0.06	0.39
12	MC	195,259	0.65	0.08	0.64
13	MC	195,287	0.64	0.07	0.40
14	MC	195,308	0.68	0.06	0.47
15	MC	195,191	0.73	0.12	0.51
16	MC	195,239	0.50	0.09	0.40
17	MC	195,161	0.50	0.13	0.50
18	MC	195,095	0.64	0.12	0.51
19	MC	195,187	0.37	0.12	0.34
20	MC	195,081	0.68	0.17	0.59
21	MC	194,864	0.46	0.29	0.52
22	MC	195,365	0.78	0.03	0.60
23	MC	195,385	0.90	0.02	0.44
24	MC	195,352	0.75	0.04	0.34
25	MC	195,319	0.68	0.05	0.46
26	MC	195,317	0.65	0.05	0.43
27	MC	195,225	0.71	0.10	0.55
28	MC	195,306	0.59	0.05	0.48
29	MC	195,309	0.70	0.06	0.56
30	MC	195,350	0.71	0.04	0.39
31	MC	195,330	0.70	0.05	0.62
32	MC	195,294	0.68	0.06	0.59
33	MC	195,306	0.87	0.06	0.44
34	MC	195,286	0.92	0.07	0.36
35	MC	195,264	0.57	0.08	0.47
36	MC	195,251	0.54	0.09	0.49
37	MC	195,318	0.75	0.05	0.60

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

Item	Item Type	N-count	P-value	% Omit	PbisKey
38	MC	195,186	0.44	0.12	0.33
39	MC	195,336	0.78	0.04	0.50
40	MC	195,292	0.64	0.07	0.42
41	MC	195,206	0.54	0.11	0.40
42	MC	195,235	0.72	0.10	0.59
43	MC	195,227	0.74	0.10	0.50
44	MC	194,870	0.86	0.28	0.39
45	CR	195,281	0.70	0.07	
46	CR	195,320	0.79	0.05	
47	CR	194,797	0.63	0.32	
48	CR	195,087	0.44	0.17	
49	CR	195,233	0.63	0.10	
50	CR	195,271	0.56	0.08	
51	CR	195,131	0.65	0.15	

Table 6D. Item Analysis, Grade 6

Item	Item Type	N-count	P-value	% Omit	PbisKey
01	MC	198,262	0.87	0.04	0.52
02	MC	198,106	0.75	0.12	0.48
03	MC	198,232	0.74	0.06	0.55
04	MC	198,198	0.72	0.07	0.57
05	MC	198,207	0.37	0.07	0.23
06	MC	198,184	0.73	0.08	0.39
07	MC	198,217	0.63	0.06	0.58
08	MC	198,089	0.57	0.13	0.40
09	MC	198,189	0.47	0.08	0.31
10	MC	198,026	0.32	0.16	0.39
11	MC	198,178	0.81	0.08	0.46
12	MC	197,777	0.55	0.28	0.44

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

Item	Item Type	N-count	P-value	% Omit	PbisKey
13	MC	198,060	0.61	0.14	0.54
14	MC	197,674	0.37	0.34	0.26
15	MC	198,208	0.72	0.07	0.60
16	MC	198,202	0.60	0.07	0.35
17	MC	197,894	0.42	0.23	0.42
18	MC	198,096	0.72	0.12	0.53
19	MC	198,151	0.57	0.10	0.51
20	MC	198,121	0.72	0.11	0.40
21	MC	197,932	0.67	0.21	0.48
22	MC	197,903	0.63	0.22	0.65
23	MC	198,058	0.81	0.14	0.39
24	MC	197,953	0.80	0.20	0.44
25	MC	198,312	0.94	0.02	0.25
26	MC	198,260	0.70	0.04	0.61
27	MC	198,084	0.65	0.13	0.49
28	MC	198,202	0.78	0.07	0.36
29	MC	198,256	0.84	0.04	0.57
30	MC	198,245	0.71	0.05	0.37
31	MC	198,223	0.73	0.06	0.54
32	MC	198,251	0.82	0.05	0.44
33	MC	198,254	0.52	0.04	0.56
34	MC	198,267	0.75	0.04	0.49
35	MC	198,082	0.48	0.13	0.45
36	MC	198,137	0.44	0.10	0.53
37	MC	194,382	0.78	0.07	0.48
38	MC	198,101	0.67	0.12	0.41
39	MC	198,233	0.79	0.05	0.48
40	MC	198,206	0.79	0.07	0.58

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

Item	Item Type	N-count	P-value	% Omit	PbisKey
41	MC	198,223	0.75	0.06	0.43
42	MC	198,174	0.67	0.08	0.43
43	MC	198,253	0.83	0.04	0.34
44	MC	198,196	0.44	0.07	0.31
45	MC	198,200	0.78	0.07	0.53
46	MC	198,159	0.62	0.09	0.46
47	MC	198,088	0.78	0.13	0.52
48	MC	197,811	0.75	0.27	0.57
49	CR	198,121	0.74	0.11	
50	CR	198,110	0.78	0.12	
51	CR	198,018	0.67	0.16	
52	CR	197,924	0.41	0.21	
53	CR	198,219	0.80	0.06	
54	CR	198,005	0.56	0.17	
55	CR	197,955	0.51	0.20	
56	CR	197,654	0.52	0.35	
57	CR	198,056	0.60	0.14	

Table 6E. Item Analysis, Grade 7

Item	Item Type	N-count	P-value	% Omit	PbisKey
01	MC	196,145	0.84	0.04	0.36
02	MC	196,161	0.65	0.03	0.19
03	MC	196,141	0.79	0.04	0.43
04	MC	196,022	0.60	0.10	0.38
05	MC	196,114	0.51	0.06	0.55
06	MC	196,122	0.62	0.05	0.43
07	MC	196,128	0.67	0.05	0.39
08	MC	195,936	0.60	0.15	0.32

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

Item	Item Type	N-count	P-value	% Omit	PbisKey
09	MC	196,110	0.75	0.06	0.41
10	MC	196,037	0.80	0.10	0.47
11	MC	196,008	0.62	0.11	0.51
12	MC	196,151	0.80	0.04	0.35
13	MC	196,007	0.38	0.11	0.20
14	MC	196,106	0.53	0.06	0.29
15	MC	195,915	0.48	0.16	0.52
16	MC	196,093	0.74	0.07	0.44
17	MC	196,139	0.69	0.05	0.42
18	MC	196,073	0.75	0.08	0.43
19	MC	196,111	0.62	0.06	0.52
20	MC	196,006	0.62	0.11	0.49
21	MC	195,971	0.66	0.13	0.42
22	MC	195,947	0.44	0.14	0.53
23	MC	195,685	0.91	0.28	0.44
24	MC	196,075	0.81	0.08	0.44
25	MC	196,186	0.86	0.02	0.32
26	MC	196,112	0.65	0.06	0.51
27	MC	195,992	0.42	0.12	0.49
28	MC	195,967	0.71	0.13	0.49
29	MC	196,143	0.85	0.04	0.39
30	MC	195,794	0.50	0.22	0.38
31	MC	196,085	0.63	0.07	0.57
32	MC	196,119	0.80	0.06	0.56
33	MC	196,139	0.48	0.05	0.49
34	MC	196,118	0.62	0.06	0.46
35	MC	196,067	0.57	0.08	0.46
36	MC	196,107	0.64	0.06	0.42
37	MC	196,008	0.64	0.11	0.38

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

Item	Item Type	N-count	P-value	% Omit	PbisKey
38	MC	196,101	0.50	0.06	0.52
39	MC	196,121	0.54	0.05	0.43
40	MC	196,089	0.53	0.07	0.42
41	MC	196,127	0.90	0.05	0.35
42	MC	196,102	0.84	0.06	0.41
43	MC	196,004	0.71	0.11	0.48
44	MC	196,071	0.62	0.08	0.49
45	MC	196,032	0.73	0.10	0.54
46	MC	195,965	0.74	0.13	0.48
47	CR	194,683	0.33	0.79	
48	CR	193,612	0.59	1.33	
49	CR	195,510	0.77	0.37	
50	CR	195,817	0.78	0.21	
51	CR	194,732	0.53	0.76	
52	CR	194,426	0.21	0.92	
53	CR	195,276	0.73	0.49	
54	CR	195,224	0.66	0.51	
55	CR	192,596	0.45	1.85	

Table 6F. Item Analysis, Grade 8

Item	Item Type	N-count	P-value	% Omit	PbisKey
01	MC	196,379	0.89	0.03	0.34
02	MC	196,392	0.90	0.02	0.39
03	MC	196,329	0.70	0.05	0.60
04	MC	196,340	0.76	0.05	0.46
05	MC	196,297	0.67	0.07	0.52
06	MC	196,020	0.39	0.21	0.29
07	MC	196,339	0.36	0.05	0.45
08	MC	196,300	0.63	0.07	0.37

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

Item	Item Type	N-count	P-value	% Omit	PbisKey
03	MC	196,329	0.70	0.05	0.60
04	MC	196,340	0.76	0.05	0.46
05	MC	196,297	0.67	0.07	0.52
06	MC	196,020	0.39	0.21	0.29
07	MC	196,339	0.36	0.05	0.45
08	MC	196,300	0.63	0.07	0.37
09	MC	196,168	0.49	0.14	0.46
10	MC	196,371	0.86	0.03	0.37
11	MC	196,333	0.83	0.05	0.43
12	MC	196,343	0.87	0.05	0.35
13	MC	196,259	0.61	0.09	0.47
14	MC	195,585	0.53	0.06	0.48
15	MC	196,288	0.73	0.07	0.35
16	MC	196,260	0.61	0.06	0.54
17	MC	196,061	0.37	0.19	0.50
18	MC	196,295	0.87	0.07	0.43
19	MC	196,290	0.82	0.07	0.48
20	MC	196,278	0.67	0.08	0.59
21	MC	196,104	0.25	0.17	0.31
22	MC	196,061	0.88	0.19	0.39
23	MC	196,351	0.86	0.04	0.48
24	MC	196,393	0.72	0.02	0.54
25	MC	196,356	0.71	0.04	0.36
26	MC	195,654	0.53	0.03	0.50
27	MC	195,866	0.39	0.29	0.28
28	MC	196,363	0.69	0.04	0.59
29	MC	195,912	0.49	0.27	0.44
30	MC	196,356	0.89	0.04	0.36
31	MC	196,380	0.54	0.03	0.41

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

Item	Item Type	N-count	P-value	% Omit	PbisKey
32	MC	196,283	0.62	0.08	0.58
33	MC	196,257	0.55	0.09	0.55
34	MC	196,295	0.73	0.07	0.51
35	MC	196,335	0.89	0.05	0.44
36	MC	196,305	0.54	0.07	0.60
37	MC	196,339	0.90	0.05	0.32
38	MC	196,217	0.69	0.11	0.57
39	MC	196,318	0.75	0.06	0.49
40	MC	196,295	0.63	0.07	0.36
41	MC	196,367	0.85	0.03	0.38
42	MC	196,315	0.78	0.06	0.48
43	MC	196,245	0.73	0.10	0.58
44	MC	196,284	0.77	0.08	0.57
45	MC	196,343	0.79	0.05	0.39
46	MC	196,205	0.61	0.12	0.47
47	CR	195,432	0.51	0.51	
48	CR	195,674	0.74	0.39	
49	CR	193,843	0.38	1.32	
50	CR	193,933	0.39	1.27	
51	CR	194,910	0.54	0.78	
52	CR	190,383	0.35	3.08	
53	CR	194,252	0.25	1.11	
54	CR	195,557	0.56	0.45	
55	CR	193,632	0.57	1.43	

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 0.20, 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 0.19–0.65. For Grade 3, the point biserials were between 0.20 and 0.59. For Grade 4, the point biserials were between 0.22 and 0.60. For Grade 5, the point biserials were between 0.30 and 0.64. For Grade 6, the point biserials were between 0.23 and 0.65. For Grade 7, the point biserials were between 0.19 and 0.57. For Grade 8, the point biserials were between 0.28 and 0.60.

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 0.92– 0.94. Feldt-Raju reliability coefficients ranged 0.93–0.95. 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 6 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 2012 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	60	44.94	10.71	0.76	0.27	0.94	0.92	0.93
4	69	47.83	13.06	0.73	0.38	0.96	0.93	0.94
5	61	40.79	13.09	0.68	0.37	0.93	0.93	0.94
6	70	45.46	15.85	0.66	0.32	0.94	0.94	0.95
7	68	42.41	14.34	0.63	0.21	0.91	0.93	0.94
8	68	41.43	15.07	0.65	0.25	0.90	0.94	0.95

Speededness

Speededness is the term used to refer to interference in test score observation due to insufficient testing time. Test developers considered speededness in the development of the NYSTP tests. NYSED believes that achievement tests should not be speeded; little or no useful instructional information can be obtained from the fact that a student did not finish a test, while a great deal can be learned from student responses to questions. Further, NYSED prefers all scores to be based on actual student performance, because all students should have ample opportunity to demonstrate that performance to enhance the validity of their scores. Test reliability is directly impacted by the number of test questions, so excluding questions that were impacted by a lack of timing would negatively impact reliability. For these reasons, sufficient administration time

limits were set for the NYSTP tests. The research department at Pearson 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 differential item functioning (DIF) was evaluated using two methods. First, the standardized mean difference (SMD) was computed for all items. The SMD statistic (Dorans, Schmitt, and Bleistein, 1992) compares the mean scores of reference and focal groups, after adjusting for ability differences. A moderate amount of significant DIF, for or against the focal group, is represented by an SMD with an absolute value between 0.10 and 0.19, inclusive. A large amount of practically significant DIF is represented by an SMD with an absolute value of 0.20 or greater. Then, the Mantel-Haenszel method is employed to compute DIF statistics for MC items. This non-parametric DIF method partitions the sample of examinees into categories based on total raw test scores. It then compares the log-odds ratio of keyed responses for the focal and reference groups. The Mantel-Haenszel method has a critical value of 6.63 (degrees of freedom = 1 for MC items; alpha = 0.01) and is compared to its corresponding delta-value (significant when absolute value of delta > 1.50) to factor in effect size (Zwick, Donoghue, and Grima, 1993). It is important to recognize that the two methods differ in assumptions and computation; therefore, the results from both methods may not be in agreement. It should be noted that two methods of classical DIF computation 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 2012 Classical DIF Sample N-Counts

Grade	Ethnicity				Gender		Needs/Resource Capacity Category		Test Language	
	Black/African American	Hispanic	Asian	White	Female	Male	High	Low	Spanish	English
3	31,749	46,419	16,279	93,032	91,655	95,824	104,192	83,287	3,092	184,387
4	31,742	45,007	15,935	92,164	90,468	94,380	101,296	83,552	3,080	181,768
5	32,620	44,299	15,587	93,626	91,163	94,969	101,128	85,004	2,867	183,265
6	33,567	43,684	16,331	95,705	92,362	96,925	100,856	88,431	3,291	185,996
7	34,012	42,909	15,170	95,921	91,654	96,358	993,20	88,692	3,306	184,706
8	34,314	42,813	15,245	97,159	93,148	96,383	992,27	90,304	2,797	186,734

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	1
4	5
5	4
6	5
7	7
8	11

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

Section VI: IRT Scaling and Equating

IRT Models and Rationale for Use

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

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

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

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

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

where

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

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

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

where

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

and

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

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

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

and

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

where

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

and

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

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

Calibration Sample

The cleaned sample data were used for calibration and scaling of New York State Math Tests. It should be noted that the scaling was done on nearly all of the New York State public school student population in each tested grade and that exclusion of some cases during the data cleaning process had minimal effect on parameter estimation. As shown in Tables 10 through 12, the 2012 OP samples were comparable to 2011 populations in terms of the Needs/Resource Capacity Category (NRC), student race and ethnicity, proportions of English language learners, proportions of students with disabilities, and proportions of students using testing accommodations.

Table 10. Grades 3 and 4 Demographic Statistics

Demographics	2011 Grade 3 Population	2012 Grade 3 Sample	2011 Grade 4 Population	2012 Grade 4 Sample
	%	%	%	%
NRC SUBGROUPS				
NYC	37.24	36.90	37.37	36.64
Big 4 Cities	4.24	4.11	4.38	4.19
Urban/Suburban	7.82	7.70	7.67	7.33
Rural	5.34	5.25	4.99	5.39
Average Needs	28.67	28.90	29.11	29.43
Low Needs	14.21	14.20	14.51	14.65
Charter	2.49	2.94	1.96	2.38
ETHNICITY				
Asian	8.41	8.54	8.29	8.57
Black	18.60	18.04	19.03	18.07
Hispanics	23.74	24.26	23.43	23.86
American Indian	0.56	0.55	0.44	0.54
Multiracial	0.79	1.00	0.69	0.86
White	47.76	47.40	47.98	47.93
Other	0.15	0.21	0.14	0.17
ELL STATUS				
No	91.61	90.56	92.38	90.99
Yes	8.39	9.44	7.62	9.01
DISABILITY				
No	85.84	85.91	84.96	84.93
Yes	14.16	14.09	15.04	15.07
ACCOMMODATIONS				
No	74.69	78.80	74.15	78.55
Yes	25.31	21.20	25.85	21.45

Table 11. Grades 5 and 6 Demographic Statistics

Demographics	2011 Grade 5 Population	2012 Grade 5 Sample	2011 Grade 6 Population	2012 Grade 6 Sample
	%	%	%	%
NRC SUBGROUPS				
NYC	36.36	35.64	36.14	35.31
Big 4 Cities	4.11	4.13	4.09	3.96
Urban/Suburban	7.53	7.46	7.32	7.24
Rural	5.03	5.51	5.56	5.41
Average Needs	29.06	29.75	29.26	30.12
Low Needs	15.26	14.49	15.21	15.18
Charter	2.65	3.01	2.43	2.78
ETHNICITY				
Asian	8.73	8.27	8.08	8.57
Black	18.98	18.69	19.34	18.71
Hispanics	22.88	23.47	22.83	22.84
American Indian	0.45	0.50	0.48	0.51
Multiracial	0.64	0.76	0.62	0.71
White	48.20	48.13	48.51	48.48
Other	0.13	0.16	0.14	0.18
ELL STATUS				
No	93.67	92.31	94.69	93.57
Yes	6.33	7.69	5.31	6.43
DISABILITY				
No	84.83	84.52	84.85	84.88
Yes	15.17	15.48	15.15	15.12
ACCOMMODATIONS				
No	74.80	78.58	76.72	80.49
Yes	25.20	21.42	23.28	19.51

Table 12. Grades 7 and 8 Demographic Statistics

Demographics	2011 Grade 7 Population	2012 Grade 7 Sample	2011 Grade 8 Population	2012 Grade 8 Sample
	%	%	%	%
NRC SUBGROUPS				
NYC	35.34	34.91	35.95	35.57
Big 4 Cities	3.90	3.92	3.78	3.76
Urban/Suburban	7.66	7.23	7.48	6.71
Rural	5.79	5.63	5.80	5.57
Average Needs	31.21	29.93	30.87	30.27
Low Needs	14.65	16.00	14.95	16.38
Charter	1.46	2.38	1.16	1.73
ETHNICITY				
Asian	7.88	8.09	7.91	8.19
Black	19.18	18.89	18.80	18.64
Hispanics	20.99	22.54	20.96	22.23
American Indian	0.48	0.51	0.45	0.53
Multiracial	0.36	0.69	0.30	0.57
White	51.08	49.11	51.53	49.66
Other	0.04	0.17	0.05	0.18
ELL STATUS				
No	95.59	93.73	95.84	93.79
Yes	4.41	6.27	4.16	6.21
DISABILITY				
No	85.00	85.04	85.21	85.21
Yes	15.00	14.96	14.79	14.79
ACCOMMODATIONS				
No	78.57	81.35	78.90	82.03
Yes	21.43	18.65	21.10	17.97

Calibration Process

The item parameters were estimated using MULTILOG software (Thissen, 1991). MC and CR items were calibrated simultaneously using marginal maximum likelihood procedures.

The NYSTP Mathematics Tests did not incur anything problematic during item calibration. The number of estimation cycles was set to 120 for all grades with convergence criterion of 0.001 for all grades. The estimated parameters were in the original theta metric, and all the items were well within the prescribed parameter ranges. For the Grades 3–8 Mathematics Tests, all calibration estimation results are reasonable. The summary of calibration results is presented in Table 13.

Table 13. NYSTP ELA 2012 Calibration Results

Grade	Largest *a*-parameter	*b*-parameter/ Gamma Range		Theta Mean	Theta Standard Deviation	# Students
3	2.397	–3.237	3.510	–0.01	0.922	197,344
4	2.716	–2.512	3.767	–0.00	0.943	193,123
5	2.731	–2.087	3.218	–0.00	0.930	195,421
6	3.478	–3.331	3.056	0.01	0.947	198,342
7	2.601	–3.665	3.441	0.01	0.946	196,228
8	3.473	–4.052	3.025	0.01	0.956	196,435

Item-Model Fit

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

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

with

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

A modification of this procedure was used to measure fit to the 2PPC model. For the 2PPC model, Q_{1j} was assumed to have approximately a chi-square distribution with the following degrees of freedom (*df*):

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

where

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

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

where

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

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

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

where

N is the calibration sample size.

To compute the Q1 and related statistics, a stratified sampling procedure was implemented in a way that a representative sample with the size of approximately 700,000 students were drawn at each grade level. Items were considered to have poor fit if the value of the obtained Z_{Q_1} was greater than the value of Z_{Q_1} critical. If the obtained Z_{Q_1} was less than Z_{Q_1} critical, the items were rated as having acceptable fit. All items in the NYSTP 2012 Mathematics Tests for Grades 3 and 8 demonstrated good model fit. Item 50 in Grade 4, item 48 in Grade 5, items 50 and 55 in Grade 6, and item 2 in Grade 7 exhibited poor item-model fit statistics. The fact that so few items were flagged for poor fit across all NYSTP 2012 Mathematics Tests further supports the use of the chosen models. Item fit statistics are presented in Tables 14–19.

Table 14. Mathematics Grade 3 Item Fit Statistics

Item	Model	Chi Square	DF	Z_{Q1}	Z_{Q1} critical	Fit OK?
1	3PL	54.01	7	12.56	184.19	Y
2	3PL	39.33	7	8.64	184.19	Y
3	3PL	220.95	7	57.18	184.19	Y
4	3PL	252.60	7	65.64	184.19	Y
5	3PL	449.20	7	118.18	184.19	Y
6	3PL	170.73	7	43.76	184.19	Y
7	3PL	266.67	7	69.40	184.19	Y
8	3PL	82.64	7	20.22	184.19	Y
9	3PL	93.89	7	23.22	184.19	Y
10	3PL	105.10	7	26.22	184.19	Y
11	3PL	237.83	7	61.69	184.19	Y
12	3PL	85.17	7	20.89	184.19	Y
13	3PL	79.00	7	19.24	184.19	Y
14	3PL	199.60	7	51.48	184.19	Y
15	3PL	122.34	7	30.83	184.19	Y
16	3PL	181.60	7	46.66	184.19	Y
17	3PL	123.51	7	31.14	184.19	Y
18	3PL	66.67	7	15.95	184.19	Y
19	3PL	69.51	7	16.71	184.19	Y
20	3PL	276.58	7	72.05	184.19	Y
21	3PL	120.09	7	30.22	184.19	Y
22	3PL	94.13	7	23.29	184.19	Y
23	3PL	39.41	7	8.66	184.19	Y
24	3PL	125.74	7	31.73	184.19	Y
25	3PL	193.98	7	49.97	184.19	Y
26	3PL	109.42	7	27.37	184.19	Y
27	3PL	68.16	7	16.35	184.19	Y
28	3PL	294.89	7	76.94	184.19	Y
29	3PL	87.29	7	21.46	184.19	Y
30	3PL	450.54	7	118.54	184.19	Y
31	3PL	499.24	7	131.56	184.19	Y
32	3PL	211.29	7	54.60	184.19	Y
33	3PL	173.22	7	44.42	184.19	Y
34	3PL	337.71	7	88.39	184.19	Y
35	3PL	57.65	7	13.54	184.19	Y
36	3PL	194.70	7	50.16	184.19	Y
37	3PL	92.11	7	22.75	184.19	Y
38	3PL	161.86	7	41.39	184.19	Y
39	3PL	89.29	7	21.99	184.19	Y

Table 14. Mathematics Grade 3 Item Fit Statistics (cont.)

