New York State Testing Program 2012: English Language Arts, Grades 3–8

Technical Report

Pearson
2012

Developed and published under contract with the New York State Education Department by Pearson, 2510 North Dodge Street, Iowa City, Iowa 52245.

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 English Language Arts (ELA) 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 Grades 3–8 ELA Tests target student progress toward three of the four content standards as described in Section II, "Test Design and Development," subsection "Content Rationale." The ELA Tests are written for all students to have the opportunity to demonstrate their knowledge and skills in these standards. The established cut scores classify student proficiency into one of four levels based on their test performance.

Target Population

Students in New York State public school Grades 3, 4, 5, 6, 7, and 8 (and ungraded students of equivalent ages) are the target population for the Grades 3–8 testing program. 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 *NYSTP Grades 3–8 English Language Arts and Mathematics Tests 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 ELA Tests are used to measure the extent to which individual students achieve the New York State Learning Standards in ELA and to determine whether schools, districts, and the state meet the required progress targets specified in the New York State accountability system. There are several types of scores available from the Grades 3–8 ELA Tests and they are discussed in this section.

Scale Scores

The scale score is a quantification of the ability measured by the Grades 3–8 ELA Tests at each grade level. The scale scores are comparable within each grade level but not across grades because the Grades 3–8 ELA Tests are not on a vertical scale. The test scores are reported at the individual level and can also be aggregated. Detailed information on the derivation and properties of scale scores is provided in Section VI, "IRT Scaling and Equating." The

Grades 3–8 ELA Tests scores are used to determine student progress within schools and districts, support registration of schools and districts, determine eligibility of students for additional instruction time, and provide teachers with indicators of a student's need, or lack of need, for remediation in specific content-area knowledge.

Proficiency Level Cut Scores and Classification

Students are classified as Level I (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 between 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 the 2009–2010 school years, and a decision to align the proficiency standards with Grade 8 student performance on the New York State Regents ELA 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, "Proficiency Level Cut Score Adjustment" of the *New York State Testing Program 2010: English Language Arts, 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 English Language Arts* and the *New York State ELA Measurement Review Technical Report 2006 for English Language Arts.*

Standard Performance Index Scores

Standard performance index (SPI) scores are obtained from the Grades 3–8 ELA Tests. The SPI score is an indicator of student ability, knowledge, and skills in specific learning standards, and it 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 use of SPI scores are provided in Section VI, "IRT Scaling and Equating."

Testing Accommodations

In accordance with federal law under the Americans with Disabilities Act and 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 2012 *SAM*.

Test Transcriptions

For visually impaired students, large type and braille editions of the test books are provided. The students dictate and/or record their responses, the teachers transcribe student responses to multiple-choice (MC) questions onto scannable answer sheets, and the teachers transcribe the

responses to 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; they all 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 assessments of proficiency in English language arts, the Grades 3–8 ELA Tests are not translated into any other language.

Section II: Test Design and Development

Test Description

The Grades 3–8 ELA Tests are New York State Learning Standards-based criterion-referenced tests composed of multiple-choice (MC) and constructed-response (CR) items. The tests were administered in New York 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 Configuration

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 English Language Arts Test Book 1 and Book 2 contained literary and informational reading passages and multiple-choice questions based on the passages. In addition, Book 2 contained multiple-choice questions, short-response questions, and an extended-response question based on a listening selection read aloud to the class. Book 3 contained reading passages with short-response questions and an extended-response question based on those passages.

To allow students sufficient time to demonstrate what they had learned, schools were instructed to schedule 90 minutes for each session, on each day and 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 English Language Arts 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 was a field test question that did not count toward their scores or an operational test question that did count toward their scores. The specific locations of the embedded items on a test form were not disclosed. These data are free from the effects of differential student motivation that may characterize stand-alone field-test designs because the items were 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 it.

Table 1 provides information on the number and type of items in each book. The 2012 *Teacher's Directions* (http://www.p12.nysed.gov/apda/ei/directions/2012/ela3-5-td-12w.pdf and http://www.p12.nysed.gov/apda/ei/directions/2012/ela6-8-td-12w.pdf) as well as the 2012 *SAM* is available online at (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 ELA 2012 Test Configuration

Grade	Day	Book	Number of Items				
			Multiple-Choice		Constructed-Response		Total*
			Operational	Embedded	Operational	Embedded	
3	1	1	30	6	0	0	36
	2	2	17	0	4	0	21
	3	3	0	0	5	0	5
	Totals		47	6	9	0	62
4	1	1	31	6	0	0	37
	2	2	20	0	4	0	24
	3	3	0	0	5	0	5
	Totals		51	6	9	0	66
5	1	1	33	6	0	0	39
	2	2	18	0	4	0	22
	3	3	0	0	5	0	5
	Totals		51	6	9	0	66
6	1	1	33	6	0	0	39
	2	2	18	0	4	0	22
	3	3	0	0	5	0	5
	Totals		51	6	9	0	66
7	1	1	33	6	0	0	39
	2	2	18	0	4	0	22
	3	3	0	0	5	0	5
	Totals		51	6	9	0	66
8	1	1	33	6	0	0	39
	2	2	18	0	4	0	22
	3	3	0	0	5	0	5
	Totals		51	6	9	0	66

*Reflects actual items in the test books.

Test Blueprint

The NYSTP Grades 3–8 ELA Tests assess students on three learning standards (S1—Information and Understanding, S2—Literary Response and Expression, and S3— Critical Analysis and Evaluation). The test items are indicators used to assess a variety of reading, writing, and listening skills against each of the three Learning Standards. Standard 1 is assessed primarily by the use of test items associated with informational passages; Standard 2 is assessed primarily by the use of test items associated with literary passages; and Standard 3 is assessed by the use of test items associated with a combination of genres. The distribution of score points across the Learning Standards was determined during blueprint-specifications meetings held with panels of New York State educators at the start of the testing program, prior to item development. The distribution in each grade reflects the number of assessable performance indicators in each Learning Standard at that grade and the emphasis placed on those performance

indicators by the blueprint-specifications panel members. Table 2 shows the Grades 3–8 ELA Tests blueprint and actual number of score points in the 2012 OP tests.

Table 2. NYSTP ELA 2012 Test Blueprint

Grade	Total Points on OP Test	Standard	Target Points	Selected Points	Target % of Test	Selected % of Test
3	67	S1	22	25	32.8	37.3
		S2	32	28	47.8	41.8
		S3	13	14	19.4	20.9
4	73	S1	26	27	35.6	37.0
		S2	33	34	45.2	46.6
		S3	14	12	19.2	16.4
5	73	S1	28	26	38.4	35.6
		S2	28	28	38.4	38.4
		S3	17	19	23.3	26.0
6	73	S1	26	26	35.6	35.6
		S2	33	31	45.2	42.5
		S3	14	16	19.2	21.9
7	73	S1	28	25	38.4	34.2
		S2	28	27	38.4	37.0
		S3	17	21	23.3	28.8
8*	67	S1	28	30	38.4	44.8
		S2	28	19	38.4	28.4
		S3	17	18	23.3	26.9

*For grade 8, one passage was exposed to the public during test administration. Therefore, the passage and the items associated with the passage were removed from the analyses, scoring, and reporting.

New York State Educators' Involvement in Test Development

New York State educators are actively involved in ELA test development at different test stages, including the 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 the performance indicators to make the following determinations:

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

The committees were composed of teachers from around the state who were selected for their grade-level expertise, were grouped by grade band (i.e., Grades 3/4, 5/6, 7/8) and met for four days. The committees were composed of approximately ten participants per grade band. Upon completion of the committee meetings, NYSED reviewed the committees' determinations and approved them, with minor adjustments when necessary to maintain consistency across the grades.

Item Development and Review

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

The content specialists at Pearson then selected passages that would best elicit the types of items outlined during the test specifications meetings and distributed writing assignments to experienced item writers. The writers' assignments outlined the number and type of items (including depth-of-knowledge or thinking skill level) to write for each passage. Item writers were trained in the New York State Learning Standards and specifications (which provide

information such as limitations and examples for assessing performance indicators) and were provided with item-writing guidelines (see Appendix B), sample New York State test items, and the New York State Style Guide.