Item	Model	Chi Square	DF	Z_{Q1}	Z_{Q1} critical	Fit OK?
40	3PL	100.47	7	24.98	184.19	Y
41	3PL	189.31	7	48.73	184.19	Y
42	3PL	163.05	7	41.71	184.19	Y
43	3PL	121.89	7	30.70	184.19	Y
44	2PP	543.74	17	93.29	184.19	Y
45	2PP	390.06	17	66.12	184.19	Y
46	2PP	827.50	17	143.45	184.19	Y
47	2PP	165.74	17	26.47	184.19	Y
48	2PP	432.64	26	57.65	184.19	Y
49	2PP	507.50	26	68.24	184.19	Y
50	2PP	819.06	26	112.30	184.19	Y

Table 15. Mathematics Grade 4 Item Fit Statistics

Item	Model	Chi Square	DF	Z_{Q1}	Z_{Q1} critical	Fit OK?
1	3PL	39.94	7	8.80	180.25	Y
2	3PL	71.17	7	17.15	180.25	Y
3	3PL	50.37	7	11.59	180.25	Y
4	3PL	86.83	7	21.34	180.25	Y
5	3PL	131.13	7	33.17	180.25	Y
6	3PL	264.85	7	68.91	180.25	Y
7	3PL	38.43	7	8.40	180.25	Y
8	3PL	59.46	7	14.02	180.25	Y
9	3PL	78.70	7	19.16	180.25	Y
10	3PL	28.86	7	5.84	180.25	Y
11	3PL	101.03	7	25.13	180.25	Y
12	3PL	94.04	7	23.26	180.25	Y
13	3PL	111.20	7	27.85	180.25	Y
14	3PL	143.78	7	36.56	180.25	Y
15	3PL	380.10	7	99.72	180.25	Y
16	3PL	137.72	7	34.94	180.25	Y
17	3PL	182.71	7	46.96	180.25	Y
18	3PL	130.66	7	33.05	180.25	Y
19	3PL	74.00	7	17.91	180.25	Y
20	3PL	84.96	7	20.83	180.25	Y
21	3PL	45.34	7	10.25	180.25	Y
22	3PL	223.03	7	57.74	180.25	Y
23	3PL	105.92	7	26.44	180.25	Y

Table 15. Mathematics Grade 4 Item Fit Statistics (cont.)

Item	Model	Chi Square	DF	Z_{Q1}	Z_{Q1} critical	Fit OK?
24	3PL	82.21	7	20.10	180.25	Y
25	3PL	150.06	7	38.23	180.25	Y
26	3PL	69.87	7	16.80	180.25	Y
27	3PL	95.21	7	23.58	180.25	Y
28	3PL	67.74	7	16.23	180.25	Y
29	3PL	187.06	7	48.12	180.25	Y
30	3PL	256.23	7	66.61	180.25	Y
31	3PL	79.51	7	19.38	180.25	Y
32	3PL	115.47	7	28.99	180.25	Y
33	3PL	100.42	7	24.97	180.25	Y
34	3PL	86.09	7	21.14	180.25	Y
35	3PL	63.57	7	15.12	180.25	Y
36	3PL	99.57	7	24.74	180.25	Y
37	3PL	54.14	7	12.60	180.25	Y
38	3PL	73.16	7	17.68	180.25	Y
39	3PL	43.56	7	9.77	180.25	Y
40	3PL	93.30	7	23.06	180.25	Y
41	3PL	121.35	7	30.56	180.25	Y
42	3PL	117.57	7	29.55	180.25	Y
43	3PL	42.58	7	9.51	180.25	Y
44	3PL	119.91	7	30.18	180.25	Y
45	3PL	75.00	7	18.17	180.25	Y
46	3PL	130.18	7	32.92	180.25	Y
47	3PL	201.48	7	51.98	180.25	Y
48	2PP	235.38	17	38.78	180.25	Y
49	2PP	145.96	17	22.97	180.25	Y
50	2PP	1,399.93	17	244.65	180.25	N
51	2PP	740.71	17	128.11	180.25	Y
52	2PP	720.04	17	124.46	180.25	Y
53	2PP	588.32	26	79.67	180.25	Y
54	2PP	822.07	26	112.72	180.25	Y
55	2PP	778.86	26	106.61	180.25	Y
56	2PP	652.15	26	88.69	180.25	Y

Table 16. Mathematics Grade 5 Item Fit Statistics

Item	Model	Chi Square	DF	Z_{Q1}	Z_{Q1} critical	Fit OK?
1	3PL	75.75	7	18.38	182.39	Y
2	3PL	89.82	7	22.13	182.39	Y
3	3PL	145.45	7	37.00	182.39	Y
4	3PL	123.84	7	31.23	182.39	Y
5	3PL	123.22	7	31.06	182.39	Y
6	3PL	223.05	7	57.74	182.39	Y
7	3PL	100.27	7	24.93	182.39	Y
8	3PL	147.69	7	37.60	182.39	Y
9	3PL	88.60	7	21.81	182.39	Y
10	3PL	154.63	7	39.46	182.39	Y
11	3PL	143.86	7	36.58	182.39	Y
12	3PL	178.93	7	45.95	182.39	Y
13	3PL	78.90	7	19.22	182.39	Y
14	3PL	82.76	7	20.25	182.39	Y
15	3PL	84.58	7	20.73	182.39	Y
16	3PL	129.42	7	32.72	182.39	Y
17	3PL	193.96	7	49.97	182.39	Y
18	3PL	164.76	7	42.16	182.39	Y
19	3PL	286.11	7	74.60	182.39	Y
20	3PL	131.59	7	33.30	182.39	Y
21	3PL	186.66	7	48.02	182.39	Y
22	3PL	188.29	7	48.45	182.39	Y
23	3PL	211.72	7	54.71	182.39	Y
24	3PL	121.69	7	30.65	182.39	Y
25	3PL	109.81	7	27.48	182.39	Y
26	3PL	167.63	7	42.93	182.39	Y
27	3PL	134.45	7	34.06	182.39	Y
28	3PL	191.02	7	49.18	182.39	Y
29	3PL	134.46	7	34.06	182.39	Y
30	3PL	150.55	7	38.37	182.39	Y
31	3PL	177.25	7	45.50	182.39	Y
32	3PL	207.44	7	53.57	182.39	Y
33	3PL	93.31	7	23.07	182.39	Y
34	3PL	70.21	7	16.89	182.39	Y
35	3PL	165.50	7	42.36	182.39	Y
36	3PL	164.60	7	42.12	182.39	Y
37	3PL	118.86	7	29.90	182.39	Y
38	3PL	150.24	7	38.28	182.39	Y
39	3PL	115.36	7	28.96	182.39	Y
40	3PL	117.31	7	29.48	182.39	Y
41	3PL	239.13	7	62.04	182.39	Y
42	3PL	489.66	7	129.00	182.39	Y

Table 16. Mathematics Grade 5 Item Fit Statistics (cont.)

Item	Model	Chi Square	DF	Z_{Q1}	Z_{Q1} critical	Fit OK?
43	3PL	106.82	7	26.68	182.39	Y
44	3PL	77.79	7	18.92	182.39	Y
45	2PP	1,020.61	17	177.59	182.39	Y
46	2PP	174.94	17	28.10	182.39	Y
47	2PP	391.55	17	66.39	182.39	Y
48	2PP	1,383.08	17	241.67	182.39	N
49	2PP	810.30	26	111.06	182.39	Y
50	2PP	329.49	26	43.06	182.39	Y
51	2PP	907.42	26	124.79	182.39	Y

Table 17. Mathematics Grade 6 Item Fit Statistics

Item	Model	Chi Square	DF	Z_{Q1}	Z_{Q1} critical	Fit OK?
1	3PL	105.20	7	26.24	185.12	Y
2	3PL	54.47	7	12.69	185.12	Y
3	3PL	79.55	7	19.39	185.12	Y
4	3PL	71.32	7	17.19	185.12	Y
5	3PL	122.39	7	30.84	185.12	Y
6	3PL	66.70	7	15.96	185.12	Y
7	3PL	76.43	7	18.56	185.12	Y
8	3PL	82.04	7	20.05	185.12	Y
9	3PL	131.90	7	33.38	185.12	Y
10	3PL	261.00	7	67.88	185.12	Y
11	3PL	87.52	7	21.52	185.12	Y
12	3PL	121.81	7	30.68	185.12	Y
13	3PL	205.61	7	53.08	185.12	Y
14	3PL	139.63	7	35.45	185.12	Y
15	3PL	399.54	7	104.91	185.12	Y
16	3PL	61.15	7	14.47	185.12	Y
17	3PL	198.93	7	51.30	185.12	Y
18	3PL	109.97	7	27.52	185.12	Y
19	3PL	79.93	7	19.49	185.12	Y
20	3PL	95.15	7	23.56	185.12	Y
21	3PL	89.08	7	21.94	185.12	Y
22	3PL	122.76	7	30.94	185.12	Y
23	3PL	72.33	7	17.46	185.12	Y
24	3PL	76.29	7	18.52	185.12	Y
25	3PL	75.52	7	18.31	185.12	Y
26	3PL	105.06	7	26.21	185.12	Y
27	3PL	128.87	7	32.57	185.12	Y
28	3PL	49.72	7	11.42	185.12	Y
29	3PL	287.22	7	74.89	185.12	Y
30	3PL	99.56	7	24.74	185.12	Y

Table 17. Mathematics Grade 6 Item Fit Statistics (cont.)

Item	Model	Chi Square	DF	Z_{Q1}	Z_{Q1} critical	Fit OK?
31	3PL	72.98	7	17.63	185.12	Y
32	3PL	50.08	7	11.51	185.12	Y
33	3PL	115.72	7	29.06	185.12	Y
34	3PL	97.37	7	24.15	185.12	Y
35	3PL	152.34	7	38.84	185.12	Y
36	3PL	168.58	7	43.18	185.12	Y
37	3PL	70.54	7	16.98	185.12	Y
38	3PL	110.27	7	27.60	185.12	Y
39	3PL	81.58	7	19.93	185.12	Y
40	3PL	120.08	7	30.22	185.12	Y
41	3PL	137.52	7	34.88	185.12	Y
42	3PL	122.18	7	30.78	185.12	Y
43	3PL	57.91	7	13.61	185.12	Y
44	3PL	653.40	7	172.76	185.12	Y
45	3PL	109.55	7	27.41	185.12	Y
46	3PL	100.78	7	25.06	185.12	Y
47	3PL	98.90	7	24.56	185.12	Y
48	3PL	75.81	7	18.39	185.12	Y
49	2PP	242.88	17	40.11	185.12	Y
50	2PP	1,363.01	17	238.12	185.12	N
51	2PP	237.73	17	39.20	185.12	Y
52	2PP	599.61	17	103.17	185.12	Y
53	2PP	486.04	17	83.09	185.12	Y
54	2PP	795.66	26	108.99	185.12	Y
55	2PP	1,551.72	26	215.91	185.12	N
56	2PP	350.18	26	45.99	185.12	Y
57	2PP	1,231.01	26	170.55	185.12	Y

Table 18. Mathematics Grade 7 Item Fit Statistics

Item	Model	Chi Square	DF	Z_{Q1}	Z_{Q1} critical	Fit OK?
1	3PL	66.57	7	15.92	183.14	Y
2	3PL	831.50	7	220.36	183.14	N
3	3PL	81.37	7	19.88	183.14	Y
4	3PL	240.34	7	62.36	183.14	Y
5	3PL	177.52	7	45.57	183.14	Y
6	3PL	96.58	7	23.94	183.14	Y
7	3PL	60.71	7	14.35	183.14	Y
8	3PL	36.75	7	7.95	183.14	Y
9	3PL	115.95	7	29.12	183.14	Y
10	3PL	78.02	7	18.98	183.14	Y
11	3PL	136.12	7	34.51	183.14	Y
12	3PL	154.28	7	39.36	183.14	Y

Table 18. Mathematics Grade 7 Item Fit Statistics (cont.)

Item	Model	Chi Square	DF	Z_{Q1}	Z_{Q1} critical	Fit OK?
13	3PL	399.87	7	105.00	183.14	Y
14	3PL	86.43	7	21.23	183.14	Y
15	3PL	210.82	7	54.47	183.14	Y
16	3PL	88.83	7	21.87	183.14	Y
17	3PL	88.71	7	21.84	183.14	Y
18	3PL	50.33	7	11.58	183.14	Y
19	3PL	89.91	7	22.16	183.14	Y
20	3PL	99.67	7	24.77	183.14	Y
21	3PL	117.17	7	29.44	183.14	Y
22	3PL	123.01	7	31.01	183.14	Y
23	3PL	116.24	7	29.20	183.14	Y
24	3PL	225.86	7	58.49	183.14	Y
25	3PL	95.99	7	23.78	183.14	Y
26	3PL	102.98	7	25.65	183.14	Y
27	3PL	98.74	7	24.52	183.14	Y
28	3PL	91.36	7	22.55	183.14	Y
29	3PL	91.32	7	22.54	183.14	Y
30	3PL	131.37	7	33.24	183.14	Y
31	3PL	110.20	7	27.58	183.14	Y
32	3PL	175.42	7	45.01	183.14	Y
33	3PL	146.50	7	37.28	183.14	Y
34	3PL	83.57	7	20.47	183.14	Y
35	3PL	86.79	7	21.33	183.14	Y
36	3PL	77.98	7	18.97	183.14	Y
37	3PL	118.92	7	29.91	183.14	Y
38	3PL	124.51	7	31.41	183.14	Y
39	3PL	125.64	7	31.71	183.14	Y
40	3PL	111.58	7	27.95	183.14	Y
41	3PL	262.75	7	68.35	183.14	Y
42	3PL	307.98	7	80.44	183.14	Y
43	3PL	129.10	7	32.63	183.14	Y
44	3PL	107.12	7	26.76	183.14	Y
45	3PL	128.28	7	32.41	183.14	Y
46	3PL	55.72	7	13.02	183.14	Y
47	2PP	514.85	17	88.19	183.14	Y
48	2PP	594.15	17	102.20	183.14	Y
49	2PP	624.50	17	107.57	183.14	Y
50	2PP	263.80	17	43.81	183.14	Y
51	2PP	559.21	17	96.03	183.14	Y
52	2PP	660.62	26	89.89	183.14	Y
53	2PP	532.18	26	71.73	183.14	Y
54	2PP	766.87	26	104.92	183.14	Y
55	2PP	603.97	26	81.88	183.14	Y

Table 19. Mathematics Grade 8 Item Fit Statistics

Item	Model	Chi Square	DF	Z_{Q1}	Z_{Q1} critical	Fit OK?
1	3PL	117.00	7	29.40	183.34	Y
2	3PL	169.17	7	43.34	183.34	Y
3	3PL	101.14	7	25.16	183.34	Y
4	3PL	113.54	7	28.48	183.34	Y
5	3PL	63.18	7	15.01	183.34	Y
6	3PL	128.73	7	32.53	183.34	Y
7	3PL	192.17	7	49.49	183.34	Y
8	3PL	145.66	7	37.06	183.34	Y
9	3PL	127.87	7	32.30	183.34	Y
10	3PL	44.00	7	9.89	183.34	Y
11	3PL	108.42	7	27.10	183.34	Y
12	3PL	163.40	7	41.80	183.34	Y
13	3PL	118.57	7	29.82	183.34	Y
14	3PL	99.06	7	24.60	183.34	Y
15	3PL	51.19	7	11.81	183.34	Y
16	3PL	90.58	7	22.34	183.34	Y
17	3PL	177.98	7	45.70	183.34	Y
18	3PL	168.40	7	43.14	183.34	Y
19	3PL	124.65	7	31.44	183.34	Y
20	3PL	97.07	7	24.07	183.34	Y
21	3PL	135.99	7	34.47	183.34	Y
22	3PL	94.64	7	23.42	183.34	Y
23	3PL	176.53	7	45.31	183.34	Y
24	3PL	86.97	7	21.37	183.34	Y
25	3PL	189.40	7	48.75	183.34	Y
26	3PL	122.54	7	30.88	183.34	Y
27	3PL	515.90	7	136.01	183.34	Y
28	3PL	79.02	7	19.25	183.34	Y
29	3PL	203.73	7	52.58	183.34	Y
30	3PL	56.80	7	13.31	183.34	Y
31	3PL	186.25	7	47.91	183.34	Y
32	3PL	125.68	7	31.72	183.34	Y
33	3PL	138.32	7	35.10	183.34	Y
34	3PL	83.31	7	20.39	183.34	Y
35	3PL	173.34	7	44.46	183.34	Y
36	3PL	149.17	7	38.00	183.34	Y
37	3PL	73.93	7	17.89	183.34	Y
38	3PL	121.29	7	30.55	183.34	Y
39	3PL	59.47	7	14.02	183.34	Y
40	3PL	150.39	7	38.32	183.34	Y

Table 19. Mathematics Grade 8 Item Fit Statistics (cont.)

Item	Model	Chi Square	DF	Z_{Q1}	Z_{Q1} critical	Fit OK?
41	3PL	70.99	7	17.10	183.34	Y
42	3PL	77.55	7	18.85	183.34	Y
43	3PL	84.84	7	20.80	183.34	Y
44	3PL	170.11	7	43.59	183.34	Y
45	3PL	72.63	7	17.54	183.34	Y
46	3PL	146.74	7	37.35	183.34	Y
47	2PP	628.10	17	108.20	183.34	Y
48	2PP	294.18	17	49.18	183.34	Y
49	2PP	508.57	17	87.07	183.34	Y
50	2PP	561.84	17	96.49	183.34	Y
51	2PP	445.68	17	75.96	183.34	Y
52	2PP	649.84	26	88.37	183.34	Y
53	2PP	425.92	26	56.70	183.34	Y
54	2PP	323.13	26	42.16	183.34	Y
55	2PP	1,073.60	26	148.29	183.34	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 significant for local dependency. 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 primary concern about locally dependent items is that they contribute less psychometric information about examinee proficiency than do locally independent items and they inflate score reliability estimates.

The Q_3 statistics were examined on all the Grades 3–8 Mathematics Tests and no items were found to be significant for local dependency in Grades 4. In Grade 3, three pairs of items were found to be significant for local dependency: items 5 and 31 (*Q3* = 0.350), items 36 and 37 (*Q3* = 0.248), and items 44 and 45 (*Q3* = 0.207). In Grade 5, one pair of items was found to be significant for local dependency: items 20 and 22 (*Q3* = 0.218). In Grade 6, three pairs of items were found to be significant for local dependency: items 2 and 11 (*Q3* = 0.266), items 3 and 26 (*Q3* = 0.215), and items 29 and 47 (*Q3* = 0.201). In Grade 7, one pair of items was found to be significant for local dependency: items 6 and 17 (*Q3* = 0.247). In Grade 8, five pairs of items were found to be significant for local dependency: items 14 and 26 (*Q3* = 0.284), items 20 and 28 (*Q3* = 0.247), items 20 and 43 (*Q3* = 0.249), items 36 and 47 (*Q3* = 0.270), and items 51 and 55 (*Q3* = 0.450). The magnitudes of these statistics were not sufficient to warrant any concern.

Scaling and Equating

For the 2012 equating, all the viable multiple choice items on the operational test form are eligible to be anchor items. The IRT linking is conducted through the equated field test item parameter estimates and newly calibrated operational item parameter estimates. That is, equated item parameter estimates from 2011 stand alone field testing and newly calibrated item parameter estimates from 2012 operational testing are used to establish the equating relationship. Students' motivation tends to be different at stand alone field testing compared with operational testing, and such motivation effect maybe impact the equating relationship.
In an attempt to control for the field test motivation effects, an evaluation of the 2011 stand alone field test data was conducted. In this analysis, Pearson psychometricians identified a percentage of the students within each grade with the largest relative differences in performance between their 2011 operational test performance and their 2011 stand alone field test performance. In discussions with the NYSED, two testing experts serving on the NYSED's Technical Advisory Committee, and a principal scientist from HumRRO, a decision was made to remove these students from the field test data and re-calibrated the field-test data. Approximately fix percent of the students were removed from the field test data for math grades 3 to 4 and fifteen percent of the students from the field test samples for grades 5 to 8. By removing students that showed low motivation to perform their best on the 2011 field test, it was hypothesized that potential biasing effects on the 2012 equating could be mitigated.

For the initial item equating process, all the anchor items were used. Procedurally, item parameters for the anchor items obtained in 2011 field test equating were compared with the item parameters calibrated using the 2011 OP data. The equating of 2012 OP data to the New York

State scale was performed using a test characteristic curve (TCC) method (Stocking and Lord, 1983) and implemented in STUIRT (Kim, & Kolen, 2004).

For all the OP items, item parameters in scale score metric (i.e., the New York State scale) were obtained via linear transformation of theta metric parameters using the M1 and M2 transformation constants.

This equating process was repeated during the "anchor set evaluation" step (see Equating Specification Section 6.6). The final M1 and M2 were obtained after the final anchor set is determined. Table 20 presents the 2012 OP transformation constants for New York State Grades 3–8 Math Tests.

Table 20. NYSTP Mathematics 2012 Final Transformation Constants

Grade	*M1*	*M2*
3	17.86	687.67
4	30.40	690.06
5	30.49	687.74
6	29.05	683.63
7	27.15	679.56
8	25.74	680.24

Anchor Item Evaluation

Anchor item set was evaluated using several procedures. Note that we used the first two procedures to conduct an overall evaluation and the procedures 3 and 4 for the evaluation at the item level.

1. Anchor set previous and current estimates of TCC alignment. The overall alignment of TCCs for the anchor set previous and current estimates were evaluated to determine the overall stability of anchor item parameters between the 2011 FT administration and 2012 OP administration.

2. Correlations of anchor previous and current estimates of *a*- and *b*-parameters. Correlations of anchor previous and current estimate of *a*- and *b*-parameters were evaluated for magnitude. Ideally, the correlations between the two sets of estimates for the *a*-parameter should be at least 0.80, and the correlations for the *b*-parameters should be at least 0.90. Scatter plots were generated for checking on outlier items.

3. Item Fit Plots. Item-fit plots (using theta values from MULTILOG) were used to evaluate the appropriateness of using an item in the 3PL or 2PPC model. Poor-fit items were be flagged, and decisions were made whether or not to include the poor-fit item(s) in the stability check (see Step 4).

4. Stability Check (i.e., Iterative evaluation of difference in ICCs). This procedure minimizes the weighted squared differences between the two ICCs for each MC item: one based on 2011 FT

item parameter estimates and the other on 2012 estimates. The differential item performance was evaluated by examining previous and current item parameters. Primarily the following steps were taken:

1. Before the iterative procedures start, the initial equating should be performed using all the ***eligible*** OP MC items as anchor items. The initial M1 and M2 were obtained through the Stocking-Lord method (save as v0). Create the raw-to-scale score table, and save this table as the first version (v0). Identify the raw score cut associated with each level of the cut score (save as v0). Particular attention should be given to Level 3 cut.

2. For each anchor item, calculate a weighted sum of the squared deviation between the ICCs based on old (x) and new (y) parameters at each point of a normal theta distribution:

$$d_i^{\,2} = \sum^{k} \left[P_{ix}(\theta_k) - P_{iy}(\theta_k)\right]^2 \bullet g(\theta_k).$$

3. Create a table (Excel or SAS dataset) which includes for each of the anchor items, the original set of parameter values, the new set of parameter values, the associated d^2, the position of the item on the test in 2011 and 2012, and the deviation in the position of the anchor item from field testing in 2011 to operational testing in 2012 (save as v0).