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 passages and items into field test (FT) forms and submitted the FT forms to NYSED for their review and approval. NYSED verified that the passages and 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 2011 census sample 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 English Language Arts 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; the second, within the limits set by these requirements, was for developers to select 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* and two sets of *Teacher's Directions*—one for Grades 3, 4, and 5, and one for Grades 6, 7, and

8—so that the tests could be administered in a standardized fashion across the state. These documents are located at the following web site: http://www.p12.nysed.gov/apda/english/ela-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 ELA examinations, led to changes in the proficiency cut scores. The results were reviewed by the New York State Technical Advisory Group and were approved by the Board of Regents in July 2010. For each grade level, there are four proficiency levels. Three cut points demarcate the performance standards needed to demonstrate each ascending level of proficiency.

Section III: Validity

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

The American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) *Standards for Educational and Psychological Testing* (1999) addressed the concept of validity in testing. Validity is the most important consideration in test evaluation. The concept refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores. Test validation is the process for accumulating evidence to support any particular inference. Validity, however, is a unitary concept. Although evidence may be accumulated in many ways, validity refers to the degree to which evidence supports the inferences made from test scores.

Content Validity

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

Logical analysis of test content indicates the degree to which the content of a test covers the domain of content the test is intended to measure. In the case of 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 ELA Tests was conducted using Norman Webb's method. The results of the study provided additional evidence of test content validity (refer to *An External Alignment Study for New York State's Assessment Program*, April 2006, Educational Testing Services).

Construct (Internal Structure) Validity

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

Internal Consistency

Empirical studies of the internal structure of the test provide one type of evidence of construct validity. For example, high internal consistency constitutes evidence of validity. This is because high coefficients imply that the test questions are measuring the same domain of skill and are reliable and consistent. Reliability coefficients of the tests for total populations and subgroups of students are presented in Section VII, "Reliability and Standard Error of Measurement." For the total population, the reliability coefficients (Cronbach's alpha) ranged from 0.90–0.92, and for all subgroups the reliability coefficient was equal to or greater than 0.84. Overall, high internal consistency of the NYSTP ELA Tests provided sound evidence of construct validity.

Unidimensionality

Other evidence comes from analyses of the degree to which the test questions conform to the requirements of the statistical models used to scale and equate the tests, as well as to generate student scores. Among other things, the models require that the items fit the model well and the questions in a test measure a single domain of skill; that is, 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 on the 2012 Grades 3–8 ELA Tests displayed good item-model fit, which provided solid evidence for the appropriateness of IRT models used to calibrate and scale the test data. Another evidence for the efficacy of modeling ability was provided by demonstrating that the questions on NYSTP ELA Tests were related. What relates the questions is most parsimoniously claimed to be the common ability acquired by students studying the subject. Factor analysis of the test data is one way of modeling the common ability. This analysis may show that there is a single or main factor that can account for much of the variability among responses to test questions. A large first component would provide evidence of the latent ability that students have in common with respect to the particular questions asked. A large main factor found from a factor analysis of an achievement test would suggest a primary-ability construct that may be considered related to what the questions were designed to have in common, i.e., English language arts ability.

To demonstrate the common factor (ability) underlying student responses to ELA test items, a principal component factor analysis was conducted on a correlation matrix of individual items for each test. Factoring a correlation matrix rather than actual item response data is preferable when dichotomous variables are in the analyzed data set. Because the NYSTP ELA Tests contain both MC and CR items, the matrix of polychoric correlations was used as input for factor analysis (polychoric correlation is an extension of tetrachoric correlations that are appropriate only for MC items). The study was conducted on the total population of New York State public- and charter-school students in each grade. A large first principal component was evident in each analysis.

More than one factor with an eigenvalue greater than 1.0 present in each data set would suggest the presence of small additional factors. However, the ratio of the variance accounted for by the

first factor to the remaining factors was sufficiently large to support the claim that these tests were essentially unidimensional. These ratios showed that the first eigenvalues were at least four times as large as the second eigenvalues for all the grades. In addition, the total amount of variance accounted for by the main factor was evaluated. According to M. Reckase (1979),

> ***. . .*** *the 1PL and the 3PL models estimate different abilities when a test measures independent factors, but . . . both estimate the first principal component when it is large relative to the other factors. In this latter case, good ability estimates can be obtained from the models, even when the first factor accounts for less than 10 percent of the test variance, although item calibration results will be unstable.*

It was found that all the NYSTP Grades 3–8 ELA Tests exhibited first principal components accounting for more than 17% 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 ELA Tests (Total Population)

	Initial Eigenvalues			
Grade	Component	Total	% of Variance	Cumulative %
3	**1**	**11.67**	**20.83**	**20.83**
	2	1.64	2.93	23.76
	3	1.27	2.27	26.03
	4	1.12	2.00	28.03
	5	1.10	1.97	30.00
	6	1.05	1.87	31.87
4	**1**	**11.28**	**18.80**	**18.80**
	2	1.39	2.31	21.11
	3	1.25	2.09	23.20
	4	1.15	1.92	25.12
	5	1.06	1.76	26.88
	6	1.04	1.73	28.61
	7	1.01	1.68	30.29
5	**1**	**10.92**	**18.20**	**18.20**
	2	1.37	2.29	20.49
	3	1.23	2.05	22.54
	4	1.14	1.90	24.44
	5	1.06	1.77	26.21
	6	1.05	1.75	27.96

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

	Initial Eigenvalues			
Grade	Component	Total	% of Variance	Cumulative %
5	7	1.03	1.71	29.67
6	**1**	**10.28**	**17.14**	**17.14**
	2	1.81	3.01	20.15
	3	1.17	1.95	22.10
	4	1.10	1.84	23.94
	5	1.05	1.75	25.69
	6	1.02	1.70	27.39
	7	1.00	1.67	29.06
7	**1**	**11.48**	**19.14**	**19.14**
	2	1.54	2.57	21.71
	3	1.30	2.17	23.88
	4	1.15	1.92	25.80
	5	1.04	1.74	27.54
	6	1.02	1.70	29.24
	7	1.01	1.69	30.93
8	**1**	**9.50**	**17.59**	**17.59**
	2	1.58	2.93	20.52
	3	1.18	2.19	22.71
	4	1.07	1.98	24.69
	5	1.05	1.94	26.63
	6	1.02	1.88	28.51
	7	1.01	1.86	30.37

This evidence supports the claim that there is a construct ability underlying the items/tasks in each ELA Test and that scores from each test would be representing performance primarily determined by that ability. Construct-irrelevant variance does not appear to create significant nuisance factors.

As additional evidence for construct validity, the same factor analysis procedure was employed to assess dimensionality of ELA construct for selected subgroups of students in each grade: English language learners (ELL), students with disabilities (SWD), and students using test accommodations (SUA). The results were comparable to the results obtained from the total population data. Evaluation of eigenvalue magnitude and proportions of variance explained by the main and secondary factors provide evidence of essential unidimensionality of the construct

measured by the ELA Tests for the analyzed subgroups. Factor analysis results for ELL, SWD, SUA, ELL/SUA, and SWD/SUA classifications are provided in Table D1 of Appendix D. The ELL/SUA subgroup is defined as examinees whose ELL statuses are true and who use one or more ELL-related accommodation. The SWD/SUA subgroup includes with examinees who are classified with disabilities and use one or more disability-related accommodations.

Minimization of Bias

Minimizing item bias contributes to minimization of construct-irrelevant variance and contributes to improved test validity. The developers of the NYSTP tests gave careful attention to questions of possible ethnic, gender, and socioeconomic status (SES) bias. All materials were written and reviewed to conform to Pearson's editorial policies and guidelines for equitable assessment, as well as NYSED's guidelines for item development. At the same time, all materials were written to NYSED's specifications and carefully checked by groups of trained New York State educators during the item review process.

Four procedures were used to eliminate bias and minimize differential item functioning (DIF) in the NYSTP ELA Tests.

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

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

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

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

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

Section IV: Test Administration and Scoring

Listed in this section are brief summaries of New York State test administration and scoring procedures. For further information, refer to the *New York State Scoring Leader Handbooks* and *School Administrator's Manual* (SAM). In addition, please refer to the *Scoring Site Operations Manual* (2012) located at http://www.p12.nysed.gov/apda/ei/ssom/ssom-12w.pdf.