4. Proceed through the following process for <u>five iterations</u>:

    a. Sort the table created in Step 3 by d^2 in descending order. Remove the item having the largest d^2;
    b. Recalculate the Stocking-Lord constants (M1 and M2) with the remaining anchor items (save as v1v5 corresponding to each iteration);
    c. Apply the new constants to the operational parameters,
    d. Recalculate the weighted sum of the squared deviation between the ICCs. Because the relative item ranking of d^2 may change, this step must be done for each iteration in order to eliminate the correct anchor for the next iteration.
    e. Calculate the new RS to SS table (save as v1–v5).
    f. Identify the raw score associated with each cut (save as v1–v5).
    g. Iterate through this process until the first five items with the largest d^2 were removed.

5. Identify the point in the iterative process where:
    - the raw score associated with the Level 3 cut score does not change for two iterations in a row, OR
    - the raw score associated with the Level 3 cut score changes back to a previously established value

The items flagged based on each of the procedures described above were summarized and evaluated. Based on the evaluation results a decision was made to remove items with D square values at or above 0.05, which led to one items being removed from the anchor set for math grades 6 and 8 and five items being removed from the anchor set for grade 7.

Item Parameters

The OP test item parameters were estimated by the software MULTILOG (Thissen, 1991) and are presented in Tables 21 to 26. The parameter estimates are expressed in scale score metric and are defined below:

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

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

Table 21. Grade 3 2012 Operational Item Parameter Estimates

Item	Max Pts	a-par/ alpha	b-par/ gamma1	c-par/ gamma2	gamma3
01	1	0.047	651.072	0.046	
02	1	0.044	640.415	0.025	
03	1	0.045	648.473	0.005	
04	1	0.057	683.672	0.242	
05	1	0.071	691.220	0.263	
06	1	0.053	683.058	0.559	
07	1	0.061	683.464	0.127	
08	1	0.030	653.610	0.064	
09	1	0.036	661.303	0.277	
10	1	0.042	657.970	0.016	
11	1	0.068	681.322	0.100	
12	1	0.048	666.125	0.146	
13	1	0.040	656.057	0.282	
14	1	0.056	683.309	0.251	
15	1	0.066	670.128	0.088	
16	1	0.034	677.972	0.432	
17	1	0.053	673.424	0.242	
18	1	0.053	659.159	0.238	

Table 21. Grade 3 2012 Operational Item Parameter Estimates (cont.)

Item	Max Pts	a-par/ alpha	b-par/ gamma1	c-par/ gamma2	gamma3
19	1	0.052	659.619	0.074	
20	1	0.049	686.273	0.308	
21	1	0.073	664.991	0.162	
22	1	0.038	671.853	0.144	
23	1	0.046	646.104	0.024	
24	1	0.040	670.819	0.558	
25	1	0.040	658.292	0.005	
26	1	0.056	680.975	0.505	
27	1	0.011	663.453	0.021	
28	1	0.052	687.436	0.129	
29	1	0.035	664.331	0.263	
30	1	0.046	710.293	0.069	
31	1	0.079	689.479	0.231	
32	1	0.059	674.040	0.267	
33	1	0.060	676.318	0.179	
34	1	0.038	698.300	0.146	
35	1	0.040	651.203	0.022	
36	1	0.067	675.504	0.108	
37	1	0.066	669.810	0.137	
38	1	0.026	637.823	0.011	
39	1	0.072	666.534	0.286	
40	1	0.052	669.868	0.207	
41	1	0.076	676.116	0.137	
42	1	0.047	681.274	0.172	
43	1	0.035	672.867	0.054	
44	2	0.077	50.448	51.832	
45	2	0.087	57.340	58.662	
46	2	0.056	35.454	40.912	
47	2	0.052	32.887	34.393	

Table 21. Grade 3 2012 Operational Item Parameter Estimates (cont.)

Item	Max Pts	a-par/ alpha	b-par/ gamma1	c-par/ gamma2	gamma3
48	3	0.050	32.521	33.290	34.545
49	3	0.064	42.397	42.201	43.969
50	3	0.059	39.405	39.917	40.130

Table 22. Grade 4 2012 Operational Item Parameter Estimates

Item	Max Pts	a-par/alpha	b-par/ gamma1	c-par/ gamma2	gamma3
01	1	0.024	598.367	0.023	
02	1	0.032	674.266	0.252	
03	1	0.027	618.500	0.039	
04	1	0.030	649.966	0.217	
05	1	0.027	688.352	0.169	
06	1	0.045	701.907	0.107	
07	1	0.029	617.125	0.125	
08	1	0.040	640.589	0.168	
09	1	0.024	628.845	0.243	
10	1	0.029	646.147	0.478	
11	1	0.040	673.373	0.310	
12	1	0.020	661.070	0.012	
13	1	0.039	686.981	0.254	
14	1	0.036	674.322	0.132	
15	1	0.031	708.109	0.168	
16	1	0.027	693.705	0.219	
17	1	0.030	630.640	0.007	
18	1	0.025	694.280	0.146	
19	1	0.039	662.673	0.279	
20	1	0.025	667.714	0.242	
21	1	0.028	656.281	0.179	
22	1	0.036	730.810	0.362	
23	1	0.036	676.907	0.247	

Table 22. Grade 4 2012 Operational Item Parameter Estimates (cont.)

Item	Max Pts	a-par/alpha	b-par/ gamma1	c-par/ gamma2	gamma3
24	1	0.022	611.755	0.014	
25	1	0.043	676.082	0.213	
26	1	0.031	644.808	0.084	
27	1	0.026	667.499	0.202	
28	1	0.023	654.174	0.228	
29	1	0.031	683.236	0.229	
30	1	0.041	697.466	0.137	
31	1	0.038	644.883	0.124	
32	1	0.028	695.518	0.448	
33	1	0.030	660.840	0.085	
34	1	0.019	630.164	0.031	
35	1	0.024	628.257	0.037	
36	1	0.039	659.920	0.236	
37	1	0.031	651.657	0.263	
38	1	0.016	696.884	0.161	
39	1	0.024	639.652	0.323	
40	1	0.032	670.056	0.205	
41	1	0.053	671.507	0.219	
42	1	0.017	639.826	0.010	
43	1	0.034	668.830	0.451	
44	1	0.036	659.907	0.091	
45	1	0.041	652.272	0.237	
46	1	0.026	657.041	0.050	
47	1	0.042	690.603	0.049	
48	2	0.037	24.077	25.301	
49	2	0.024	15.521	15.803	
50	2	0.028	18.910	19.633	
51	2	0.037	23.634	25.391	
52	2	0.046	30.197	34.551	

Table 22. Grade 4 2012 Operational Item Parameter Estimates (cont.)

Item	Max Pts	a-par/alpha	b-par/ gamma1	c-par/ gamma2	gamma3
53	3	0.034	22.297	24.445	24.128
54	3	0.034	22.739	24.489	22.947
55	3	0.030	20.685	20.407	19.737
56	3	0.052	34.709	37.881	38.086

Table 23. Grade 5 2012 Operational Item Parameter Estimates

Item	Max Pts	a-par/alpha	b-par/ gamma1	c-par/ gamma2	gamma3
01	1	0.024	608.284	0.016	
02	1	0.027	645.678	0.294	
03	1	0.037	684.993	0.200	
04	1	0.043	677.957	0.336	
05	1	0.045	657.093	0.160	
06	1	0.043	698.797	0.220	
07	1	0.032	657.758	0.212	
08	1	0.035	670.829	0.059	
09	1	0.017	621.062	0.048	
10	1	0.045	682.712	0.170	
11	1	0.022	688.494	0.197	
12	1	0.052	677.751	0.088	
13	1	0.020	676.123	0.137	
14	1	0.027	673.690	0.154	
15	1	0.032	667.149	0.134	
16	1	0.024	700.296	0.165	
17	1	0.035	695.578	0.123	
18	1	0.048	688.341	0.285	
19	1	0.038	720.445	0.205	
20	1	0.042	673.271	0.094	
21	1	0.034	696.933	0.072	
22	1	0.048	661.743	0.086	

Table 23. Grade 5 2012 Operational Item Parameter Estimates (cont.)

Item	Max Pts	a-par/alpha	b-par/ gamma1	c-par/ gamma2	gamma3
23	1	0.035	633.594	0.007	
24	1	0.020	667.288	0.296	
25	1	0.033	680.990	0.265	
26	1	0.021	671.809	0.078	
27	1	0.038	671.470	0.150	
28	1	0.033	689.238	0.187	
29	1	0.037	671.607	0.124	
30	1	0.033	686.498	0.410	
31	1	0.053	675.079	0.147	
32	1	0.039	672.468	0.079	
33	1	0.032	641.194	0.121	
34	1	0.031	635.185	0.334	
35	1	0.035	692.809	0.215	
36	1	0.033	692.072	0.147	
37	1	0.050	668.000	0.149	
38	1	0.022	713.821	0.178	
39	1	0.041	669.129	0.310	
40	1	0.029	689.046	0.286	
41	1	0.026	698.783	0.221	
42	1	0.038	667.327	0.056	
43	1	0.049	678.969	0.354	
44	1	0.024	637.232	0.131	
45	2	0.028	19.671	17.182	
46	2	0.032	20.332	20.456	
47	2	0.040	26.744	26.666	
48	2	0.035	23.839	25.359	
49	3	0.016	10.215	10.402	10.660
50	3	0.022	14.751	15.408	14.794
51	3	0.041	26.835	27.520	28.358

Table 24. Grade 6 2012 Operational Item Parameter Estimates

Item	Max Pts	a-par/ alpha	b-par/ gamma1	c-par/ gamma2	gamma3
01	1	0.055	652.405	0.268	
02	1	0.034	665.242	0.234	
03	1	0.053	672.547	0.272	
04	1	0.045	669.529	0.175	
05	1	0.039	725.907	0.271	
06	1	0.029	676.748	0.375	
07	1	0.040	676.611	0.104	
08	1	0.036	695.821	0.303	
09	1	0.032	712.376	0.294	
10	1	0.048	713.327	0.141	
11	1	0.039	661.987	0.336	
12	1	0.045	696.766	0.285	
13	1	0.049	684.524	0.213	
14	1	0.038	722.423	0.250	
15	1	0.039	661.944	0.014	
16	1	0.017	677.723	0.130	
17	1	0.050	705.724	0.205	
18	1	0.035	664.869	0.115	
19	1	0.037	685.935	0.164	
20	1	0.021	658.663	0.115	
21	1	0.031	674.643	0.190	
22	1	0.060	678.470	0.119	
23	1	0.022	639.500	0.062	
24	1	0.029	653.151	0.205	
25	1	0.020	595.036	0.037	
26	1	0.048	669.356	0.115	
27	1	0.043	684.512	0.289	
28	1	0.028	671.979	0.444	
29	1	0.048	647.420	0.004	

Table 24. Grade 6 2012 Operational Item Parameter Estimates (cont.)

Item	Max Pts	a-par/ alpha	b-par/ gamma1	c-par/ gamma2	gamma3
30	1	0.019	656.766	0.110	
31	1	0.048	672.627	0.273	
32	1	0.041	665.646	0.428	
33	1	0.037	685.775	0.072	
34	1	0.032	662.913	0.181	
35	1	0.047	700.156	0.223	
36	1	0.048	697.096	0.117	
37	1	0.031	655.676	0.127	
38	1	0.060	692.188	0.433	
39	1	0.039	665.658	0.320	
40	1	0.055	663.209	0.226	
41	1	0.043	679.955	0.444	
42	1	0.033	682.695	0.317	
43	1	0.023	647.316	0.319	
44	1	0.070	711.452	0.295	
45	1	0.045	665.391	0.261	
46	1	0.027	677.852	0.139	
47	1	0.040	660.922	0.188	
48	1	0.049	667.429	0.212	
49	2	0.031	21.416	18.999	
50	2	0.051	32.850	33.640	
51	2	0.056	38.066	36.450	
52	2	0.031	23.254	19.902	
53	2	0.028	16.650	18.051	
54	3	0.038	25.367	25.482	25.765
55	3	0.041	27.486	28.466	27.842
56	3	0.039	26.550	27.073	26.215
57	3	0.041	27.015	28.551	26.490

Table 25. Grade 7 2012 Operational Item Parameter Estimates

Item	Max Pts	a-par/ alpha	b-par/ gamma1	c-par/ gamma2	gamma3
01	1	0.029	652.473	0.386	
02	1	0.008	630.762	0.008	
03	1	0.026	644.119	0.032	
04	1	0.028	684.868	0.267	
05	1	0.052	686.740	0.148	
06	1	0.028	677.215	0.188	
07	1	0.024	669.687	0.200	
08	1	0.020	683.909	0.249	
09	1	0.023	646.477	0.028	
10	1	0.038	655.677	0.216	
11	1	0.044	679.312	0.221	
12	1	0.020	632.438	0.009	
13	1	0.053	720.119	0.307	
14	1	0.045	706.151	0.389	
15	1	0.054	690.517	0.173	
16	1	0.033	664.540	0.269	
17	1	0.023	656.332	0.046	
18	1	0.030	658.929	0.208	
19	1	0.038	674.950	0.148	
20	1	0.032	673.920	0.125	
21	1	0.043	683.971	0.375	
22	1	0.040	689.941	0.078	
23	1	0.056	643.633	0.321	
24	1	0.029	641.253	0.024	
25	1	0.023	633.377	0.238	
26	1	0.039	673.592	0.185	
27	1	0.041	695.356	0.117	
28	1	0.043	671.375	0.269	
29	1	0.030	642.810	0.214	

Table 25. Grade 7 2012 Operational Item Parameter Estimates (cont.)

Item	Max Pts	a-par/ alpha	b-par/ gamma1	c-par/ gamma2	gamma3
30	1	0.050	700.462	0.307	
31	1	0.049	674.498	0.153	
32	1	0.049	652.884	0.075	
33	1	0.032	687.298	0.086	
34	1	0.036	680.197	0.237	
35	1	0.032	681.525	0.161	
36	1	0.041	683.583	0.338	
37	1	0.038	688.410	0.383	
38	1	0.046	688.460	0.162	
39	1	0.041	692.119	0.264	
40	1	0.042	693.348	0.263	
41	1	0.027	621.285	0.008	
42	1	0.028	634.802	0.007	
43	1	0.043	671.847	0.296	
44	1	0.056	684.349	0.310	
45	1	0.041	661.767	0.111	
46	1	0.042	668.324	0.295	
47	2	0.055	37.628	39.747	
48	2	0.050	33.644	33.258	
49	2	0.031	20.439	19.124	
50	2	0.049	32.107	31.510	
51	2	0.045	31.149	29.778	
52	3	0.039	27.735	28.682	27.005
53	3	0.035	22.224	22.501	23.105
54	3	0.038	25.080	24.940	26.088
55	3	0.052	34.279	36.509	35.177

Table 26. Grade 8 2012 Operational Item Parameter Estimates

Item	Max Pts	a-par/ alpha	b-par/ gamma1	c-par/ gamma2	gamma3
01	1	0.029	632.344	0.210	
02	1	0.035	630.955	0.009	
03	1	0.060	669.742	0.176	
04	1	0.043	669.642	0.355	
05	1	0.053	676.879	0.265	
06	1	0.044	712.906	0.261	
07	1	0.047	701.976	0.133	
08	1	0.020	666.748	0.065	
09	1	0.037	692.079	0.168	
10	1	0.032	646.304	0.326	
11	1	0.031	641.417	0.012	
12	1	0.026	627.884	0.010	
13	1	0.031	674.794	0.115	
14	1	0.046	689.378	0.213	
15	1	0.019	646.291	0.022	
16	1	0.057	681.760	0.236	
17	1	0.056	698.726	0.124	
18	1	0.035	637.326	0.010	
19	1	0.040	649.576	0.117	
20	1	0.061	673.456	0.187	
21	1	0.039	719.026	0.124	
22	1	0.033	637.309	0.171	
23	1	0.042	642.732	0.033	
24	1	0.042	663.959	0.115	
25	1	0.019	651.053	0.025	
26	1	0.043	686.973	0.181	
27	1	0.079	710.281	0.279	
28	1	0.057	670.554	0.166	
29	1	0.048	696.017	0.243	

Table 26. Grade 8 2012 Operational Item Parameter Estimates (cont.)

Item	Max Pts	a-par/ alpha	b-par/ gamma1	c-par/ gamma2	gamma3
30	1	0.030	631.527	0.126	
31	1	0.034	691.670	0.244	
32	1	0.054	677.533	0.168	
33	1	0.047	682.257	0.135	
34	1	0.041	665.725	0.195	
35	1	0.044	639.378	0.088	
36	1	0.056	682.280	0.111	
37	1	0.027	620.787	0.036	
38	1	0.060	672.638	0.221	
39	1	0.051	671.049	0.333	
40	1	0.034	690.109	0.371	
41	1	0.029	639.295	0.143	
42	1	0.040	659.733	0.214	
43	1	0.061	668.856	0.217	
44	1	0.055	660.859	0.145	
45	1	0.024	643.923	0.065	
46	1	0.038	680.694	0.223	
47	2	0.052	35.336	35.028	
48	2	0.039	26.427	24.669	
49	2	0.069	47.179	48.077	
50	2	0.040	28.332	27.184	
51	2	0.065	44.566	43.740	
52	3	0.041	29.079	28.918	28.389
53	3	0.046	32.555	32.209	32.521
54	3	0.039	24.460	26.285	27.607
55	3	0.052	35.758	34.751	35.061

Test Characteristic Curves

Test characteristic curves (TCCs) provide an overview of the tests in the IRT scale score metric. The 2011 and 2012 TCCs were generated using final OP item parameters for all reporting test items administered in 2011 and 2012. TCCs are the summation of all the item characteristic curves (ICCs) for items that contribute to the OP scale score. Conditional Standard Error of Measurement (CSEM) curves graphically show the amount of measurement error at different ability levels. The 2011 and 2012 TCCs curves and CSEM curves are presented in Figure 1 through Figure 12. Following the adoption of the chain equating method by New York State, the TCCs for new OP test forms are compared to the previous year's TCCs rather than to the baseline 2006 test form TCCs. It should be noted that the test lengths between 2011 and 2012 operational tests are slightly different. Note that in all figures red represents the 2012 OP test and green represents the 2011 OP test. The *x*-axis is the ability scale expressed in scale score metric with the lower and upper bounds established in Year 1 of test administration and presented in the lower corners of the graphs. The *y*-axis is the proportion of the test that the students can answer correctly.

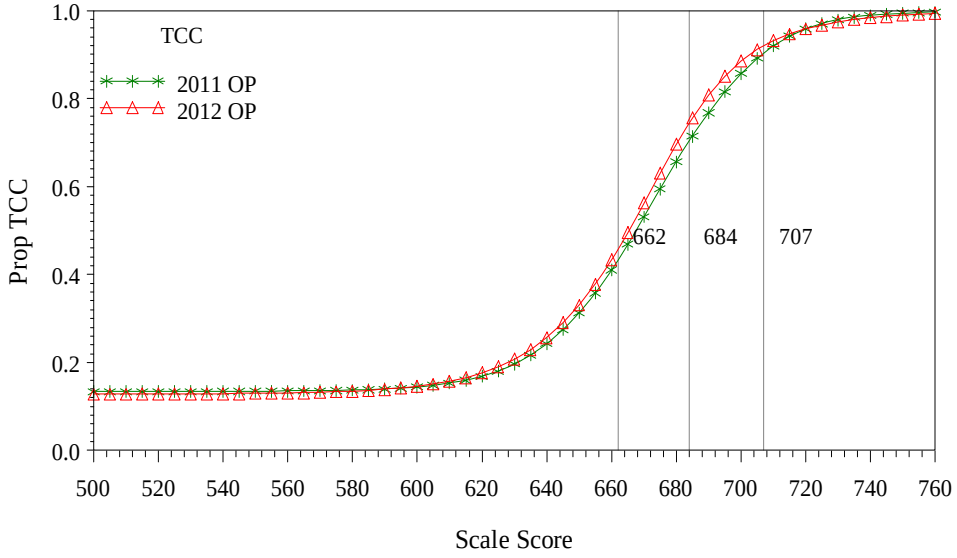

Figure 1. Grade 3 2011 and 2012 OP TCCs

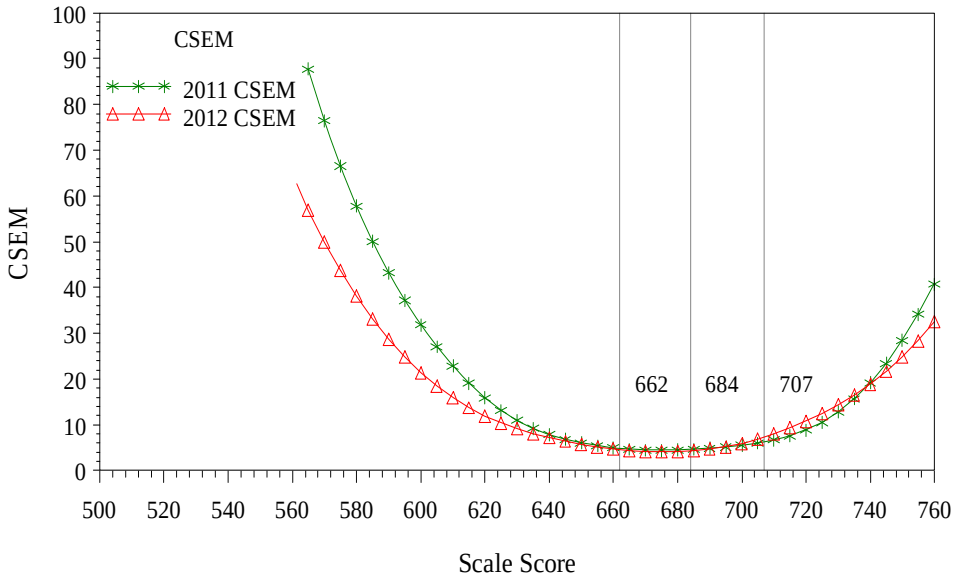

Figure 2. Grade 3 2011 and 2012 CSEM Curves

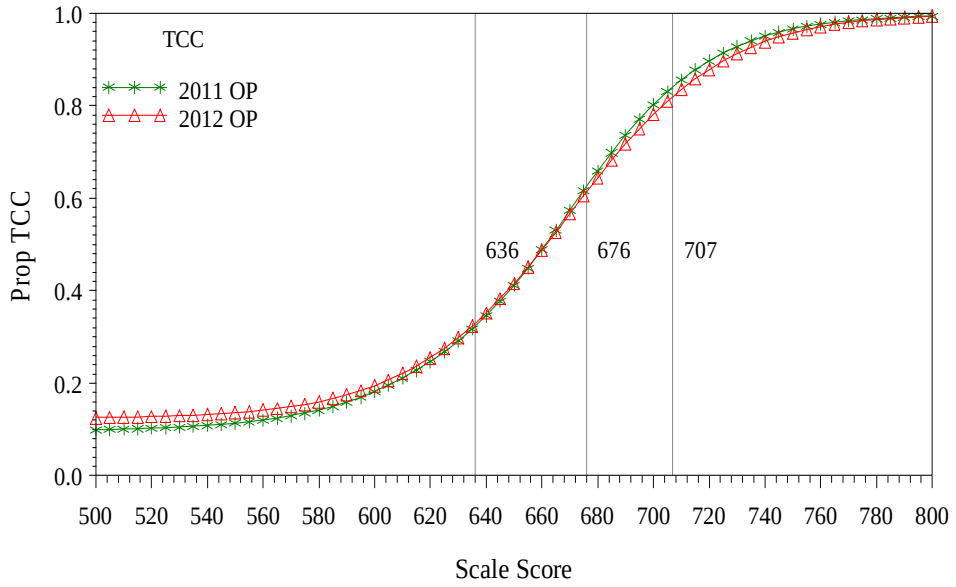

Figure 3. Grade 4 2011 and 2012 OP TCCs

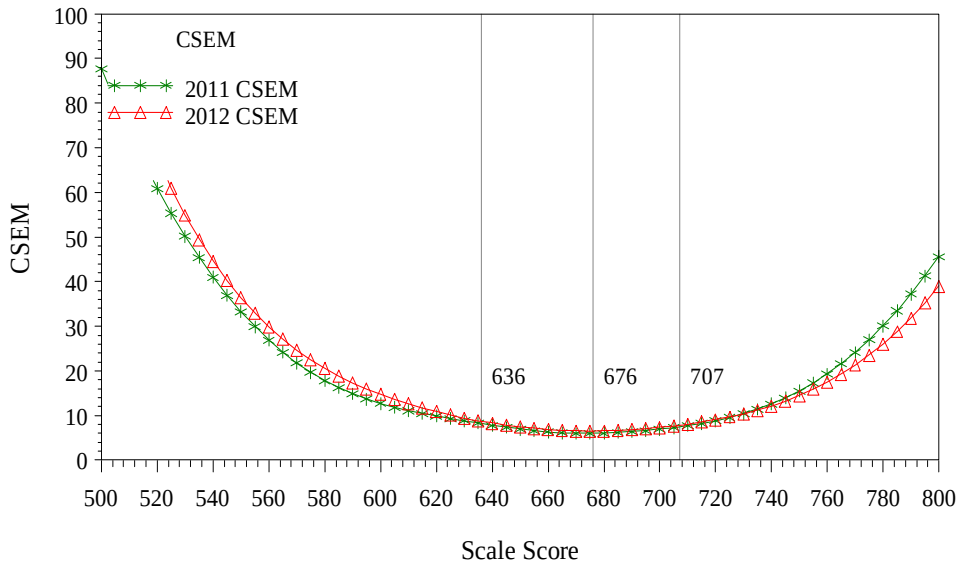

Figure 4. Grade 4 2011 and 2012 CSEM Curves

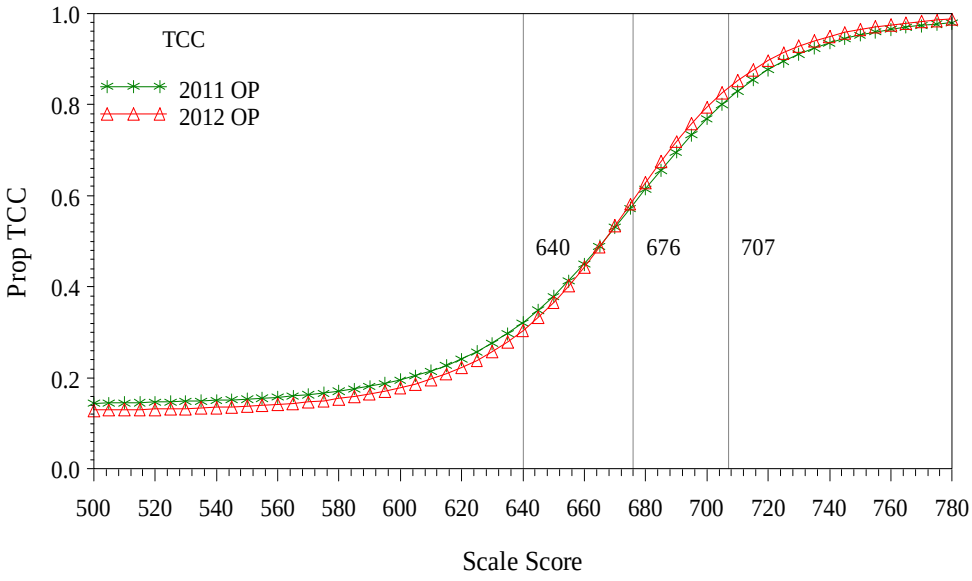

Figure 5. Grade 5 2011 and 2012 OP TCCs

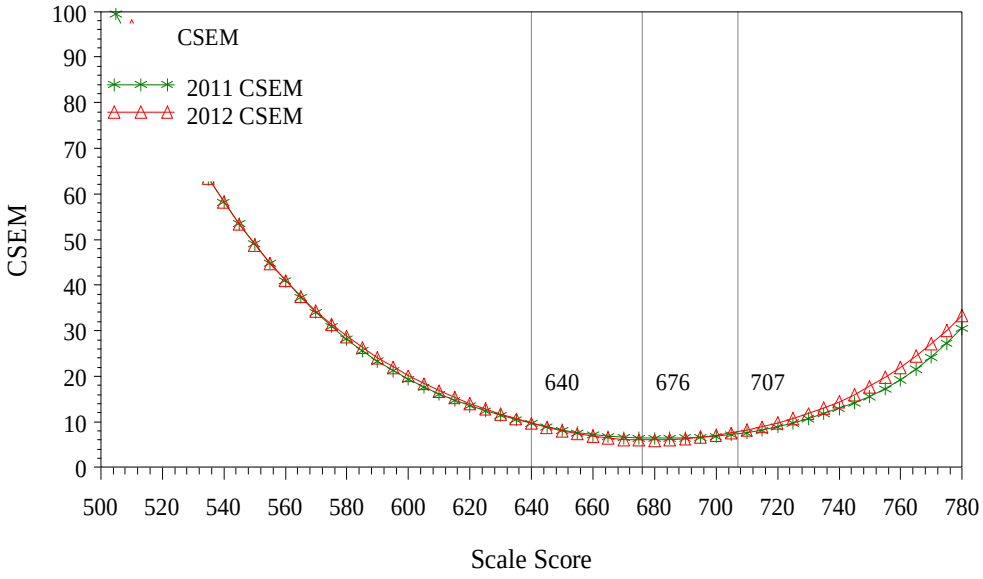

Figure 6. Grade 5 2011 and 2012 CSEM Curves

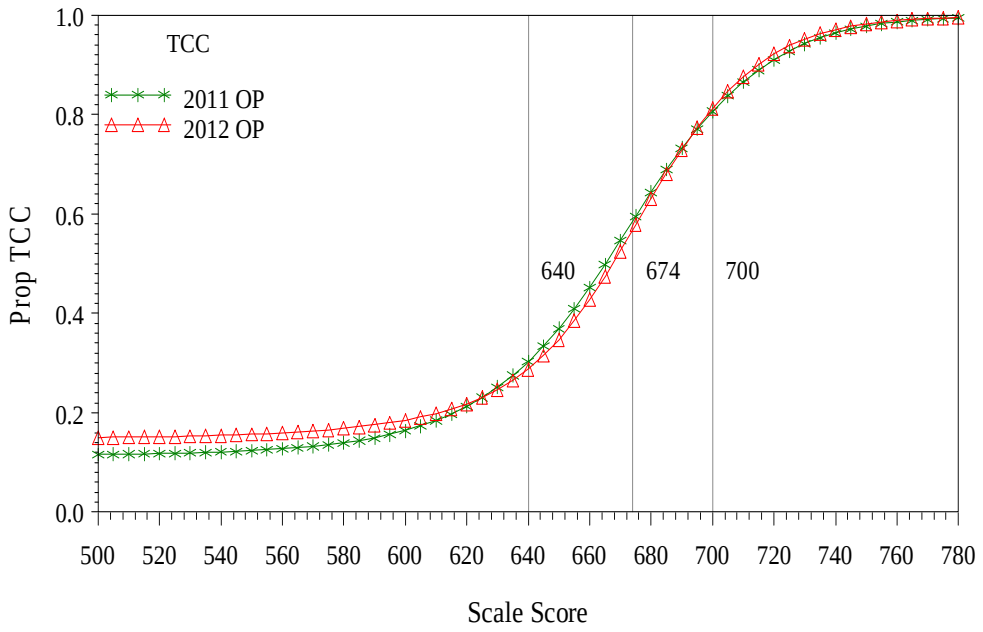

Figure 7. Grade 6 2011 and 2012 OP TCCs

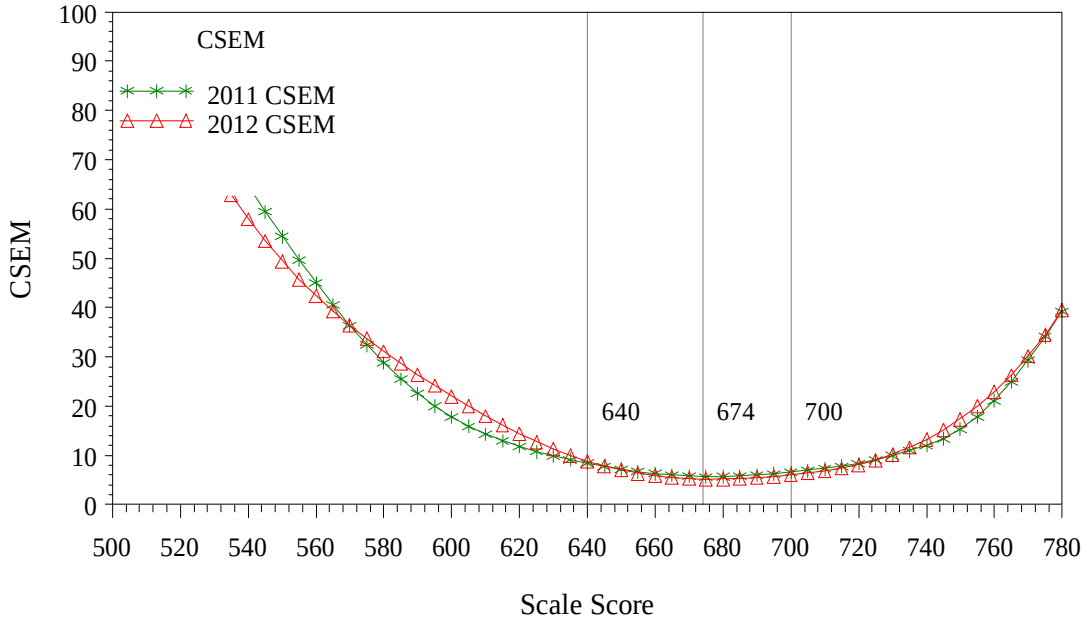

Figure 8. Grade 6 2011 and 2012 CSEM Curves

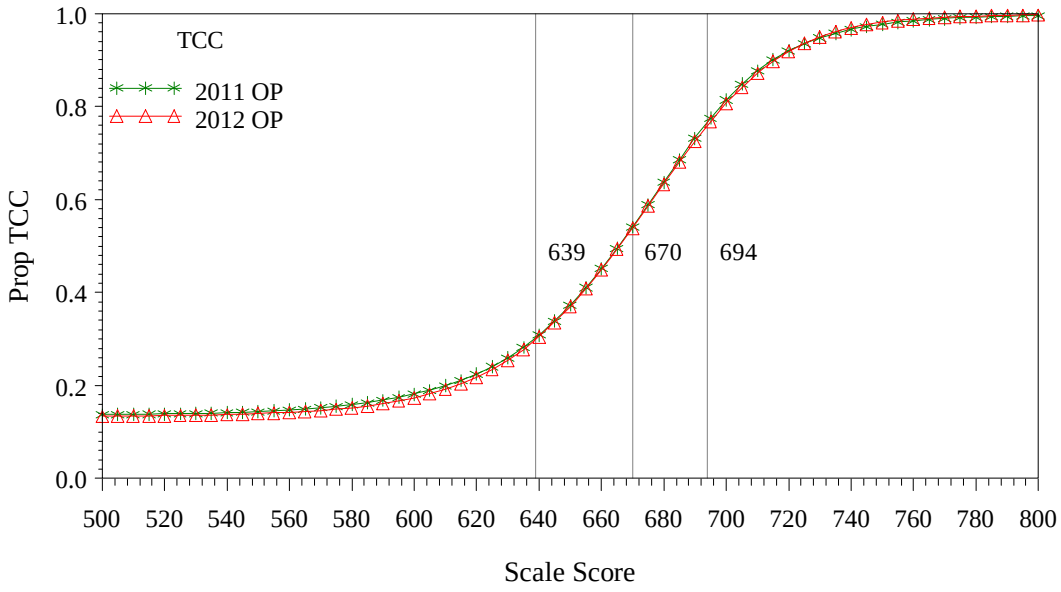

Figure 9. Grade 7 2011 and 2012 OP TCCs

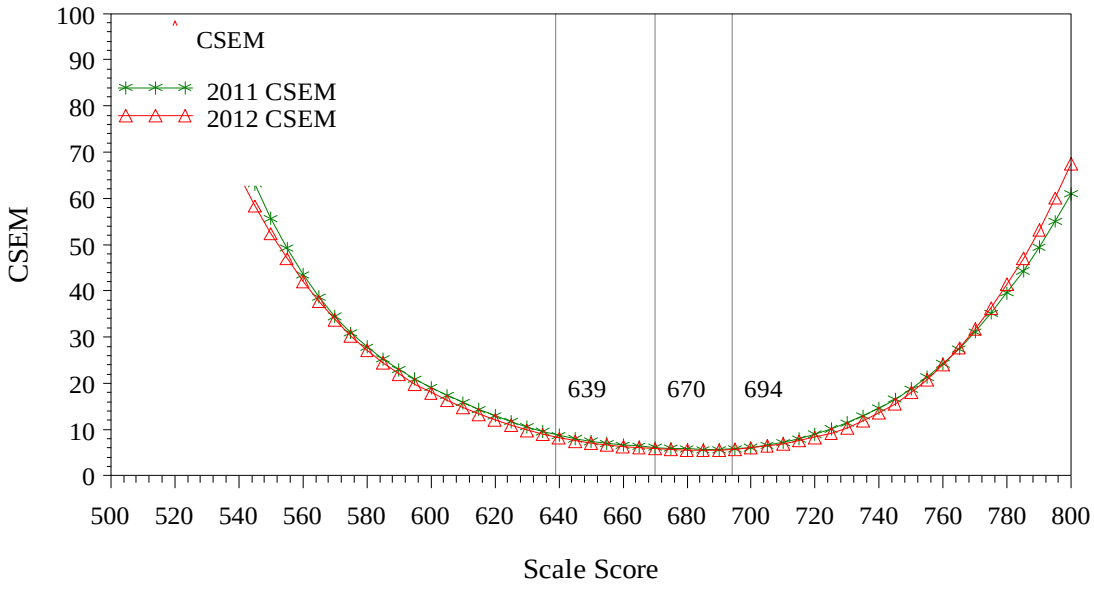

Figure 10. Grade 7 2011 and 2012 CSEM Curves

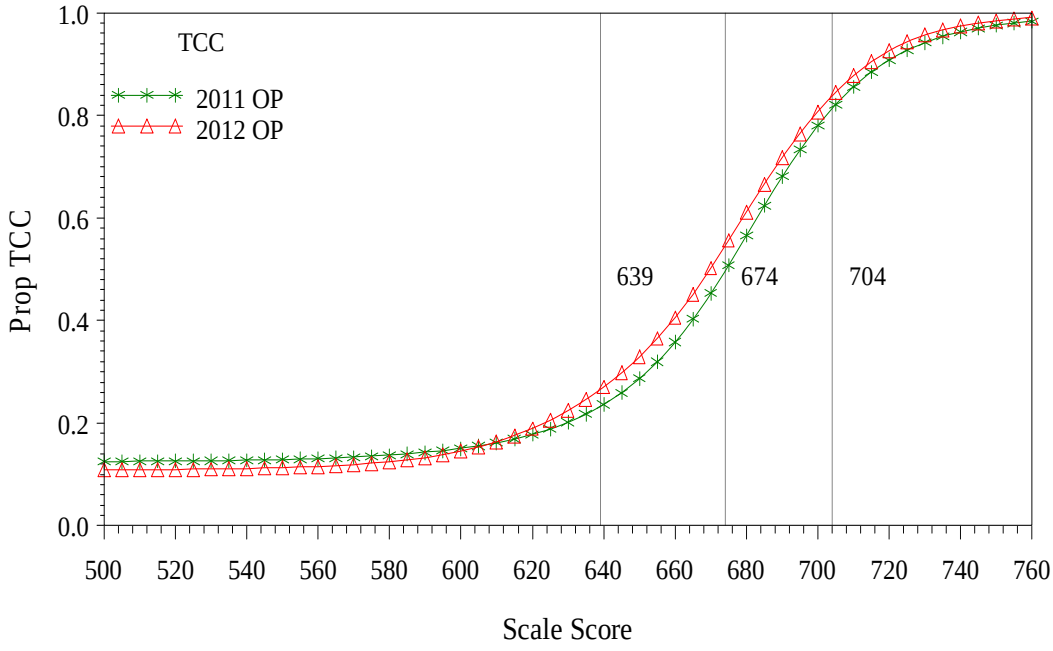

Figure 11. Grade 8 2011 and 2012 OP TCCs

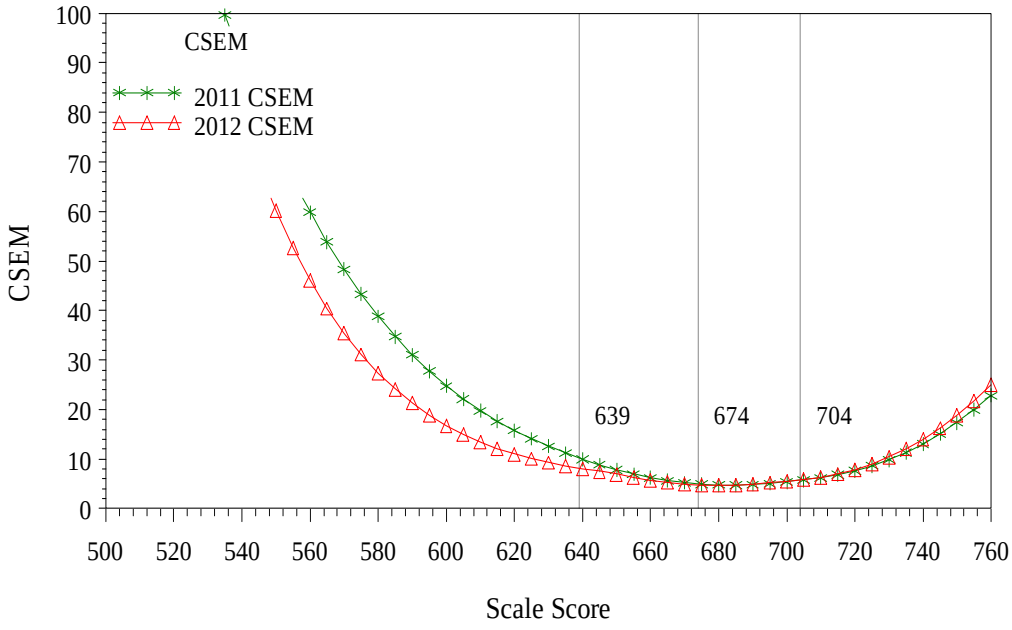

Figure 12. Grade 8 2011 and 2012 CSEM Curves

As seen in Figures 1–12, the 2012 TCCs for all grades were found to be very similar to the 2011 TCCs, indicating that the 2012 form had the same difficulty to the 2011 forms for most of the students. The CSEM curves were well aligned for all grades. It should be noted that potential differences in test form difficulty at different ability levels are accounted for in the equating and in the resulting raw score-to-scale score conversion tables, so that students of the same ability are expected to obtain the same scale score regardless of which form they took.

Scoring Procedure

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

The final item parameters in the scale score metric were used to produce raw score-to-scale score conversion tables for the Grades 3–8 Math Tests. An inverse TCC method was employed using POLYEQUATE (Kolen, 2003). The inverse of the TCC procedure produces trait values based on unweighted raw scores. These estimates show negligible bias for tests with maximum possible raw scores of at least 30 points. All New York State Mathematics Tests have a maximum raw score higher than 30 points. In the inverse TCC method, a student's trait estimate is taken to be the trait value that has an expected raw score equal to the student's observed raw score. It was found that for tests containing all MC items, the inverse of the TCC is an excellent first-order

approximation to the number correct maximum likelihood estimates (MLE) showing negligible bias for tests of at least 30 items. For tests with a mixture of MC and CR items, the MLE and TCC estimates are even more similar (Yen, 1984).

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

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

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:

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

where

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

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

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

Table 27. Grade 3 Raw Score-to-Scale Score (with Standard Error)

Weighted Raw Score	Scale Score	Standard Error
0	470	412
1	470	412
2	470	412
3	470	412
4	470	412
5	470	412
6	470	412
7	470	412
8	574	45
9	604	19
10	615	14
11	623	11
12	628	10
13	632	9
14	636	8
15	639	7
16	642	7
17	644	7
18	646	6
19	648	6
20	650	6
21	652	6
22	654	5
23	655	5
24	657	5
25	658	5
26	660	5
27	661	5
28	663	5

Table 27. Grade 3 Raw Score-to-Scale Score (with Standard Error) (cont.)

Weighted Raw Score	Scale Score	Standard Error
29	664	4
30	665	4
31	667	4
32	668	4
33	669	4
34	670	4
35	672	4
36	673	4
37	674	4
38	675	4
39	676	4
40	678	4
41	679	4
42	680	4
43	682	4
44	683	4
45	684	4
46	686	4
47	688	5
48	689	5
49	691	5
50	693	5
51	695	5
52	697	5
53	700	6
54	703	6
55	706	7
56	711	8
57	716	10

Table 27. Grade 3 Raw Score-to-Scale Score (with Standard Error) (cont.)

Weighted Raw Score	Scale Score	Standard Error
58	725	13
59	740	19
60	770	43

Table 28. Grade 4 Raw Score-to-Scale Score (with Standard Error)

Weighted Raw Score	Scale Score	Standard Error
0	485	139
1	485	139
2	485	139
3	485	139
4	485	139
5	485	139
6	485	139
7	485	139
8	485	139
9	533	52
10	565	27
11	579	21
12	589	18
13	597	16
14	603	14
15	609	13
16	614	12
17	618	11
18	622	11
19	625	10
20	628	10
21	631	9
22	634	9

Table 28. Grade 4 Raw Score-to-Scale Score (with Standard Error) (cont.)

Weighted Raw Score	Scale Score	Standard Error
23	637	9
24	639	8
25	642	8
26	644	8
27	646	8
28	649	8
29	651	7
30	653	7
31	655	7
32	657	7
33	659	7
34	661	7
35	662	7
36	664	7
37	666	7
38	668	7
39	670	7
40	672	7
41	674	7
42	675	7
43	677	7
44	679	7
45	681	7
46	683	7
47	685	7
48	687	7
49	689	7
50	691	7
51	693	7

Table 28. Grade 4 Raw Score-to-Scale Score (with Standard Error) (cont.)

Weighted Raw Score	Scale Score	Standard Error
52	695	7
53	698	7
54	700	7
55	703	8
56	705	8
57	708	8
58	711	8
59	714	8
60	718	9
61	722	9
62	726	10
63	730	10
64	735	11
65	742	13
66	749	14
67	760	18
68	779	26
69	800	39

Table 29. Grade 5 Raw Score-to-Scale Score (with Standard Error)

Weighted Raw Score	Scale Score	Standard Error
0	495	123
1	495	123
2	495	123
3	495	123
4	495	123
5	495	123
6	495	123
7	495	123

Table 29. Grade 5 Raw Score-to-Scale Score (with Standard Error) (cont.)

Weighted Raw Score	Scale Score	Standard Error
8	516	88
9	570	34
10	589	25
11	601	20
12	610	17
13	616	15
14	622	14
15	627	12
16	631	11
17	635	11
18	638	10
19	641	10
20	644	9
21	647	8
22	649	8
23	652	8
24	654	8
25	656	7
26	658	7
27	660	7
28	662	7
29	664	7
30	665	6
31	667	6
32	669	6
33	671	6
34	672	6
35	674	6
36	676	6

Table 29. Grade 5 Raw Score-to-Scale Score (with Standard Error) (cont.)