Test Administration

NYSTP Grades 3–8 ELA 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 school-wide 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 ELA leader trained the scoring-committee members and monitored the sessions, and a school ELA leader assisted in monitoring the sessions. For school-wide scoring, oversight was provided by the principal; otherwise, titles for the school-wide 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 meet their needs for the scoring of the Grades 3–8 ELA 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. 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 that were consensus scored by NYSED and New York State teachers during Rangefinding sessions. In 2012, the Pearson ELA 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 also 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 ELA leader conducted a "read-behind" of approximately 12 sets of test books per hour to verify the accuracy of scoring. If a question arose that was not covered in the training materials, facilitators or trainers were to call the New York State ELA Helpline (see the subsection "Quality Control Process").

Scorer Qualifications and Training

The scoring of the OP test was conducted by qualified administrators and teachers. Trainers used the scoring guides 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 a facilitator could not reach a decision on a paper after reviewing the scoring guides and audio files, they called the New York State ELA Helpline. This call center was established to help teachers and administrators during OP scoring. The help-line 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, it was 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 in May 2012. These data were used for all data analysis except section VII and VIII. Phase 2 involved submitting "straggler files" to Pearson in late June 2012. The straggler files were later merged with the main data sets. The straggler files contained around 2% of the total population cases and due to late submission were excluded from research data analyses. Data from nonpublic schools were excluded from any data analysis.

Data Processing

Data processing refers to the cleaning and screening procedures used to identify errors (such as out-of-range data) and the decisions made to exclude student cases or to suppress particular items in analyses. NYSED and the data repository were provided with the results of the checking, and some edits to the initial data were made; however, Pearson Psychometric and Research performs data cleaning to the delivered data and excludes some student cases in order to obtain a sample of the utmost integrity. It should be noted that the major groups of cases excluded from the data set were students from nonpublic schools and students with incorrect or incomplete grade information. Other deleted cases included duplicate record cases and no-response 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 ELA Grade 3 Data Cleaning

Exclusion Rule	# Deleted	# Cases Remain
Initial Number of Cases	0	199,632
Wrong Subject	0	199,632
No Grade	30	199,602
Wrong Grade	109	199,493
Nonpublic School	6,046	193,447
No Response	1	193,446
Invalid Score	0	193,446
Out of Range CR Scores	0	193,446
Duplicated Record	10	193,436

Table 4B. NYSTP ELA Grade 4 Data Cleaning

Exclusion Rule	# Deleted	# Cases Remain
Initial Number of Cases	0	206,953
Wrong Subject	0	206,953
No Grade	38	206,915
Wrong Grade	110	206,805
Nonpublic School	16,393	190,412
No Response	4	190,408
Invalid Score	0	190,408
Out of Range CR Scores	0	190,408
Duplicated Record	6	190,402

Table 4C. NYSTP ELA Grade 5 Data Cleaning

Exclusion Rule	# Deleted	# Cases Remain
Initial Number of Cases	0	198,823
Wrong Subject	0	198,823
No Grade	265	198,558
Wrong Grade	43	198,515
Nonpublic School	6,053	192,462
No Response	3	192,459
Invalid Score	0	192,459
Out of Range CR Scores	0	192,459
Duplicated Record	6	192,453

Table 4D. NYSTP ELA Grade 6 Data Cleaning

Exclusion Rule	# Deleted	# Cases Remain
Initial Number of Cases	0	209,313
Wrong Subject	0	209,313
No Grade	67	209,246
Wrong Grade	127	209,119
Nonpublic School	13,599	195,520
No Response	3	195,517
Invalid Score	0	195,517
Out of Range CR Scores	0	195,517
Duplicated Record	0	195,517

Table 4E. NYSTP ELA Grade 7 Data Cleaning

Exclusion Rule	# Deleted	# Cases Remain
Initial Number of Cases	0	199,832
Wrong Subject	0	199,832
No Grade	34	199,798
Wrong Grade	180	199,618
Nonpublic School	5,938	193,680
No Response	0	193,680
Invalid Score	0	193,680
Out of Range CR Scores	0	193,680
Duplicated Record	2	193,678

Table 4F. NYSTP ELA Grade 8 Data Cleaning

Exclusion Rule	# Deleted	# Cases Remain
Initial Number of Cases	0	208,330
Wrong Subject	0	208,330
No Grade	42	208,288
Wrong Grade	122	208,166
Nonpublic School	16,010	192,156
No Response	0	192,156
Invalid Score	0	192,156
Out of Range CR Scores	0	192,156
Duplicated Record	6	192,150