Weighted Raw Score	Scale Score	Standard Error
37	678	6
38	679	6
39	681	6
40	683	6
41	685	6
42	686	6
43	688	6
44	690	6
45	692	6
46	694	7
47	697	7
48	699	7
49	701	7
50	704	7
51	707	8
52	710	8
53	713	9
54	717	9
55	721	10
56	726	11
57	733	13
58	741	15
59	752	19
60	772	28
61	780	33

Table 30. Grade 6 Raw Score-to-Scale Score (with Standard Error)

Weighted Raw Score	Scale Score	Standard Error
0	500	114
1	500	114
2	500	114
3	500	114
4	500	114
5	500	114
6	500	114
7	500	114
8	500	114
9	500	114
10	500	114
11	551	49
12	584	29
13	600	22
14	611	18
15	618	15
16	624	13
17	629	12
18	633	10
19	636	10
20	640	9
21	642	8
22	645	8
23	647	8
24	649	7
25	651	7
26	653	7
27	655	6
28	657	6

Table 30. Grade 6 Raw Score-to-Scale Score (with Standard Error) (cont.)

Weighted Raw Score	Scale Score	Standard Error
29	658	6
30	660	6
31	662	6
32	663	6
33	665	6
34	666	5
35	668	5
36	669	5
37	670	5
38	672	5
39	673	5
40	674	5
41	676	5
42	677	5
43	678	5
44	680	5
45	681	5
46	683	5
47	684	5
48	685	5
49	687	5
50	688	5
51	690	6
52	691	6
53	693	6
54	695	6
55	696	6
56	698	6
57	700	6

Table 30. Grade 6 Raw Score-to-Scale Score (with Standard Error) (cont.)

Weighted Raw Score	Scale Score	Standard Error
58	702	6
59	704	6
60	707	7
61	709	7
62	712	7
63	715	7
64	718	8
65	722	8
66	726	9
67	732	11
68	741	14
69	755	20
70	780	39

Table 31. Grade 7 Raw Score-to-Scale Score (with Standard Error)

Weighted Raw Score	Scale Score	Standard Error
0	500	137
1	500	137
2	500	137
3	500	137
4	500	137
5	500	137
6	500	137
7	500	137
8	500	137
9	500	137
10	571	33
11	591	22
12	602	17

Table 31. Grade 7 Raw Score-to-Scale Score (with Standard Error) (cont.)

Weighted Raw Score	Scale Score	Standard Error
13	610	15
14	616	13
15	621	12
16	625	11
17	629	10
18	632	9
19	635	9
20	638	8
21	641	8
22	643	8
23	645	8
24	647	7
25	650	7
26	652	7
27	654	7
28	655	7
29	657	7
30	659	6
31	661	6
32	662	6
33	664	6
34	666	6
35	667	6
36	669	6
37	671	6
38	672	6
39	674	6
40	675	6
41	677	6

Table 31. Grade 7 Raw Score-to-Scale Score (with Standard Error) (cont.)

Weighted Raw Score	Scale Score	Standard Error
42	678	6
43	680	6
44	681	6
45	683	6
46	685	6
47	686	6
48	688	6
49	689	6
50	691	6
51	693	6
52	695	6
53	696	6
54	698	6
55	700	6
56	702	6
57	704	6
58	707	7
59	709	7
60	712	7
61	715	8
62	718	8
63	722	9
64	727	10
65	733	11
66	741	14
67	756	21
68	800	68

Table 32. Grade 8 Raw Score-to-Scale Score (with Standard Error)

Weighted Raw Score	Scale Score	Standard Error
0	480	356
1	480	356
2	480	356
3	480	356
4	480	356
5	480	356
6	480	356
7	480	356
8	566	39
9	589	22
10	601	16
11	609	14
12	615	12
13	620	11
14	625	10
15	629	9
16	633	9
17	636	9
18	639	8
19	642	8
20	644	8
21	647	7
22	649	7
23	651	7
24	653	7
25	655	6
26	657	6
27	659	6
28	661	6

Table 32. Grade 8 Raw Score-to-Scale Score (with Standard Error) (cont.)

Weighted Raw Score	Scale Score	Standard Error
29	662	6
30	664	5
31	665	5
32	667	5
33	668	5
34	670	5
35	671	5
36	673	5
37	674	5
38	675	5
39	677	5
40	678	5
41	679	5
42	681	5
43	682	5
44	683	5
45	685	5
46	686	5
47	687	5
48	689	5
49	690	5
50	692	5
51	693	5
52	695	5
53	697	5
54	698	5
55	700	6
56	702	6
57	704	6

Table 32. Grade 8 Raw Score-to-Scale Score (with Standard Error) (cont.)

Weighted Raw Score	Scale Score	Standard Error
58	706	6
59	708	6
60	711	6
61	713	7
62	716	7
63	720	8
64	724	9
65	730	10
66	737	13
67	751	19
68	775	38

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, Pearson'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 Pearson 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 2012 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 33 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 33. SPI Target Ranges

Grade	Objective	# Items	Total Points	Level III Cut SPI Target Range
3	1	24	29	75–84
	2	7	9	63–75
	3	6	7	54–59
	4	7	7	66–77
	5	6	8	75–84
4	1	25	31	64–75
	2	7	9	44–55
	3	5	7	52–62
	4	11	12	44–56
	5	8	10	58–67
5	1	21	25	51–64
	2	6	8	54–65
	3	12	13	57–69
	4	8	9	56–65
	5	4	6	49–58
6	1	22	25	54–65
	2	11	14	50–60
	3	9	13	45–54
	4	6	8	36–49
	5	9	10	69–76
7	1	18	21	57–69
	2	6	7	40–49
	3	8	11	40–50
	4	7	8	62–72
	5	16	21	42–54
8	1	4	7	39–46
	2	24	29	46–60
	3	22	24	66–75
	4	5	8	19–27

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 New York State 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.

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 Pearson and is included in a different 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 34 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 34. 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	199,587	50	60	0.92	3.00	0.93	2.87
4	195,956	56	69	0.93	3.45	0.94	3.27
5	198,201	51	61	0.93	3.42	0.94	3.29
6	200,802	57	70	0.94	3.83	0.95	3.56
7	199,113	55	68	0.93	3.68	0.94	3.52
8	199,783	55	68	0.94	3.69	0.95	3.44

All the coefficients for total test reliability were in the range 0.92–0.95, 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, 6, 7, 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 35 presents reliabilities for the MC subsets.

Table 35. 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	199,587	43	0.90	2.39	0.90	2.38
4	195,956	47	0.91	2.50	0.91	2.49
5	198,201	44	0.92	2.62	0.92	2.60
6	200,802	48	0.92	2.77	0.92	2.75
7	199,113	46	0.91	2.79	0.91	2.78
8	199,783	46	0.92	2.64	0.92	2.62

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 7–9 CR items depending on grade level. The results are presented in Table 36.

Table 36. 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	199,587	7	17	0.77	1.68	0.80	1.59
4	195,956	9	22	0.82	2.18	0.83	2.10
5	198,201	7	17	0.76	2.05	0.77	2.02
6	200,802	9	22	0.86	2.33	0.88	2.21
7	199,113	9	22	0.84	2.17	0.85	2.14
8	199,783	9	22	0.87	2.25	0.88	2.19

Note: Results should be interpreted with caution for Grades 3 and 5 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 Capacity Category (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 Languages. 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 0.86. Feldt-Raju reliability coefficients, which tend to be larger than the Cronbach's alpha estimates for the same group, were all larger than 0.88. Overall, the New York State Mathematics Tests were found to have very good test internal consistency (reliability) for analyzed subgroups of examinees.

Table 37A. 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	199,587	0.92	3.00	0.93	2.87
Gender	Female	97,664	0.92	3.00	0.92	2.88
	Male	101,923	0.92	2.99	0.93	2.87
Ethnicity	Asian	16,917	0.90	2.59	0.91	2.49
	Black	35,827	0.92	3.26	0.93	3.13
	Hispanic	47,940	0.92	3.19	0.93	3.05
	American Indian	1,086	0.91	3.11	0.93	3.01
	Multiracial	2,018	0.92	3.02	0.93	2.89
	Other	416	0.92	2.91	0.93	2.76
	White	95,383	0.91	2.83	0.91	2.73

Table 37A. Grade 3 Test Reliability by Subgroup (cont.)

Group	Subgroup	N-count	Cronbach's Alpha	SEM of Cronbach	Feldt-Raju	SEM of Feldt-Raju
NRC	New York City	72,824	0.92	3.07	0.93	2.93
	Big 4 Cites	8,128	0.93	3.37	0.93	3.22
	High Needs Urban/Suburban	15,732	0.92	3.20	0.93	3.07
	High Needs Rural	11,031	0.91	3.11	0.92	3.00
	Average Needs	57,884	0.91	2.91	0.91	2.80
	Low Needs	28,021	0.89	2.65	0.90	2.56
	Charter	5,967	0.89	2.95	0.90	2.85
SWD	All Codes	28,242	0.92	3.41	0.93	3.25
SUA	All Codes	23,078	0.92	3.36	0.93	3.22
SWD/ SUA	SUA=504 Plan Codes	12,368	0.92	3.45	0.93	3.30
ELL/ SUA	SUA=ELL Codes	5,943	0.92	3.37	0.92	3.24
ELL	English	15,493	0.92	3.36	0.93	3.22
	Chinese	396	0.88	2.87	0.90	2.79
	Haitian-Creole	40	0.92	3.60	0.93	3.41
	Korean	26	0.92	2.71	0.93	2.50
	Russian	63	0.94	3.18	0.95	3.00
	Spanish	2,646	0.91	3.49	0.92	3.35
	All Translations	3,171	0.92	3.43	0.93	3.29

Table 37B. 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	195,956	0.93	3.45	0.94	3.27
Gender	Female	96,155	0.93	3.47	0.93	3.28
	Male	99,801	0.93	3.44	0.94	3.25

Table 37B. Grade 4 Test Reliability by Subgroup (cont.)

Group	Subgroup	N-count	Cronbach's Alpha	SEM of Cronbach	Feldt-Raju	SEM of Feldt-Raju
Ethnicity	Asian	16,596	0.92	3.05	0.93	2.87
	Black	35,613	0.93	3.62	0.93	3.46
	Hispanic	46,302	0.92	3.57	0.93	3.40
	American Indian	1,042	0.93	3.56	0.94	3.37
	Multiracial	1,701	0.93	3.50	0.94	3.31
	Other	326	0.93	3.30	0.94	3.14
	White	94,376	0.92	3.35	0.93	3.17
NRC	New York City	70,753	0.93	3.49	0.94	3.29
	Big 4 Cites	8,111	0.93	3.68	0.94	3.52
	High Needs Urban/Suburban	15,307	0.93	3.59	0.94	3.42
	High Needs Rural	11,048	0.92	3.55	0.93	3.38
	Average Needs	57,665	0.92	3.41	0.93	3.24
	Low Needs	28,286	0.90	3.19	0.91	3.02
	Charter	4,786	0.91	3.37	0.92	3.22
SWD	All Codes	29,617	0.93	3.72	0.94	3.55
SUA	All Codes	20,724	0.93	3.70	0.94	3.53
SWD/ SUA	SUA=504 Plan Codes	12,666	0.93	3.73	0.93	3.57
ELL/ SUA	SUA=ELL Codes	3,798	0.93	3.70	0.93	3.54
ELL	English	14,168	0.93	3.68	0.93	3.52
	Chinese	468	0.90	3.17	0.91	3.04
	Haitian-Creole	56	0.93	3.71	0.93	3.49
	Korean	32	0.86	2.65	0.88	2.41
	Russian	66	0.93	3.57	0.94	3.33
	Spanish	2,705	0.92	3.70	0.93	3.55
	All Translations	3,327	0.94	3.69	0.94	3.49

Table 37C. 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	198,201	0.93	3.42	0.94	3.29
Gender	Female	97,203	0.93	3.42	0.93	3.30
	Male	100,998	0.93	3.41	0.94	3.28
Ethnicity	Asian	16,202	0.93	2.93	0.94	2.83
	Black	36,739	0.92	3.60	0.93	3.47
	Hispanic	46,034	0.93	3.54	0.93	3.42
	American Indian	996	0.93	3.55	0.94	3.42
	Multiracial	1,548	0.93	3.45	0.94	3.33
	Other	328	0.93	3.35	0.94	3.22
	White	96,354	0.92	3.33	0.93	3.22
NRC	New York City	69,660	0.93	3.40	0.94	3.28
	Big 4 Cites	8,083	0.93	3.65	0.93	3.51
	High Needs Urban/Suburban	15,122	0.92	3.59	0.93	3.46
	High Needs Rural	11,402	0.92	3.58	0.92	3.47
	Average Needs	59,175	0.92	3.41	0.93	3.30
	Low Needs	28,745	0.92	3.13	0.92	3.03
	Charter	6,014	0.91	3.47	0.92	3.37
SWD	All Codes	30,729	0.92	3.67	0.93	3.54
SUA	All Codes	26,448	0.92	3.66	0.93	3.52
SWD/ SUA	SUA=504 Plan Codes	16,219	0.92	3.67	0.92	3.54
ELL/ SUA	SUA=ELL Codes	4,925	0.92	3.66	0.93	3.53
ELL	English	12,122	0.92	3.67	0.93	3.54
	Chinese	416	0.91	3.16	0.92	3.06
	Haitian-Creole	48	0.92	3.51	0.93	3.34
	Korean	36	0.90	2.63	0.93	2.48
	Russian	71	0.93	3.58	0.94	3.42
	Spanish	2,375	0.91	3.68	0.91	3.56
	All Translations	2,946	0.93	3.63	0.94	3.49

Table 37D. 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	200,802	0.94	3.83	0.95	3.56
Gender	Female	98,129	0.94	3.82	0.95	3.54
	Male	102,673	0.94	3.83	0.95	3.56
Ethnicity	Asian	17,022	0.94	3.27	0.95	3.05
	Black	37,247	0.93	3.99	0.94	3.73
	Hispanic	45,281	0.93	3.98	0.94	3.71
	American Indian	1,016	0.94	3.96	0.95	3.69
	Multiracial	1,449	0.94	3.85	0.95	3.57
	Other	366	0.95	3.58	0.96	3.28
	White	98,421	0.93	3.71	0.94	3.48
NRC	New York City	70,043	0.95	3.88	0.95	3.57
	Big 4 Cites	7,877	0.93	3.96	0.94	3.72
	High Needs Urban/Suburban	14,781	0.93	3.99	0.94	3.73
	High Needs Rural	11,459	0.93	3.96	0.94	3.72
	Average Needs	61,005	0.93	3.80	0.94	3.56
	Low Needs	30,106	0.93	3.44	0.94	3.25
	Charter	5,531	0.93	3.79	0.94	3.55
SWD	All Codes	30,408	0.92	3.98	0.93	3.76
SUA	All Codes	19,942	0.93	3.99	0.94	3.76
SWD/ SUA	SUA=504 Plan Codes	13,186	0.92	3.96	0.93	3.76
ELL/ SUA	SUA=ELL Codes	2,506	0.93	3.97	0.94	3.74
ELL	English	9,675	0.93	4.00	0.93	3.77
	Chinese	407	0.93	3.72	0.94	3.51
	Haitian-Creole	113	0.91	3.97	0.92	3.74
	Korean	35	0.92	3.24	0.96	3.06
	Russian	92	0.96	3.91	0.97	3.58
	Spanish	2,382	0.91	3.92	0.92	3.71
	All Translations	3,029	0.94	3.96	0.94	3.70

Table 37E. 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	199,113	0.93	3.68	0.94	3.52
Gender	Female	97,037	0.93	3.67	0.94	3.50
	Male	102,076	0.93	3.68	0.94	3.53
Ethnicity	Asian	15,940	0.93	3.31	0.94	3.12
	Black	37,258	0.92	3.81	0.93	3.67
	Hispanic	44,233	0.92	3.79	0.93	3.65
	American Indian	1,019	0.92	3.78	0.93	3.65
	Multiracial	1,390	0.93	3.69	0.94	3.53
	Other	333	0.93	3.71	0.94	3.51
	White	98,940	0.92	3.57	0.93	3.43
NRC	New York City	68,533	0.94	3.74	0.94	3.55
	Big 4 Cites	7,708	0.91	3.84	0.92	3.70
	High Needs Urban/Suburban	14,623	0.92	3.79	0.93	3.66
	High Needs Rural	11,715	0.91	3.75	0.92	3.63
	Average Needs	60,217	0.92	3.61	0.93	3.48
	Low Needs	31,640	0.92	3.41	0.92	3.27
	Charter	4,677	0.92	3.62	0.92	3.49
SWD	All Codes	29,828	0.91	3.85	0.91	3.71
SUA	All Codes	20,298	0.91	3.85	0.92	3.71
SWD/SUA	SUA=504 Plan Codes	13,789	0.90	3.85	0.91	3.71
ELL/SUA	SUA=ELL Codes	2,606	0.92	3.88	0.92	3.72
ELL	English	8994	0.91	3.85	0.92	3.71
	Chinese	515	0.92	3.64	0.93	3.49
	Haitian-Creole	113	0.90	3.81	0.91	3.63
	Korean	32	0.94	3.41	0.95	3.20
	Russian	77	0.93	3.83	0.94	3.62
	Spanish	2,541	0.88	3.83	0.89	3.71
	All Translations	3,278	0.92	3.88	0.93	3.71

Table 37F. 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	199,783	0.94	3.69	0.95	3.44
Gender	Female	98,226	0.94	3.69	0.95	3.43
	Male	101,557	0.94	3.69	0.95	3.44
Ethnicity	Asian	16,226	0.94	3.34	0.95	3.08
	Black	37,062	0.93	3.72	0.94	3.53
	Hispanic	43,903	0.93	3.73	0.94	3.52
	American Indian	1,045	0.93	3.72	0.94	3.49
	Multiracial	1,146	0.94	3.69	0.95	3.42
	Other	354	0.95	3.67	0.96	3.37
	White	100,047	0.93	3.64	0.94	3.40
NRC	New York City	69,908	0.94	3.70	0.95	3.45
	Big 4 Cites	7,415	0.93	3.65	0.93	3.49
	High Needs Urban/Suburban	14,365	0.93	3.72	0.93	3.54
	High Needs Rural	11,619	0.92	3.73	0.93	3.54
	Average Needs	60,856	0.93	3.67	0.94	3.44
	Low Needs	32,185	0.92	3.48	0.94	3.25
	Charter	3,435	0.93	3.62	0.94	3.40
SWD	All Codes	29,633	0.92	3.64	0.92	3.50
SUA	All Codes	18,040	0.93	3.68	0.93	3.52
SWD/SUA	SUA=504 Plan Codes	12,807	0.92	3.64	0.92	3.50
ELL/SUA	SUA=ELL Codes	1,740	0.93	3.63	0.94	3.48
ELL	English	8,904	0.93	3.66	0.94	3.50
	Chinese	664	0.92	3.33	0.94	3.16
	Haitian-Creole	101	0.91	3.65	0.92	3.50
	Korean	41	0.87	2.77	0.88	2.60
	Russian	110	0.92	3.70	0.94	3.49
	Spanish	2,417	0.92	3.67	0.92	3.53
	All Translations	3,333	0.95	3.68	0.95	3.49

Standard Error of Measurement

SEMs, as computed from Cronbach's alpha and the Feldt-Raju reliability statistics, are presented in Table 34. SEMs based on Cronbach's alpha ranged 3.00–3.83, 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 37A–37F. SEMs associated with all reliability estimates for all subpopulations are in the range 2.41–4.00, 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.

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 38 depicts the consistency study results based on the range of performance levels for all grades. Overall, between 76 and 81% 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 0.65–0.74.

Table 38. Decision Consistency (All Cuts)

Grade	N-count	Agreement	Inconsistency	Kappa
3	199,587	0.76	0.24	0.65
4	195,956	0.78	0.22	0.69
5	198,201	0.76	0.24	0.66
6	200,802	0.79	0.21	0.71
7	199,113	0.78	0.22	0.70
8	199,783	0.81	0.19	0.74

Table 39 depicts the consistency study results based on two performance levels (passing and not passing) as defined by the Level III cut. Overall, about 91–93% 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 0.80–0.87.

Table 39. Decision Consistency (Level III Cut)

Grade	N-count	Agreement	Inconsistency	Kappa
3	199,587	0.91	0.09	0.81
4	195,956	0.91	0.09	0.81
5	198,201	0.91	0.08	0.80
6	200,802	0.92	0.08	0.84
7	199,113	0.92	0.08	0.82
8	199,783	0.93	0.07	0.87

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 82–86% of the time across all performance levels and approximately 93–95% of the time in regards to the Level III cut score.

Table 40. 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	199,587	**0.82**	0.11	0.07	**0.93**	0.02	0.04
4	195,956	**0.85**	0.07	0.08	**0.94**	0.02	0.04
5	198,201	**0.83**	0.08	0.09	**0.94**	0.02	0.04
6	200,802	**0.85**	0.09	0.06	**0.95**	0.04	0.02
7	199,113	**0.83**	0.13	0.04	**0.94**	0.05	0.01
8	199,783	**0.86**	0.10	0.05	**0.94**	0.05	0.01

Section VIII: Summary of Operational Test Results

This section summarizes the distribution of OP scale score results on the New York State 2012 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, need/resource capacity category, ELLs, SWDs, SUAs, and test language variables (Test Language) were used to calculate the results of subgroups required for federal reporting and test equity purposes. Especially, the ELL/SUA subgroup is defined as examinees whose ELL status are true and use one or more ELL-related accommodation. The SWD/SUA subgroup is defined as examinees with disabilities using one or more disability-related accommodations falling under 504 Plan. 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 41. 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- 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.

Table 41. 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	199,587	688.17	20.73	664	676	689	700	711
4	195,956	690.31	33.19	649	672	691	711	730
5	198,201	687.87	34.95	647	667	688	710	726
6	200,802	683.05	34.99	645	665	685	704	722
7	199,113	679.30	32.05	643	662	680	698	715
8	199,783	679.90	29.28	644	664	681	698	713

Grade 3

Scale score statistics and N-counts of demographic groups for Grade 3 are presented in Table 42. The population scale score mean was 688.17 with a standard deviation of 20.73. The gender subgroups performed the same, with a mean difference of 0.02 scale score points. Asian and White ethnic subgroups had scale score means that exceeded the State mean scale score on the test, as did students from Low Needs and Average Needs districts and the Charter schools. The lowest performing NRC subgroup was the Big 4 Cities, with a mean of 673.75, and the lowest performing ethnic subgroup was Black (mean scale score of 679.81). 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 663 (Haitian-Creole subgroup) to 700 (Korean subgroup), a difference that exceeds a standard deviation. The subgroup that used the Haitian-Creole translation had a scale score mean of 25 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 689: Asian (697), White (693), Average Needs (691), Low Needs (697), and students who used the Chinese (693) or Korean (700) translations.

Table 42. 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	199,587	688.17	20.73	664	676	689	700	711
Gender	Female	97,664	688.18	20.05	664	676	689	700	711
	Male	101,923	688.16	21.36	663	676	689	700	711
Ethnicity	Asian	16,917	699.09	20.40	676	688	697	711	725
	Black	35,827	679.81	20.48	655	668	680	693	703
	Hispanic	47,940	682.63	20.28	658	672	684	695	706
	American Indian	1,086	684.31	20.74	660	674	686	695	706
	Multiracial	2,018	687.88	21.16	663	675	688	700	711
	Other	416	690.67	22.66	665	679	691	703	716
	White	95,383	692.20	19.11	670	682	693	703	716
NRC	New York City	72,824	686.53	21.74	661	674	688	700	711
	Big 4 Cites	8,128	673.75	23.42	648	661	675	688	700
	High Needs Urban/Suburban	15,732	681.61	20.02	658	670	683	693	703
	High Needs Rural	11,031	684.02	18.98	663	674	684	695	706
	Average Needs	57,884	690.29	18.73	668	680	691	700	711
	Low Needs	28,021	697.16	18.16	676	686	697	706	716
	Charter	5,967	690.10	17.24	669	679	689	700	711

Table 42. Scale Score Distribution Summary, by Subgroup, Grade 3 (cont.)

Demographic Category (Subgroup)		N-count	SS Mean	SS Std Dev	10th %tile	25th %tile	50th %tile	75th %tile	90th %tile
SWD	All Codes	28,242	672.31	23.33	646	660	674	686	697
SUA	All Codes	23,078	675.05	22.18	650	663	676	689	700
SWD/SUA	SUA=504 Plan Codes	12,368	669.61	22.96	644	658	672	683	693
ELL/SUA	SUA=ELL Codes	5,943	675.89	20.30	652	665	678	688	697
ELL	English	195,831	688.43	20.59	664	676	689	700	711
	Chinese	474	692.97	17.79	672	683	693	703	716
	Haitian-Creole	50	664.34	23.53	639	650	663	683	691
	Korean	41	699.66	19.00	680	689	700	711	725
	Russian	93	682.57	20.68	652	673	686	695	703
	Spanish	3,098	671.39	22.40	646	660	673	686	695
	All Translations	3,756	674.60	23.15	648	663	676	689	700

Grade 4

Scale score statistics and N-counts of demographic groups for Grade 4 are presented in Table 43. The population scale score mean was 690.31 with a standard deviation of 33.19. The gender subgroups performed the same, with a mean difference of 0.15 scale score points. Asian and White students' scale score means exceeded the State mean scale score on the test. Asian students (the highest performing ethnic subgroup) exceeded the State mean by more than three-fifths of a standard deviation. Black, Hispanic, and American Indian ethnic subgroups had mean scale scores almost one standard deviation below the Asian subgroup. Students from Low Needs and Average Needs districts outperformed the other NRC subgroups. The lowest performing NRC subgroup was the Big 4 Cities, with a mean of 665.15, more than three-fourths 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 and a half 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, and Korean translation subgroups. At the 50th percentile, the following groups exceeded the population scale score of 691: Male (693), Asian (711), White (698), Average Needs (693), Low Needs (705), and ELL students who used the Chinese (705), Russian (695), or Korean (722) translations.

Table 43. 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	195,956	690.31	33.19	649	672	691	711	730
Gender	Female	96,155	690.23	32.29	651	672	691	711	730
	Male	99,801	690.38	34.04	649	672	693	711	730
Ethnicity	Asian	16,596	711.09	32.89	672	693	711	730	749
	Black	35,613	676.28	32.55	637	657	677	698	714
	Hispanic	46,302	681.36	31.94	642	662	683	703	718
	American Indian	1,042	681.98	33.16	639	662	683	703	722
	Multiracial	1,701	687.83	33.20	646	668	691	708	726
	Other	326	696.39	35.68	659	679	698	722	735
	White	94,376	696.45	30.53	661	679	698	714	730
NRC	New York City	70,753	688.84	34.90	646	668	689	711	730
	Big 4 Cites	8,111	665.15	36.13	622	644	668	689	705
	High Needs Urban/Suburban	15,307	678.77	32.69	639	661	681	700	718
	High Needs Rural	11,048	683.06	30.50	644	666	685	703	718
	Average Needs	57,665	692.75	29.89	657	675	693	711	726
	Low Needs	28,286	704.67	28.57	670	687	705	722	735
	Charter	4,786	693.88	27.91	659	677	693	711	730
SWD	All Codes	29,617	662.58	36.07	618	642	664	685	705
SUA	All Codes	20,724	665.65	35.16	622	646	668	689	705
SWD/SUA	SUA=504 Plan Codes	12,666	658.38	35.55	614	637	661	683	700
ELL/SUA	SUA=ELL Codes	3,798	667.46	34.40	625	649	670	689	708

Table 43. 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
ELL	English	192,150	690.71	32.99	651	672	693	711	730
	Chinese	527	702.45	27.28	670	687	705	718	735
	Haitian-Creole	66	640.94	41.24	597	614	645	666	693
	Korean	41	720.39	38.67	685	703	722	742	760
	Russian	81	687.72	33.87	642	662	695	703	730
	Spanish	3,091	663.64	34.01	622	644	666	687	703
	All Translations	3,806	669.74	36.59	622	649	672	693	711

Grade 5

Grade 5 demographic group N-counts and scale score statistics are presented in Table 44. The population scale score mean was 687.87 with a standard deviation of 34.95. The gender subgroups performed very similarly, with a mean difference of less than one scale score point. Asian and White students' scale score means exceeded the State mean scale score on the test. Asian students (the highest performing ethnic subgroup) exceeded the State mean by close to 23 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 660.55, nearly one-half of a standard deviation below the second lowest performing NRC subgroup (High Needs/Urban/ Suburban: 676.06) and 43 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 (630.82) of more than 57 units below the population mean, was the lowest performing group analyzed. The Korean translation subgroup was the highest performing group analyzed, with a scale score mean of 718.83, about four-fifths of standard deviation above the population mean. At the 50th percentile, the following groups exceeded the population scale score of 688: Asian (710), White (694), Average Needs (690), Low Needs (704), and students who used the Chinese (701) or Korean (721) translations.

Table 44. 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	198,201	687.87	34.95	647	667	688	710	726
Gender	Female	97,203	688.11	33.76	649	669	688	710	726
	Male	100,998	687.64	36.05	644	667	688	710	733

Table 44. Scale Score Distribution Summary, by Subgroup, Grade 5 (cont.)

Demographic Category (Subgroup)		N-count	SS Mean	SS Std Dev	10th %tile	25th %tile	50th %tile	75th %tile	90th %tile
Ethnicity	Asian	16,202	710.58	35.54	669	690	710	733	752
	Black	36,739	674.05	33.49	635	656	676	694	713
	Hispanic	46,034	679.14	33.49	641	660	681	699	717
	American Indian	996	677.64	36.13	638	658	678	699	717
	Multiracial	1,548	684.75	36.57	644	664	683	707	726
	Other	328	691.88	35.40	654	672	694	713	733
	White	96,354	693.63	32.66	656	674	694	713	733
NRC	New York City	69,660	687.50	36.18	647	665	688	710	733
	Big 4 Cites	8,083	660.55	37.20	616	641	662	685	704
	High Needs Urban/Suburban	15,122	676.06	33.43	638	658	678	697	713
	High Needs Rural	11,402	678.07	30.84	641	662	679	697	713
	Average Needs	59,175	689.52	32.16	652	671	690	710	726
	Low Needs	28,745	703.40	31.96	667	685	704	721	741
	Charter	6,014	686.70	28.86	654	669	688	704	721
SWD	All Codes	30,729	657.88	37.17	616	638	662	681	699
SUA	All Codes	26,448	662.35	37.29	622	644	665	685	704
SWD/SUA	SUA=504 Plan Codes	16,219	654.44	37.45	610	635	658	678	694
ELL/SUA	SUA=ELL Codes	4,925	662.60	37.21	622	644	665	685	704
ELL	English	194,659	688.29	34.72	647	669	688	710	726
	Chinese	474	701.26	29.18	667	683	701	717	741
	Haitian-Creole	55	630.82	51.98	570	610	641	671	686
	Korean	47	718.83	32.71	681	692	721	741	772
	Russian	89	676.69	32.07	635	662	678	699	713
	Spanish	2,877	658.32	36.44	616	641	662	681	697
	All Translations	3,542	664.90	39.31	622	644	667	688	710

Grade 6

Grade 6 scale score statistics and N-counts of demographic groups are presented in Table 45. The population scale score mean was 683.05 with a standard deviation of 34.99. The gender subgroups performed very similarly, with a mean difference of less than three scale score points. Asian, Multiracial, and White students' scale score means exceeded the State mean scale score. American Indian, Black, and Hispanic ethnic subgroups had mean scale scores approximately one standard deviation below the Asian subgroup. Students from Low Needs and Average Needs districts outperformed the other NRC subgroups. The lowest performing NRC subgroup was the Big 4 Cities, with a mean of 657.08. New York City, High Needs Urban/Suburban, and High Needs Rural subgroups had similar scale score means (ranging from approximately 670–676). 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 (646.57) more than 36 units below the population mean, was the lowest performing group analyzed. Asian students (the highest performing subgroup with a mean of 706.07) exceeded the State mean by 23 scale score points. At the 50th percentile, the following groups exceeded the population scale score of 685: Asian (707), White (691), Average Needs (687), Low Needs (700), and students who used the Chinese (693) or Korean (706) translations.

Table 45. 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	200,802	683.05	34.99	645	665	685	704	722
Gender	Female	98,129	684.33	33.25	647	668	685	704	722
	Male	102,673	681.82	36.53	640	663	684	704	722
Ethnicity	Asian	17,022	706.07	34.93	668	687	707	726	755
	Black	37,247	668.38	34.62	633	651	672	690	707
	Hispanic	45,281	672.06	34.34	633	655	674	693	709
	American Indian	1,016	671.63	38.32	633	657	676	693	712
	Multiracial	1,449	683.21	35.22	647	663	684	704	722
	Other	366	694.72	44.00	647	670	695	718	755
	White	98,421	689.74	31.23	655	674	691	709	726

Table 45. Scale Score Distribution Summary, by Subgroup, Grade 6 (cont.)

Demographic Category (Subgroup)		N-count	SS Mean	SS Std Dev	10th %tile	25th %tile	50th %tile	75th %tile	90th %tile
NRC	New York City	70,043	680.00	37.63	640	660	681	702	722
	Big 4 Cites	7,877	657.08	39.81	618	640	662	681	698
	High Needs Urban/Suburban	14,781	670.28	34.46	633	655	674	691	707
	High Needs Rural	11,459	676.29	31.95	642	662	680	695	712
	Average Needs	61,005	685.92	30.70	651	670	687	704	718
	Low Needs	30,106	699.72	29.04	668	684	700	718	732
	Charter	5,531	684.34	27.91	651	669	685	702	718
SWD	All Codes	30,408	651.49	39.94	611	636	657	676	691
SUA	All Codes	19,942	654.82	39.51	611	640	660	680	695
SWD/SUA	SUA=504 Plan Codes	13,186	648.75	40.29	600	633	655	673	688
ELL/SUA	SUA=ELL Codes	2,506	650.07	41.56	611	633	655	674	691
ELL	English	196,787	683.55	34.69	645	666	685	704	722
	Chinese	531	693.10	28.19	657	677	693	712	722
	Haitian-Creole	134	646.57	43.66	600	636	653	673	687
	Korean	44	703.77	36.89	665	687	706	720	755
	Russian	104	650.88	62.63	551	629	666	696	715
	Spanish	3,202	652.66	37.79	611	636	658	676	691
	All Translations	4,015	658.32	40.45	618	640	662	683	702

Grade 7

N-counts and scale score statistics of demographic groups for Grade 7 are presented in Table 46. The population scale score mean was 679.30 with a standard deviation of 32.05. The gender subgroups performed very similarly, with a mean difference of less than two scale score points. Asian and White ethnic subgroups' scale score means exceeded the State mean scale score. American Indian, Black, and Hispanic ethnic subgroups had mean scale scores between one-fifths and one-half of a standard deviation below the population. The lowest performing NRC subgroup, Big 4 Cities, had a scale score mean of 652.35, while the Low Needs subgroup's scale score mean was 694.77. 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, with a scale score mean of 645.03. At the 50th percentile, the following groups exceeded the population scale score of 680: Asian (702), White (688), Average Needs (685), Low Needs (696), Charter schools (681), and ELL students who used the English form (681), or Chinese (691) or Korean (701) translations.

Table 46. 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	199,113	679.30	32.05	643	662	680	698	715
Gender	Female	97,037	679.99	30.77	645	662	681	698	715
	Male	102,076	678.64	33.22	641	661	680	698	715
Ethnicity	Asian	15,940	701.39	34.06	662	681	702	718	741
	Black	37,258	663.89	31.13	629	650	666	683	698
	Hispanic	44,233	668.60	30.68	635	654	671	688	702
	American Indian	1,019	672.35	29.16	638	657	674	689	707
	Multiracial	1,390	678.94	30.47	645	661	680	698	715
	Other	333	682.81	30.70	647	664	681	702	718
	White	98,940	686.39	28.29	655	671	688	702	718
NRC	New York City	68,533	675.54	34.45	638	655	675	696	715
	Big 4 Cites	7,708	652.35	34.74	616	638	655	674	689
	High Needs Urban/Suburban	14,623	666.77	30.86	635	652	669	686	700
	High Needs Rural	11,715	672.74	27.25	643	659	674	689	702
	Average Needs	60,217	683.08	27.83	652	669	685	700	715
	Low Needs	31,640	694.77	27.38	664	680	695	709	727
	Charter	4,677	681.04	25.93	652	666	681	696	712
SWD	All Codes	29,828	649.61	35.46	616	635	654	671	686
SUA	All Codes	20,298	654.10	35.66	616	638	659	675	691
SWD/SUA	SUA=504 Plan Codes	13,789	648.77	35.92	610	635	654	671	685
ELL/SUA	SUA=ELL Codes	2,606	649.36	37.90	610	635	654	671	688

Table 46. 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	English	194,997	679.76	31.80	643	662	681	698	715
	Chinese	617	690.97	28.87	657	674	691	707	727
	Haitian-Creole	138	645.03	44.38	610	635	650	672	685
	Korean	44	701.20	32.79	664	679	701	718	741
	Russian	95	671.54	33.03	635	652	669	693	707
	Spanish	3,222	650.38	32.75	616	638	655	671	683
	All Translations	4,116	657.32	36.11	621	641	659	677	696

Grade 8

Grade 8 scale score statistics and N-counts of demographic groups are presented in Table 47. The population scale score mean was 679.90 with a standard deviation of 29.28. The gender subgroups performed similarly, with a mean difference of less than 3 scale score points. Asian, Multiracial, 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 below the population mean. The lowest performing NRC subgroup, Big 4 Cities, had a scale score mean of 653.08, while the Low Needs subgroup's scale score mean was 694.46, 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 681: Female (682), Asian (702), White (686), Average Needs (685), Low Needs (695), and students who used the Chinese (697) or Korean (708) translations.

Table 47. 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	199,783	679.90	29.28	644	664	681	698	713
Gender	Female	98,226	681.10	28.50	647	665	682	698	713
	Male	101,557	678.74	29.97	642	662	681	698	713

Table 47. 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
Ethnicity	Asian	16,226	700.25	29.46	665	683	702	716	737
	Black	37,062	666.23	28.80	633	651	668	685	698
	Hispanic	43,903	670.65	28.15	636	655	673	689	702
	American Indian	1,045	674.03	28.83	642	659	675	690	706
	Multiracial	1,146	680.08	28.71	644	662	681	698	713
	Other	354	681.78	31.96	642	661	681	702	724
	White	100,047	685.78	26.18	655	671	686	702	716
NRC	New York City	69,908	676.88	30.83	642	659	677	697	713
	Big 4 Cites	7,415	653.08	32.24	620	636	655	673	689
	High Needs Urban/Suburban	14,365	667.55	27.84	636	653	670	685	698
	High Needs Rural	11,619	673.37	25.16	644	659	674	689	702
	Average Needs	60,856	682.99	25.78	653	668	685	700	713
	Low Needs	32,185	694.46	25.04	665	679	695	711	724
	Charter	3,435	681.86	23.62	651	667	683	698	711
SWD	All Codes	29,633	653.22	30.78	620	636	655	673	686
SUA	All Codes	18,040	656.65	31.34	620	642	659	677	690
SWD/SUA	SUA=504 Plan Codes	12,807	652.61	31.03	620	636	655	673	686
ELL/SUA	SUA=ELL Codes	1,740	651.56	35.93	615	633	655	674	689
ELL	English	195,952	680.20	29.08	644	664	682	698	713
	Chinese	727	696.20	25.91	664	681	697	711	724
	Haitian-Creole	120	657.37	29.68	620	644	661	675	687
	Korean	49	705.71	16.29	686	697	708	716	724
	Russian	118	672.41	30.43	639	659	674	692	706
	Spanish	2,817	655.53	31.50	620	639	659	675	687
	All Translations	3,831	664.47	34.51	625	647	667	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 New York State Regents Mathematics Exam. The theoretical cut scores established in 2010 were used as the cut scores for the 2012 administration.

Table 48 shows the mathematics cut scores used for classification of students into the four performance levels in 2012.

Table 48. Mathematics Grades 3–8 Performance Level Cut Scores

Grade	2012 New York State Cut Scores		
	Level		
	II	III	IV
3	662	684	707
4	636	676	707
5	640	676	707
6	640	674	700
7	639	670	694
8	639	674	704

Tables 49–55 show the performance level distributions for all examinees from public and charter schools with valid scores. Table 49 presents performance level data for total populations of students in Grades 3–8. Tables 50–55 contain performance level data for selected subgroups of students. In general, these summaries reflect the same achievement trends as in the scale score summary discussion. Male and Female students performed similarly across grades. More White and Asian students were classified in Level III and above, as compared to their peers from other ethnic subgroups. Students from Low- and Average-Needs districts outperformed students from High-Needs districts (New York City, Big 4 Cities, High-Needs Urban/Suburban, and High-Needs Rural) and Charter schools. The subgroups that 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 49. 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	199,587	8.77	29.80	48.43	13.00	61.43
4	195,956	5.20	25.28	39.14	30.38	69.51
5	198,201	7.07	25.67	38.59	28.67	67.26
6	200,802	7.77	26.73	34.74	30.76	65.50
7	199,113	8.23	26.17	34.56	31.03	65.59
8	199,783	6.73	31.38	42.10	19.79	61.89

Grade 3

Performance level summaries and N-counts of demographic groups for Grade 3 are presented in Table 50. Statewide, 61.43% of the third-grade population was placed in Levels III and IV. American Indian, Black, Hispanic, and Multiracial 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 80% of students from Low-Needs districts were classified in Levels III and IV, whereas only about 33% of Big 4 Cities students were in Levels III and IV. Less than 35% of SWD, SUA, or those who used Haitian-Creole or Spanish translated test forms were classified in Level III or above; however, the subgroups for Korean and Chinese translations had more than 73% in Levels III and IV, with Korean students having the greatest percentage of more than 85%.

Table 50. 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	199,587	8.77	29.80	48.43	13.00	61.43
Gender	Female	97,664	8.25	30.35	48.95	12.44	61.39
	Male	101,923	9.26	29.28	47.94	13.53	61.47
Ethnicity	Asian	16,917	3.25	15.01	53.64	28.10	81.74
	Black	35,827	16.36	40.70	37.18	5.76	42.94
	Hispanic	47,940	12.67	37.31	42.87	7.16	50.03
	American Indian	1,086	11.23	35.54	44.29	8.93	53.22
	Multiracial	2,018	9.17	31.67	46.04	13.13	59.17
	Other	416	7.45	24.76	53.13	14.66	67.79
	White	95,383	4.90	24.48	54.61	16.01	70.62
NRC	New York City	72,824	10.81	31.96	44.77	12.46	57.23
	Big 4 Cites	8,128	26.37	40.71	28.89	4.04	32.92
	High Needs Urban/Suburban	15,732	14.06	38.79	40.55	6.60	47.15
	High Needs Rural	11,031	9.68	37.28	46.23	6.81	53.04
	Average Needs	57,884	5.54	27.73	53.49	13.24	66.73
	Low Needs	28,021	2.49	17.33	57.53	22.65	80.18
	Charter	5,967	4.93	29.81	52.87	12.38	65.26

Table 50. Performance Level Distributions Summary, by Subgroup, Grade 3 (cont.)

Demographic Category (Subgroup)		N-count	Level I %	Level II %	Level III %	Level IV %	Levels III & IV %
SWD	All Codes	28,242	27.83	42.32	26.68	3.17	29.85
SUA	All Codes	23,078	23.15	42.05	31.06	3.74	34.80
SWD/SUA	SUA=504 Plan Codes	12,368	31.69	43.36	23.12	1.82	24.94
ELL/SUA	SUA=ELL Codes	5,943	20.83	44.03	31.58	3.55	35.13
ELL	ELL status = Y	18,664	22.70	42.64	31.22	3.45	34.66
ELL Test Language	English	195,831	8.47	29.59	48.78	13.16	61.94
	Chinese	474	3.59	23.21	56.96	16.24	73.21
	Haitian-Creole	50	48.00	30.00	22.00	0.00	22.00
	Korean	41	4.88	9.76	56.10	29.27	85.37
	Russian	93	16.13	27.96	49.46	6.45	55.91
	Spanish	3,098	27.70	44.48	25.73	2.10	27.82
	All Translations	3,756	24.39	40.81	30.54	4.26	34.80

Grade 4

Performance level summaries and N-counts of demographic groups for Grade 4 are presented in Table 51. Statewide, 69.51% 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 1.84% of Asian students and 3.09% of White students. American Indian, Black, and Hispanic ethnic subgroups had percentages of students in Levels III and IV ranging from 52%–60%, but the percentages of the Multiracial, White, and Asian subgroups meeting standards for Levels III and IV (66.49%, 77.95%, and 87.98%, respectively) exceeded the population. Student achievement also varied widely by NRC subgroup. About 86% of students from Low-Needs districts were meeting standards for Levels III and IV, but only about 40% 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 (86.72%). The Korean translation subgroup had 92.68% 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, Charter, Chinese translation, and Korean translation.

Table 51. 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	195,956	5.20	25.28	39.14	30.38	69.51
Gender	Female	96,155	4.68	25.73	40.00	29.59	69.59
	Male	99,801	5.70	24.85	38.30	31.14	69.44

Table 51. Performance Level Distribution Summary, by Subgroup, Grade 4 (cont.)

Demographic Category (Subgroup)		N-count	Level I %	Level II %	Level III %	Level IV %	Levels III & IV %
Ethnicity	Asian	16,596	1.84	10.18	30.55	57.43	87.98
	Black	35,613	9.43	38.61	36.55	15.41	51.96
	Hispanic	46,302	7.41	33.08	40.12	19.39	59.51
	American Indian	1,042	7.58	33.59	36.47	22.36	58.83
	Multiracial	1,701	5.41	28.10	38.10	28.40	66.49
	Other	326	3.99	18.71	37.42	39.88	77.30
	White	94,376	3.09	18.97	41.20	36.75	77.95
NRC	New York City	70,753	5.91	28.12	36.75	29.22	65.97
	Big 4 Cites	8,111	17.72	42.67	29.75	9.86	39.61
	High Needs Urban/Suburban	15,307	8.75	35.14	38.50	17.61	56.11
	High Needs Rural	11,048	6.33	30.96	42.80	19.90	62.71
	Average Needs	57,665	3.54	22.25	42.56	31.65	74.21
	Low Needs	28,286	1.41	12.22	38.84	47.53	86.37
	Charter	4,786	1.98	23.00	44.38	30.63	75.01
SWD	All Codes	29,617	19.78	44.56	26.69	8.97	35.66
SUA	All Codes	20,724	17.41	42.51	30.47	9.61	40.08
SWD/SUA	SUA=504 Plan Codes	12,666	22.79	46.09	24.72	6.40	31.11
ELL/SUA	SUA=ELL Codes	3,798	15.90	42.31	31.78	10.01	41.79
ELL	ELL status = Y	17,495	15.03	42.34	32.16	10.47	42.63
ELL Test Language	English	192,150	4.98	25.03	39.29	30.71	70.00
	Chinese	527	2.28	11.01	42.50	44.21	86.72
	Haitian-Creole	66	39.39	39.39	18.18	3.03	21.21
	Korean	41	2.44	4.88	26.83	65.85	92.68
	Russian	81	6.17	30.86	40.74	22.22	62.96
	Spanish	3,091	18.86	43.55	29.89	7.70	37.59
	All Translations	3,806	16.47	38.28	31.63	13.61	45.24

Grade 5

Performance level summaries and N-counts of demographic groups for Grade 5 are presented in Table 52. Statewide, 67.26% 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 Multiracial ethnic subgroups were below the State average of students meeting standards for Levels III and IV (ranging 51–61%), as compared to the percentage of Asian and White students meeting standards for Levels III and IV (87% and 75%, respectively). Over 84% of students from Low-Needs districts were in Levels III or IV, but only about 35% of the Big 4 Cities students were in those levels. Only about 7–9% of SWD or SUA subgroups were placed in Level IV, compared to the population's 28.67% 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 (19.10%) and Chinese and Korean translation subgroups that had very high percentages of students in Level IV (44.51% and 63.83%, 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, Charter, Chinese translation, and Korean translation.

Table 52. 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	198,201	7.07	25.67	38.59	28.67	67.26
Gender	Female	97,203	6.35	25.75	39.52	28.37	67.90
	Male	100,998	7.76	25.59	37.69	28.95	66.64
Ethnicity	Asian	16,202	2.70	10.50	30.51	56.28	86.79
	Black	36,739	12.45	36.51	36.33	14.71	51.04
	Hispanic	46,034	9.71	32.52	38.73	19.03	57.77
	American Indian	996	11.85	34.34	35.44	18.37	53.82
	Multiracial	1,548	8.40	30.68	34.63	26.29	60.92
	Other	328	6.10	22.26	39.94	31.71	71.65
	White	96,354	4.43	20.66	40.84	34.08	74.92
NRC	New York City	69,660	7.48	27.00	36.83	28.69	65.51
	Big 4 Cites	8,083	24.52	40.88	25.72	8.88	34.60
	High Needs Urban/Suburban	15,122	11.43	34.77	37.38	16.41	53.80
	High Needs Rural	11,402	8.81	34.35	40.66	16.17	56.83
	Average Needs	59,175	5.26	24.16	41.82	28.76	70.58
	Low Needs	28,745	2.45	12.66	38.42	46.47	84.89
	Charter	6,014	4.49	27.52	44.50	23.50	67.99
SWD	All Codes	30,729	25.74	42.37	25.33	6.55	31.89
SUA	All Codes	26,448	22.03	40.82	28.34	8.81	37.15
SWD/SUA	SUA=504 Plan Codes	16,219	28.68	43.11	23.24	4.96	28.21
ELL/SUA	SUA=ELL Codes	4,925	21.69	41.69	27.74	8.89	36.63
ELL	ELL status = Y	15,068	21.26	41.63	28.44	8.67	37.11
ELL Test Language	English	194,659	6.81	25.46	38.76	28.97	67.73
	Chinese	474	1.90	15.40	38.19	44.51	82.70
	Haitian-Creole	55	49.09	32.73	18.18	0.00	18.18
	Korean	47	0.00	6.38	29.79	63.83	93.62
	Russian	89	12.36	32.58	35.96	19.10	55.06
	Spanish	2,877	24.50	41.85	27.98	5.67	33.65
	All Translations	3,542	21.23	37.46	29.42	11.89	41.30

Grade 6

Performance level summaries and N-counts of demographic groups for Grade 6 are presented in Table 53. Statewide, 65.50% of the sixth-grade population was placed in Levels III and IV. There was a slight performance differentiation by gender subgroup with less than 4% difference between each level. There were marked differences across ethnic and test translation subgroups. About 11–14% 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 47–54%), as compared to the percentage of Asian and White students meeting standards for Levels III and IV (86.75% and 75.18%, respectively). About 86% of students from Low-Needs districts were in Levels III or IV, but only about 35% of the Big 4 Cities students were. Only about 6–7% of SWD and SUA subgroups were placed in Level IV, compared to the population's 30.76% in Level IV. Less than 8% 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 (41.81% and 61.36%, 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, Charter, Chinese translation, and Korean translation.

Table 53. 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	200,802	7.77	26.73	34.74	30.76	65.50
Gender	Female	98,129	6.60	26.11	35.91	31.38	67.29
	Male	102,673	8.89	27.32	33.63	30.16	63.79
Ethnicity	Asian	17,022	2.78	10.46	26.47	60.28	86.75
	Black	37,247	13.88	39.15	32.00	14.97	46.97
	Hispanic	45,281	11.92	36.07	34.00	18.02	52.01
	American Indian	1,016	12.20	33.86	34.74	19.19	53.94
	Multiracial	1,449	6.21	30.57	33.33	29.88	63.22
	Other	366	5.74	24.59	24.86	44.81	69.67
	White	98,421	4.39	20.43	37.61	37.57	75.18
NRC	New York City	70,043	9.91	30.44	31.40	28.26	59.65
	Big 4 Cites	7,877	23.70	41.59	25.05	9.66	34.71
	High Needs Urban/Suburban	14,781	12.61	37.35	33.75	16.29	50.04
	High Needs Rural	11,459	8.19	33.57	38.04	20.20	58.24
	Average Needs	61,005	5.06	23.98	39.22	31.75	70.96
	Low Needs	30,106	2.18	11.94	34.31	51.57	85.88
	Charter	5,531	4.47	26.99	39.76	28.78	68.54
SWD	All Codes	30,408	27.57	44.72	21.69	6.02	27.71

Table 53. 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 %
SUA	All Codes	19,942	24.40	43.89	24.45	7.26	31.71
SWD/SUA	SUA=504 Plan Codes	13,186	29.46	45.60	20.45	4.49	24.94
ELL/SUA	SUA=ELL Codes	2,506	28.85	45.29	19.27	6.58	25.86
ELL	ELL status = Y	12,704	26.65	44.25	21.58	7.53	29.10
ELL Test Language	English	196,787	7.45	26.43	34.95	31.16	66.11
	Chinese	531	4.14	17.51	36.53	41.81	78.34
	Haitian-Creole	134	26.12	49.25	22.39	2.24	24.63
	Korean	44	4.55	11.36	22.73	61.36	84.09
	Russian	104	29.81	28.85	19.23	22.12	41.35
	Spanish	3,202	26.20	45.75	22.80	5.25	28.04
	All Translations	4,015	23.14	41.32	24.51	11.03	35.54

Grade 7

Performance level summaries and N-counts of demographic groups for Grade 7 are presented in Table 54. Statewide, 65.59% 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 44–56% of students meeting standards for Levels III and IV, with less than 22% of those students in Level IV, whereas over 85% of Asian students were meeting standards for Levels III and IV (and over 60% were in Level IV.) About 29% of Big 4 Cities students were meeting standards for Levels III and IV, with less than 8% in Level IV, yet over 86% of students from Low-Needs districts were meeting standards for Levels III and IV (about 52% in Level IV). Less than 8% of SWD and SUA subgroups were placed in Level IV, and over 25% were in Level I. Less than 8% 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 (45.87% and 61.36%, respectively). Across all subgroups, the Haitian-Creole translation subgroup had the largest percentage of students placed in Level I (31.16%), and the Korean translation subgroup had the largest percentage of students (84.09%) 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, Multiracial, White, Average Needs, Low Needs, Charter, Chinese translation, and Korean translation.

Table 54. 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	199,113	8.23	26.17	34.56	31.03	65.59
Gender	Female	97,037	7.40	25.95	35.46	31.19	66.65
	Male	102,076	9.03	26.39	33.71	30.87	64.59

Table 54. Performance Level Distribution Summary, by Subgroup, Grade 7 (cont.)

Demographic Category (Subgroup)		N-count	Level I %	Level II %	Level III %	Level IV %	Levels III & IV%
Ethnicity	Asian	15,940	3.24	11.02	25.39	60.36	85.75
	Black	37,258	15.83	40.13	30.63	13.41	44.04
	Hispanic	44,233	12.86	35.44	34.19	17.51	51.70
	American Indian	1,019	10.01	34.25	34.45	21.30	55.74
	Multiracial	1,390	6.98	28.85	34.96	29.21	64.17
	Other	333	6.61	26.43	34.53	32.43	66.97
	White	98,940	4.11	19.10	37.69	39.10	76.79
NRC	New York City	68,533	10.97	31.34	30.59	27.10	57.69
	Big 4 Cites	7,708	27.96	42.68	22.26	7.10	29.36
	High Needs Urban/Suburban	14,623	13.49	37.74	33.09	15.69	48.78
	High Needs Rural	11,715	8.54	32.60	40.05	18.80	58.86
	Average Needs	60,217	4.73	21.83	39.73	33.72	73.45
	Low Needs	31,640	2.22	11.53	34.30	51.95	86.25
	Charter	4,677	4.21	26.11	39.19	30.49	69.68
SWD	All Codes	29,828	30.14	43.22	21.11	5.53	26.64
SUA	All Codes	20,298	25.13	42.31	24.69	7.87	32.56
SWD/SUA	SUA=504 Plan Codes	13,789	29.89	44.19	21.10	4.81	25.91
ELL/SUA	SUA=ELL Codes	2,606	31.58	42.98	17.96	7.48	25.44
ELL	ELL status = Y	12,272	29.81	43.42	19.76	7.01	26.77
ELL Test Language	English	194,997	7.91	25.84	34.79	31.45	66.25
	Chinese	617	3.89	16.05	34.20	45.87	80.06
	Haitian-Creole	138	31.16	37.68	26.09	5.07	31.16
	Korean	44	2.27	13.64	22.73	61.36	84.09
	Russian	95	11.58	38.95	27.37	22.11	49.47
	Spanish	3,222	27.25	47.61	21.57	3.57	25.14
	All Translations	4,116	23.25	41.98	23.76	11.01	34.77

Grade 8

Performance level summaries and N-counts of demographic groups for Grade 8 are presented in Table 55. Statewide, 61.89% 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 9–13% 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 Multiracial ethnic subgroups had around 42–49% of students meeting standards for Levels III and IV, respectively, whereas about 85% of Asian students were meeting Levels III and IV standards. About 25% of Big 4 Cities students were in Levels III and IV, yet over 83%

of students from Low Needs districts were classified in these proficiency levels. Approximately 22–25% of SWD, SUA, and ELL students were placed in Level I. Less than 6% 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 (39.75% and 61.22%, respectively). Across all subgroups, the Spanish translation subgroup had the largest percentage of students placed in Level I (22.36%), and the Korean translation subgroup had the largest percentage of students placed in Level IV (61.22%). The following subgroups had a higher percentage of students meeting standards for Levels III and IV than the State population: Female, Asian, White, Average Needs, Low Needs, Charter, Chinese translation, and Korean translation.

Table 55. Performance Level Distribution Summary, by Subgroup, Grade 8

Demographic Category (Subgroup)		N-count	Level I %	Level II %	Level III %	Level IV %	Levels III & IV %
State	All Students	199,783	6.73	31.38	42.10	19.79	61.89
Gender	Female	98,226	5.87	30.44	43.23	20.47	63.69
	Male	101,557	7.56	32.29	41.01	19.14	60.16
Ethnicity	Asian	16,226	2.18	13.32	36.93	47.57	84.50
	Black	37,062	13.40	44.70	34.41	7.48	41.90
	Hispanic	43,903	10.21	40.82	39.22	9.75	48.97
	American Indian	1,045	8.71	38.37	39.90	13.01	52.92
	Multiracial	1,146	6.98	31.68	42.15	19.20	61.34
	Other	354	8.47	31.07	35.59	24.86	60.45
	White	100,047	3.43	25.15	47.10	24.32	71.41
NRC	New York City	69,908	8.40	36.00	37.57	18.03	55.60
	Big 4 Cites	7,415	26.45	48.67	20.81	4.07	24.88
	High Needs Urban/Suburban	14,365	11.54	44.36	37.03	7.07	44.10
	High Needs Rural	11,619	6.37	41.02	43.30	9.31	52.61
	Average Needs	60,856	4.09	27.69	48.19	20.04	68.22
	Low Needs	32,185	1.83	15.20	46.54	36.44	82.98
	Charter	3,435	3.81	30.13	48.12	17.93	66.06
SWD	All Codes	29,633	25.25	50.41	22.02	2.33	24.34
SUA	All Codes	18,040	21.90	48.94	25.47	3.70	29.17
SWD/SUA 3	SUA=504 Plan Codes	12,807	25.31	51.14	21.53	2.02	23.55
ELL/SUA 2	SUA=ELL Codes	1,740	29.43	45.34	20.75	4.48	25.23
ELL	ELL status = Y	12,237	21.88	45.26	26.93	5.92	32.86
ELL Test Language	English	195,952	6.52	31.16	42.35	19.97	62.32
	Chinese	727	1.10	16.37	42.78	39.75	82.53
	Haitian-Creole	120	17.50	54.17	25.83	2.50	28.33
	Korean	49	0.00	4.08	34.69	61.22	95.92
	Russian	118	9.32	39.83	39.83	11.02	50.85
	Spanish	2,817	22.36	49.73	25.45	2.45	27.90
	All Translations	3,831	17.49	42.65	29.31	10.55	39.86

Section IX: Longitudinal Comparison of Results

This section provides a longitudinal comparison of OP scale score results on the NYSTP 2006–2012 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 56.

Table 56. 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	2012	199,587	688.17	20.73	8.77	29.80	48.43	13.00	61.43
	2011	198,574	686.66	20.98	9.09	31.22	46.27	13.43	59.70
	2010	198,549	692.72	32.85	9.30	31.50	35.16	24.05	59.20
	2009	200,058	692.06	37.02	0.98	5.98	66.06	26.98	93.04
	2008	197,306	688.36	34.39	2.26	7.80	63.60	26.34	89.94
	2007	200,071	684.93	36.64	4.09	10.61	55.97	29.33	85.30
	2006	201,908	677.49	37.75	6.35	13.13	55.42	25.11	80.52
4	2012	195,956	690.31	33.19	5.20	25.28	39.14	30.38	69.51
	2011	199,134	687.96	32.21	5.54	27.73	39.99	26.74	66.73
	2010	201,418	686.99	34.69	5.26	30.84	38.14	25.75	63.90
	2009	197,379	689.59	38.28	3.69	9.00	51.82	35.49	87.31
	2008	198,509	683.13	38.11	4.70	11.37	54.49	29.45	83.93
	2007	199,181	679.91	39.85	6.02	13.97	52.52	27.49	80.01
	2006	202,695	676.55	40.81	7.41	14.59	52.12	25.88	78.00
5	2012	198,201	687.87	34.95	7.07	25.67	38.59	28.67	67.26
	2011	202,346	686.12	30.49	5.75	27.89	42.85	23.51	66.36
	2010	199,254	684.79	32.48	5.99	29.25	40.85	23.91	64.76
	2009	199,180	686.32	33.80	2.16	9.67	52.29	35.89	88.18
	2008	199,474	679.65	36.38	3.77	12.93	56.27	27.04	83.31
	2007	203,670	673.69	37.93	5.78	18.01	54.10	22.11	76.20
	2006	209,200	665.59	39.85	10.29	21.24	49.31	19.16	68.47
6	2012	200,802	683.05	34.99	7.77	26.73	34.74	30.76	65.50
	2011	200,076	682.16	31.60	7.91	28.99	36.72	26.38	63.10
	2010	200,415	680.25	33.85	7.96	30.58	34.27	27.19	61.46
	2009	199,605	679.91	35.21	3.56	13.30	55.02	28.12	83.14
	2008	201,719	674.85	38.21	5.45	15.04	53.21	26.31	79.52
	2007	205,976	667.96	40.34	8.71	19.94	51.33	20.02	71.35
	2006	211,376	655.94	40.44	13.32	26.23	47.26	13.19	60.45

Table 56. 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
7	2012	199,113	679.30	32.05	8.23	26.17	34.56	31.03	65.59
	2011	202,109	678.66	30.59	7.86	27.44	34.24	30.46	64.70
	2010	202,359	676.91	31.78	8.10	29.40	33.32	29.18	62.49
	2009	204,292	680.84	32.27	1.42	11.16	57.65	29.76	87.41
	2008	208,694	674.60	38.30	3.82	17.15	51.25	27.77	79.02
	2007	213,165	662.84	38.16	7.46	26.06	48.13	18.35	66.48
	2006	217,225	651.08	40.55	13.19	31.12	43.52	12.17	55.69
8	2012	199,783	679.90	29.28	6.73	31.38	42.10	19.79	61.89
	2011	203,235	677.25	33.66	8.63	31.39	42.29	17.69	59.98
	2010	206,346	677.18	32.37	9.19	35.94	36.60	18.27	54.87
	2009	208,835	674.99	33.75	3.47	16.18	61.09	19.27	80.36
	2008	210,265	666.44	38.19	7.31	22.69	53.10	16.89	69.99
	2007	215,108	656.93	38.62	12.21	28.90	46.97	11.92	58.89
	2006	219,294	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 2012 administrations and the scale scores from the 2006 and 2012 administrations can be directly compared, the performance level results between 2006–2009 and 2010–2012 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 56, an increase in scale score means was observed for all mathematics grades between 2006 and 2012. The least gain was observed for Grades 3 and 4, for which the total gain was about 11 and 14 scale score points, respectively, between the 2006 and 2012 test administrations. The greatest gain in scale score points between the 2006 and 2012 test administrations was noted for Grades 6, 7, and 8 (27, 28, and 28 scale score points, respectively).

The scale score standard deviation for Grade 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. There was a slight increase between 2011 and 2012.

The variability of scale score distribution decreased steadily across years for mathematics Grades 5, 6, and 7 from 2006 to 2011. 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. Overall, the scale score standard deviation decreased for Grades 3 and 8, and slightly increased for all other grades from 2011 to 2012.

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-stereotypical 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

Appendix B—Psychometric Guidelines for Operational Item Selection

It is primarily up to the Content Development department to select items for the 2012 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 2012 item selection included 2011 FT items and items owned by Pearson. 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 2012 OP forms and the 2011 OP forms.
- To the extent possible, select both easy and difficult items to provide good measurement information at both ends of the performance scale.
- To the extent possible, get the best scale coverage with selected items.
- Provide Research with the following item selection information:
    - Percentage of score points per learning standard (target, 2012 full selection, 2012 MC items only)
    - Item number in 2012 OP book
    - Item unique identification number, item type, FT year, FT form, and FT item number
    - Item classical statistics (p-values, point biserials, etc.)
    - 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)

		Initial Eigenvalues			
Grade	Subgroups	Component	Total	% of Variance	Cumulative %
3	ELL	**1**	**10.75**	**21.49**	**21.49**
		2	1.55	3.11	24.60
		3	1.15	2.30	26.90
		4	1.11	2.23	29.13
		5	1.05	2.11	31.24
		6	1.02	2.04	33.28
	SWD	**1**	**11.20**	**22.40**	**22.40**
		2	1.58	3.16	25.56
		3	1.16	2.32	27.88
		4	1.07	2.15	30.03
		5	1.03	2.06	32.09
		6	1.00	2.01	34.10
	SUA	**1**	**11.34**	**22.67**	**22.67**
		2	1.55	3.10	25.77
		3	1.14	2.28	28.05
		4	1.08	2.16	30.21
		5	1.02	2.05	32.26
	SWD/SUA	**1**	**10.77**	**21.54**	**21.54**
		2	1.58	3.15	24.69
		3	1.18	2.35	27.04
		4	1.06	2.13	29.17

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

		Initial Eigenvalues			
Grade	Subgroups	Component	Total	% of Variance	Cumulative %
3	SWD/SUA	5	1.05	2.11	31.28
		6	1.01	2.02	33.30
	ELL/SUA	**1**	**10.81**	**21.62**	**21.62**
		2	1.56	3.12	24.74
		3	1.15	2.29	27.03
		4	1.10	2.21	29.24
		5	1.06	2.12	31.36
		6	1.02	2.04	33.40
4	ELL	**1**	**11.98**	**21.39**	**21.39**
		2	1.68	3.00	24.39
		3	1.16	2.07	26.46
		4	1.10	1.97	28.43
		5	1.02	1.83	30.26
	SWD	**1**	**12.39**	**22.13**	**22.13**
		2	1.68	2.99	25.12
		3	1.17	2.09	27.21
		4	1.11	1.98	29.19
		5	1.03	1.83	31.02
	SUA	**1**	**12.80**	**22.86**	**22.86**
		2	1.68	3.00	25.86
		3	1.15	2.05	27.91
		4	1.11	1.98	29.89
		5	1.01	1.80	31.69
	SWD/SUA	**1**	**12.06**	**21.53**	**21.53**
		2	1.66	2.97	24.50
		3	1.18	2.11	26.61
		4	1.11	1.99	28.60
		5	1.03	1.83	30.43

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

		Initial Eigenvalues			
Grade	Subgroups	Component	Total	% of Variance	Cumulative %
4	SWD/SUA	6	1.00	1.79	32.22
	ELL/SUA	**1**	**11.95**	**21.34**	**21.34**
		2	1.68	2.99	24.33
		3	1.17	2.09	26.42
		4	1.11	1.98	28.40
		5	1.03	1.84	30.24
5	ELL	**1**	**11.22**	**22.00**	**22.00**
		2	1.47	2.88	24.88
		3	1.16	2.28	27.16
		4	1.05	2.06	29.22
		5	1.02	1.99	31.21
	SWD	**1**	**10.97**	**21.50**	**21.50**
		2	1.45	2.84	24.34
		3	1.19	2.34	26.68
		4	1.07	2.10	28.78
	SUA	**1**	**11.87**	**23.28**	**23.28**
		2	1.45	2.83	26.11
		3	1.16	2.28	28.39
		4	1.05	2.05	30.44
	SWD/SUA	**1**	**10.70**	**20.99**	**20.99**
		2	1.45	2.84	23.83
		3	1.21	2.36	26.19
		4	1.08	2.11	28.30
		5	1.00	1.97	30.27
	ELL/SUA	**1**	**11.37**	**22.29**	**22.29**
		2	1.49	2.92	25.21
		3	1.16	2.28	27.49
		4	1.06	2.07	29.56

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

		Initial Eigenvalues			
Grade	Subgroups	Component	Total	% of Variance	Cumulative %
5	ELL/SUA	5	1.01	1.98	31.54
6	ELL	**1**	**12.10**	**21.23**	**21.23**
		2	1.77	3.10	24.33
		3	1.22	2.14	26.47
		4	1.11	1.95	28.42
		5	1.09	1.91	30.33
		6	1.04	1.82	32.15
		7	1.02	1.79	33.94
		8	1.01	1.77	35.71
	SWD	**1**	**11.82**	**20.73**	**20.73**
		2	1.66	2.91	23.64
		3	1.22	2.14	25.78
		4	1.12	1.97	27.75
		5	1.09	1.91	29.66
		6	1.02	1.79	31.45
		7	1.01	1.76	33.21
	SUA	**1**	**12.70**	**22.28**	**22.28**
		2	1.73	3.03	25.31
		3	1.22	2.14	27.45
		4	1.10	1.93	29.38
		5	1.08	1.89	31.27
		6	1.00	1.76	33.03
	SWD/SUA	**1**	**11.28**	**19.79**	**19.79**
		2	1.61	2.82	22.61
		3	1.24	2.17	24.78
		4	1.12	1.97	26.75
		5	1.10	1.93	28.68
		6	1.03	1.82	30.50

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

		Initial Eigenvalues			
Grade	Subgroups	Component	Total	% of Variance	Cumulative %
6	SWD/SUA	7	1.02	1.80	32.30
		8	1.01	1.77	34.07
	ELL/SUA	**1**	**12.18**	**21.36**	**21.36**
		2	1.81	3.18	24.54
		3	1.22	2.14	26.68
		4	1.12	1.97	28.65
		5	1.11	1.94	30.59
		6	1.03	1.81	32.40
		7	1.02	1.79	34.19
		8	1.02	1.78	35.97
7	ELL	**1**	**10.24**	**18.61**	**18.61**
		2	2.02	3.66	22.27
		3	1.29	2.35	24.62
		4	1.08	1.96	26.58
		5	1.05	1.91	28.49
		6	1.04	1.89	30.38
		7	1.00	1.82	32.20
	SWD	**1**	**9.72**	**17.68**	**17.68**
		2	1.81	3.29	20.97
		3	1.36	2.47	23.44
		4	1.13	2.05	25.49
		5	1.05	1.92	27.41
		6	1.04	1.89	29.30
		7	1.00	1.82	31.12
	SUA	**1**	**10.65**	**19.36**	**19.36**
		2	1.91	3.48	22.84
		3	1.35	2.45	25.29
		4	1.11	2.02	27.31

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

		Initial Eigenvalues			
Grade	Subgroups	Component	Total	% of Variance	Cumulative %
7	SUA	5	1.03	1.87	29.18
		6	1.01	1.83	31.01
	SWD/SUA	**1**	**9.36**	**17.02**	**17.02**
		2	1.80	3.27	20.29
		3	1.36	2.48	22.77
		4	1.13	2.05	24.82
		5	1.06	1.93	26.75
		6	1.05	1.90	28.65
		7	1.01	1.84	30.49
		8	1.00	1.82	32.31
	ELL/SUA	**1**	**10.46**	**19.01**	**19.01**
		2	2.04	3.71	22.72
		3	1.29	2.35	25.07
		4	1.09	1.99	27.06
		5	1.06	1.93	28.99
		6	1.04	1.90	30.89
		7	1.01	1.84	32.73
8	ELL	**1**	**12.37**	**22.50**	**22.50**
		2	1.87	3.39	25.89
		3	1.48	2.69	28.58
		4	1.18	2.15	30.73
		5	1.09	1.97	32.70
		6	1.07	1.95	34.65
		7	1.04	1.90	36.55
		8	1.02	1.85	38.40
	SWD	**1**	**10.98**	**19.96**	**19.96**
		2	1.99	3.61	23.57
		3	1.56	2.83	26.40

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

		Initial Eigenvalues			
Grade	Subgroups	Component	Total	% of Variance	Cumulative %
8	SWD	4	1.25	2.28	28.68
		5	1.11	2.02	30.70
		6	1.07	1.94	32.64
		7	1.03	1.87	34.51
	SUA	**1**	**12.10**	**22.00**	**22.00**
		2	2.01	3.65	25.65
		3	1.54	2.80	28.45
		4	1.21	2.20	30.65
		5	1.08	1.97	32.62
		6	1.05	1.92	34.54
		7	1.02	1.85	36.39
	SWD/SUA	**1**	**10.69**	**19.43**	**19.43**
		2	2.00	3.63	23.06
		3	1.55	2.81	25.87
		4	1.24	2.25	28.12
		5	1.12	2.03	30.15
		6	1.07	1.95	32.10
		7	1.03	1.88	33.98
		8	1.01	1.84	35.82
	ELL/SUA	**1**	**12.63**	**22.96**	**22.96**
		2	1.87	3.40	26.36
		3	1.45	2.64	29.00
		4	1.20	2.18	31.18
		5	1.09	1.99	33.17
		6	1.07	1.95	35.12
		7	1.06	1.93	37.05
		8	1.03	1.87	38.92

Appendix D—Items Flagged for DIF

Table D1 supports the DIF information in Section V, "Operational Test Data Collection and Classical Analysis." It shows items flagged by the SMD and Mantel-Haenszel methods and includes item numbers, focal groups, and directions of DIF and DIF statistics. 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 2012 Classical DIF Item Flags

Grade	Item #	Subgroup	DIF	SMD	Mantel-Haenszel	Delta
3	48	ELL	Against	–0.1436	No Flag	No Flag
3	48	Spanish	Against	–0.1162	No Flag	No Flag
4	49	Female	In Favor	0.1055	No Flag	No Flag
4	53	Asian	Against	–0.1247	No Flag	No Flag
4	54	ELL	In Favor	0.1383	No Flag	No Flag
4	54	Spanish	In Favor	0.1415	No Flag	No Flag
4	55	ELL	In Favor	0.1081	No Flag	No Flag
4	55	Spanish	In Favor	0.1206	No Flag	No Flag
4	56	Black	Against	–0.1050	No Flag	No Flag
5	46	Black	Against	–0.1300	No Flag	No Flag
5	47	Black	In Favor	0.1088	No Flag	No Flag
5	49	Black	In Favor	0.1066	No Flag	No Flag
5	50	Female	In Favor	0.1163	No Flag	No Flag
6	52	ELL	In Favor	0.1361	No Flag	No Flag
6	52	Spanish	In Favor	0.1576	No Flag	No Flag
6	52	Black	In Favor	0.1013	No Flag	No Flag
6	52	Hispanic	In Favor	0.1029	No Flag	No Flag
6	54	Spanish	In Favor	0.1013	No Flag	No Flag
6	54	Female	In Favor	0.1672	No Flag	No Flag
6	55	ELL	In Favor	0.1237	No Flag	No Flag
6	55	Spanish	In Favor	0.1408	No Flag	No Flag

Table D1. NYSTP Mathematics 2012 Classical DIF Item Flags (cont.)

Grade	Item #	Subgroup	DIF	SMD	Mantel-Haenszel	Delta
6	56	Female	In Favor	0.1166	No Flag	No Flag
6	57	High Needs	Against	–0.1147	No Flag	No Flag
6	57	Black	Against	–0.1671	No Flag	No Flag
6	57	Hispanic	Against	–0.1375	No Flag	No Flag
7	03	ELL	Against	–0.1535	1377.77	–1.81
7	19	ELL	Against	–0.1194	911.13	–1.58
7	19	Asian	Against	No Flag	770.17	–1.51
7	19	Hispanic	Against	–0.1131	No Flag	No Flag
7	48	Female	In Favor	0.1298	No Flag	No Flag
7	50	High Needs	Against	–0.1063	No Flag	No Flag
7	50	Black	Against	–0.1725	No Flag	No Flag
7	53	ELL	Against	–0.1864	No Flag	No Flag
7	53	Spanish	Against	–0.2108	No Flag	No Flag
7	54	ELL	Against	–0.1147	No Flag	No Flag
7	54	Female	In Favor	0.1034	No Flag	No Flag
7	54	High Needs	Against	–0.1692	No Flag	No Flag
7	54	Black	Against	–0.1939	No Flag	No Flag
7	55	Female	In Favor	0.1118	No Flag	No Flag
8	12	Spanish	Against	–0.119	259.53	–1.65
8	14	ELL	Against	–0.1034	No Flag	No Flag
8	14	Spanish	Against	–0.1283	232.44	–1.64
8	21	Female	Against	–0.1003	No Flag	No Flag
8	26	Spanish	In Favor	0.159	357.81	1.73
8	27	Asian	In Favor	0.1128	No Flag	No Flag
8	38	Spanish	Against	–0.107	No Flag	No Flag
8	47	ELL	In Favor	0.1509	No Flag	No Flag
8	47	Spanish	In Favor	0.1454	No Flag	No Flag
8	47	Black	In Favor	0.1079	No Flag	No Flag
8	48	ELL	Against	–0.1384	No Flag	No Flag

Table D1. NYSTP Mathematics 2012 Classical DIF Item Flags (cont.)

Grade	Item #	Subgroup	DIF	SMD	Mantel-Haenszel	Delta
8	48	Spanish	Against	–0.182	No Flag	No Flag
8	50	ELL	Against	–0.122	No Flag	No Flag
8	50	Spanish	Against	–0.1167	No Flag	No Flag
8	50	Female	Against	–0.2253	No Flag	No Flag
8	50	Asian	Against	–0.1064	No Flag	No Flag
8	50	Black	Against	–0.1378	No Flag	No Flag
8	50	Hispanic	Against	–0.1271	No Flag	No Flag
8	52	ELL	Against	–0.1624	No Flag	No Flag
8	52	Asian	Against	–0.1381	No Flag	No Flag
8	52	Hispanic	Against	–0.1176	No Flag	No Flag
8	53	High Needs	Against	–0.1015	No Flag	No Flag
8	55	Spanish	In Favor	0.1021	No Flag	No Flag

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 $\beta(r_j, s_j)$:

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

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

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

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

where

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

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

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

with

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

and

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

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

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

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

Estimation of the Prior Distribution of T_j

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

The 3PL model is

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

where

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

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

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

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

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

and

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

Solving these equations for r_j and s_j produces

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

and

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

where

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

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

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

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

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

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, $x \in L_h$ ($h = 1, 2, \ldots, \text{H}$), and a latent score, $\theta \in \Gamma_w$ ($w = 1, 2, \ldots, \text{H}$), 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 H1–H6 depict the scale score (SS) distributions by N-count (frequency), percent, cumulative frequency, and cumulative percent for each grade (total population of students from public and charter schools).

Table G1. Grade 3 Mathematics 2012 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
470	123	0.06	123	0.06
574	71	0.04	194	0.10
604	137	0.07	331	0.17
615	205	0.10	536	0.27
623	291	0.15	827	0.41
628	347	0.17	1174	0.59
632	458	0.23	1632	0.82
636	542	0.27	2174	1.09
639	593	0.30	2767	1.39
642	761	0.38	3528	1.77
644	835	0.42	4363	2.19
646	854	0.43	5217	2.61
648	974	0.49	6191	3.10
650	1,115	0.56	7306	3.66
652	1,170	0.59	8476	4.25
654	1,229	0.62	9705	4.86
655	1,383	0.69	11088	5.56
657	1,470	0.74	12558	6.29
658	1,586	0.79	14144	7.09
660	1,623	0.81	15767	7.90
661	1,729	0.87	17496	8.77
663	1,931	0.97	19427	9.73
664	2,016	1.01	21443	10.74

Table G1. Grade 3 Mathematics 2012 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
665	2,176	1.09	23619	11.83
667	2,334	1.17	25953	13.00
668	2,510	1.26	28463	14.26
669	2,623	1.31	31086	15.58
670	2,787	1.40	33873	16.97
672	3,074	1.54	36947	18.51
673	3,189	1.60	40136	20.11
674	3,536	1.77	43672	21.88
675	3,712	1.86	47384	23.74
676	4,015	2.01	51399	25.75
678	4,415	2.21	55814	27.96
679	4,600	2.30	60414	30.27
680	5,067	2.54	65481	32.81
682	5,571	2.79	71052	35.60
683	5,928	2.97	76980	38.57
684	6,268	3.14	83248	41.71
686	6,894	3.45	90142	45.16
688	7,280	3.65	97422	48.81
689	7,933	3.97	105355	52.79
691	8,688	4.35	114043	57.14
693	9,083	4.55	123126	61.69
695	9,504	4.76	132630	66.45
697	10,169	5.10	142799	71.55
700	10,360	5.19	153159	76.74
703	10,433	5.23	163592	81.97
706	10,055	5.04	173647	87.00
711	9,183	4.60	182830	91.60
716	7,700	3.86	190530	95.46

Table G1. Grade 3 Mathematics 2012 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
725	5,460	2.74	195990	98.20
740	2,800	1.40	198790	99.60
770	797	0.40	199587	100.00

Table G2. Grade 4 Mathematics 2012 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
485	165	0.08	165	0.08
533	153	0.08	318	0.16
565	236	0.12	554	0.28
579	340	0.17	894	0.46
589	453	0.23	1347	0.69
597	517	0.26	1864	0.95
603	564	0.29	2428	1.24
609	751	0.38	3179	1.62
614	773	0.39	3952	2.02
618	845	0.43	4797	2.45
622	939	0.48	5736	2.93
625	991	0.51	6727	3.43
628	1,087	0.55	7814	3.99
631	1,132	0.58	8946	4.57
634	1,246	0.64	10192	5.20
637	1,364	0.70	11556	5.90
639	1,423	0.73	12979	6.62
642	1,551	0.79	14530	7.41
644	1,597	0.81	16127	8.23
646	1,770	0.90	17897	9.13
649	1,849	0.94	19746	10.08
651	1,904	0.97	21650	11.05
653	2,053	1.05	23703	12.10

Table G2. Grade 4 Mathematics 2012 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
655	2,205	1.13	25908	13.22
657	2,341	1.19	28249	14.42
659	2,406	1.23	30655	15.64
661	2,598	1.33	33253	16.97
662	2,689	1.37	35942	18.34
664	2,893	1.48	38835	19.82
666	3,006	1.53	41841	21.35
668	3,214	1.64	45055	22.99
670	3,490	1.78	48545	24.77
672	3,558	1.82	52103	26.59
674	3,752	1.91	55855	28.50
675	3,883	1.98	59738	30.49
677	4,086	2.09	63824	32.57
679	4,334	2.21	68158	34.78
681	4,465	2.28	72623	37.06
683	4,704	2.40	77327	39.46
685	4,894	2.50	82221	41.96
687	5,322	2.72	87543	44.67
689	5,416	2.76	92959	47.44
691	5,744	2.93	98703	50.37
693	5,977	3.05	104680	53.42
695	6,119	3.12	110799	56.54
698	6,271	3.20	117070	59.74
700	6,277	3.20	123347	62.95
703	6,617	3.38	129964	66.32
705	6,465	3.30	136429	69.62
708	6,668	3.40	143097	73.03
711	6,528	3.33	149625	76.36

Table G2. Grade 4 Mathematics 2012 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
714	6,546	3.34	156171	79.70
718	6,421	3.28	162592	82.97
722	6,115	3.12	168707	86.09
726	5,813	2.97	174520	89.06
730	5,450	2.78	179970	91.84
735	4,795	2.45	184765	94.29
742	3,940	2.01	188705	96.30
749	3,115	1.59	191820	97.89
760	2,295	1.17	194115	99.06
779	1,313	0.67	195428	99.73
800	528	0.27	195956	100.00

Table G3. Grade 5 Mathematics 2012 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
495	347	0.18	347	0.18
516	329	0.17	676	0.34
570	552	0.28	1228	0.62
589	691	0.35	1919	0.97
601	955	0.48	2874	1.45
610	1,109	0.56	3983	2.01
616	1,350	0.68	5333	2.69
622	1,470	0.74	6803	3.43
627	1,595	0.80	8398	4.24
631	1,776	0.90	10174	5.13
635	1,885	0.95	12059	6.08
638	1,956	0.99	14015	7.07
641	2,126	1.07	16141	8.14

Table G3. Grade 5 Mathematics 2012 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
644	2,248	1.13	18389	9.28
647	2,282	1.15	20671	10.43
649	2,401	1.21	23072	11.64
652	2,538	1.28	25610	12.92
654	2,644	1.33	28254	14.26
656	2,776	1.40	31030	15.66
658	2,949	1.49	33979	17.14
660	3,075	1.55	37054	18.70
662	3,093	1.56	40147	20.26
664	3,172	1.60	43319	21.86
665	3,331	1.68	46650	23.54
667	3,467	1.75	50117	25.29
669	3,482	1.76	53599	27.04
671	3,650	1.84	57249	28.88
672	3,787	1.91	61036	30.80
674	3,862	1.95	64898	32.74
676	4,018	2.03	68916	34.77
678	4,231	2.13	73147	36.91
679	4,336	2.19	77483	39.09
681	4,420	2.23	81903	41.32
683	4,651	2.35	86554	43.67
685	4,725	2.38	91279	46.05
686	4,899	2.47	96178	48.53
688	4,961	2.50	101139	51.03
690	5,224	2.64	106363	53.66
692	5,487	2.77	111850	56.43
694	5,699	2.88	117549	59.31
697	5,693	2.87	123242	62.18

Table G3. Grade 5 Mathematics 2012 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
699	6,020	3.04	129262	65.22
701	6,033	3.04	135295	68.26
704	6,091	3.07	141386	71.33
707	6,350	3.20	147736	74.54
710	6,425	3.24	154161	77.78
713	6,296	3.18	160457	80.96
717	6,358	3.21	166815	84.16
721	6,057	3.06	172872	87.22
726	5,938	3.00	178810	90.22
733	5,539	2.79	184349	93.01
741	4,967	2.51	189316	95.52
752	4,165	2.10	193481	97.62
772	3,013	1.52	196494	99.14
780	1,707	0.86	198201	100.00

Table G4. Grade 6 Mathematics 2012 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
500	1,188	0.59	1188	0.59
551	751	0.37	1939	0.97
584	1,050	0.52	2989	1.49
600	1,243	0.62	4232	2.11
611	1,517	0.76	5749	2.86
618	1,720	0.86	7469	3.72
624	1,844	0.92	9313	4.64
629	1,994	0.99	11307	5.63
633	2,116	1.05	13423	6.68
636	2,175	1.08	15598	7.77

Table G4. Grade 6 Mathematics 2012 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
640	2,153	1.07	17751	8.84
642	2,229	1.11	19980	9.95
645	2,312	1.15	22292	11.10
647	2,246	1.12	24538	12.22
649	2,325	1.16	26863	13.38
651	2,453	1.22	29316	14.60
653	2,430	1.21	31746	15.81
655	2,561	1.28	34307	17.08
657	2,468	1.23	36775	18.31
658	2,679	1.33	39454	19.65
660	2,726	1.36	42180	21.01
662	2,781	1.38	44961	22.39
663	2,739	1.36	47700	23.75
665	2,888	1.44	50588	25.19
666	2,909	1.45	53497	26.64
668	3,026	1.51	56523	28.15
669	3,103	1.55	59626	29.69
670	3,127	1.56	62753	31.25
672	3,265	1.63	66018	32.88
673	3,258	1.62	69276	34.50
674	3,380	1.68	72656	36.18
676	3,476	1.73	76132	37.91
677	3,537	1.76	79669	39.68
678	3,595	1.79	83264	41.47
680	3,835	1.91	87099	43.38
681	3,812	1.90	90911	45.27
683	3,798	1.89	94709	47.17
684	4,063	2.02	98772	49.19

Table G4. Grade 6 Mathematics 2012 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
685	4,189	2.09	102961	51.27
687	4,179	2.08	107140	53.36
688	4,330	2.16	111470	55.51
690	4,400	2.19	115870	57.70
691	4,507	2.24	120377	59.95
693	4,544	2.26	124921	62.21
695	4,620	2.30	129541	64.51
696	4,647	2.31	134188	66.83
698	4,853	2.42	139041	69.24
700	4,938	2.46	143979	71.70
702	4,835	2.41	148814	74.11
704	4,860	2.42	153674	76.53
707	4,862	2.42	158536	78.95
709	5,154	2.57	163690	81.52
712	5,009	2.49	168699	84.01
715	5,120	2.55	173819	86.56
718	4,922	2.45	178741	89.01
722	4,843	2.41	183584	91.43
726	4,533	2.26	188117	93.68
732	4,260	2.12	192377	95.80
741	3,774	1.88	196151	97.68
755	2,850	1.42	199001	99.10
780	1,801	0.90	200802	100.00

Table G5. Grade 7 Mathematics 2012 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
500	957	0.48	957	0.48
571	523	0.26	1480	0.74
591	768	0.39	2248	1.13
602	964	0.48	3212	1.61
610	1,107	0.56	4319	2.17
616	1,309	0.66	5628	2.83
621	1,417	0.71	7045	3.54
625	1,627	0.82	8672	4.36
629	1,742	0.87	10414	5.23
632	1,824	0.92	12238	6.15
635	2,071	1.04	14309	7.19
638	2,082	1.05	16391	8.23
641	2,301	1.16	18692	9.39
643	2,362	1.19	21054	10.57
645	2,573	1.29	23627	11.87
647	2,666	1.34	26293	13.21
650	2,833	1.42	29126	14.63
652	3,003	1.51	32129	16.14
654	2,986	1.50	35115	17.64
655	3,186	1.60	38301	19.24
657	3,440	1.73	41741	20.96
659	3,492	1.75	45233	22.72
661	3,555	1.79	48788	24.50
662	3,762	1.89	52550	26.39
664	3,835	1.93	56385	28.32
666	3,952	1.98	60337	30.30
667	4,001	2.01	64338	32.31
669	4,168	2.09	68506	34.41
671	4,275	2.15	72781	36.55

Table G5. Grade 7 Mathematics 2012 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
672	4,411	2.22	77192	38.77
674	4,507	2.26	81699	41.03
675	4,506	2.26	86205	43.29
677	4,568	2.29	90773	45.59
678	4,596	2.31	95369	47.90
680	4,631	2.33	100000	50.22
681	4,594	2.31	104594	52.53
683	4,607	2.31	109201	54.84
685	4,632	2.33	113833	57.17
686	4,626	2.32	118459	59.49
688	4,646	2.33	123105	61.83
689	4,801	2.41	127906	64.24
691	4,719	2.37	132625	66.61
693	4,704	2.36	137329	68.97
695	4,518	2.27	141847	71.24
696	4,611	2.32	146458	73.56
698	4,529	2.27	150987	75.83
700	4,557	2.29	155544	78.12
702	4,491	2.26	160035	80.37
704	4,614	2.32	164649	82.69
707	4,385	2.20	169034	84.89
709	4,185	2.10	173219	87.00
712	4,095	2.06	177314	89.05
715	3,870	1.94	181184	91.00
718	3,814	1.92	184998	92.91
722	3,414	1.71	188412	94.63
727	2,999	1.51	191411	96.13
733	2,741	1.38	194152	97.51

Table G5. Grade 7 Mathematics 2012 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
741	2,248	1.13	196400	98.64
756	1,719	0.86	198119	99.50
800	994	0.50	199113	100.00

Table G6. Grade 8 Mathematics 2012 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
480	372	0.19	372	0.19
566	323	0.16	695	0.35
589	503	0.25	1198	0.60
601	749	0.37	1947	0.97
609	949	0.48	2896	1.45
615	1,167	0.58	4063	2.03
620	1,460	0.73	5523	2.76
625	1,741	0.87	7264	3.64
629	1,806	0.90	9070	4.54
633	2,092	1.05	11162	5.59
636	2,278	1.14	13440	6.73
639	2,467	1.23	15907	7.96
642	2,554	1.28	18461	9.24
644	2,684	1.34	21145	10.58
647	2,841	1.42	23986	12.01
649	2,849	1.43	26835	13.43
651	2,980	1.49	29815	14.92
653	3,186	1.59	33001	16.52
655	3,178	1.59	36179	18.11
657	3,341	1.67	39520	19.78
659	3,364	1.68	42884	21.47

Table G6. Grade 8 Mathematics 2012 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
661	3,506	1.75	46390	23.22
662	3,520	1.76	49910	24.98
664	3,539	1.77	53449	26.75
665	3,705	1.85	57154	28.61
667	3,735	1.87	60889	30.48
668	3,653	1.83	64542	32.31
670	3,794	1.90	68336	34.21
671	3,908	1.96	72244	36.16
673	3,885	1.94	76129	38.11
674	3,954	1.98	80083	40.08
675	3,976	1.99	84059	42.08
677	4,076	2.04	88135	44.12
678	4,038	2.02	92173	46.14
679	4,150	2.08	96323	48.21
681	4,120	2.06	100443	50.28
682	4,163	2.08	104606	52.36
683	4,148	2.08	108754	54.44
685	4,280	2.14	113034	56.58
686	4,229	2.12	117263	58.70
687	4,263	2.13	121526	60.83
689	4,299	2.15	125825	62.98
690	4,234	2.12	130059	65.10
692	4,401	2.20	134460	67.30
693	4,273	2.14	138733	69.44
695	4,259	2.13	142992	71.57
697	4,340	2.17	147332	73.75
698	4,325	2.16	151657	75.91
700	4,248	2.13	155905	78.04

Table G6. Grade 8 Mathematics 2012 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
702	4,334	2.17	160239	80.21
704	4,235	2.12	164474	82.33
706	4,270	2.14	168744	84.46
708	4,188	2.10	172932	86.56
711	4,186	2.10	177118	88.66
713	3,994	2.00	181112	90.65
716	3,861	1.93	184973	92.59
720	3,639	1.82	188612	94.41
724	3,254	1.63	191866	96.04
730	2,984	1.49	194850	97.53
737	2,417	1.21	197267	98.74
751	1,642	0.82	198909	99.56
775	874	0.44	199783	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.

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

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

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

Dorans, N.J. & Holland, P.W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum.

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.

Fleiss, J.L. & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33: 613–619.

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, S., & Kolen, M. J. (2004). *STUIRT: A computer program for scale transformation under unidimensional item response theory models*. Iowa City: Iowa Testing Programs, The University of Iowa.

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.
Thissen, D. (1991). *MULTILOG* [Computer program]. Chicago, IL: Scientific Software, Inc.
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.