Classical Analysis and Calibration Sample Characteristics

The demographic characteristics of students in the cleaned calibration and equating data sets are presented in the proceeding tables. The clean data sets included over 95% of New York State students and were used for classical analyses presented in the calibrations in this section. 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 = 193,436)

Demographic Category		N-count	% of Total N-count
NRC	NYC	71,611	37.02
	Big 4 Cities	7,998	4.13
	Urban/Suburban	12,327	6.37
	Rural	11,016	5.69
	Average Needs	57,042	29.49
	Low Needs	27,566	14.25
	Charter	5,876	3.04
Ethnicity	Asian	16,318	8.44
	Black	34,103	17.63
	Hispanic	45,763	23.66
	American Indian	1,081	0.56
	Multiracial	1,940	1.00
	Other	399	0.21
	White	93,832	48.51
ELL	No	176,988	91.50
	Yes	16,448	8.50
SWD	No	166,086	85.86
	Yes	27,350	14.14
SUA	No	176,988	91.50
	Yes	16,448	8.50

Table 5B. Grade 4 Sample Characteristics (N = 190,402)

Demographic Category		N-count	% of Total N-count
NRC	NYC	69,618	36.56
	Big 4 Cities	7,981	4.19
	Urban/Suburban	12,610	6.62
	Rural	11,026	5.79
	Average Needs	56,849	29.86
	Low Needs	27,541	14.46
	Charter	4,777	2.51
Ethnicity	Asian	16,001	8.40
	Black	34,265	18.00
	Hispanic	44,433	23.34
	American Indian	1,031	0.54
	Multiracial	1,643	0.86
	Other	314	0.16
	White	92,715	48.69
ELL	No	174,961	91.89
	Yes	15,441	8.11
SWD	No	161,536	84.84
	Yes	28,866	15.16
SUA	No	165,627	86.99
	Yes	24,775	13.01

Table 5C. Grade 5 Sample Characteristics (N = 192,453)

Demographic Category		N-count	% of Total N-count
NRC	NYC	68,557	35.62
	Big 4 Cities	7,953	4.13
	Urban/Suburban	12,360	6.42
	Rural	11,383	5.91
	Average Needs	58,016	30.15
	Low Needs	28,289	14.70
	Charter	5,895	3.06
Ethnicity	Asian	15,676	8.15
	Black	35,257	18.32
	Hispanic	44,093	22.91
	American Indian	984	0.51
	Multiracial	1,483	0.77
	Other	312	0.16
	White	94,648	49.18
ELL	No	179,298	93.16
	Yes	13,155	6.84
SWD	No	162,525	84.45
	Yes	29,928	15.55
SUA	No	167,024	86.79
	Yes	25,429	13.21

Table 5D. Grade 6 Sample Characteristics (N = 195,517)

Demographic Category		N-count	% of Total N-count
NRC	NYC	68,957	35.27
	Big 4 Cities	7,746	3.96
	Urban/Suburban	12,136	6.21
	Rural	11,406	5.83
	Average Needs	59,912	30.64
	Low Needs	29,862	15.27
	Charter	5,498	2.81
Ethnicity	Asian	16,567	8.47
	Black	35,724	18.27
	Hispanic	43,540	22.27
	American Indian	998	0.51
	Multiracial	1,409	0.72
	Other	347	0.18
	White	96,932	49.58
ELL	No	184,500	94.37
	Yes	11,017	5.63
SWD	No	165,846	84.82
	Yes	29,671	15.18
SUA	No	171,130	87.53
	Yes	24,387	12.47

Table 5E. Grade 7 Sample Characteristics (N = 193,678)

Demographic Category		N-count	% of Total N-count
NRC	NYC	67,415	34.81
	Big 4 Cities	7,590	3.92
	Urban/Suburban	12,099	6.25
	Rural	11,703	6.04
	Average Needs	59,072	30.50
	Low Needs	31,184	16.10
	Charter	4,615	2.38
Ethnicity	Asian	15,422	7.96
	Black	35,673	18.42
	Hispanic	42,482	21.93
	American Indian	1,004	0.52
	Multiracial	1,349	0.70
	Other	318	0.16
	White	97,430	50.31
ELL	No	183,137	94.56
	Yes	10,541	5.44
SWD	No	164,535	84.95
	Yes	29,143	15.05
SUA	No	169,876	87.71
	Yes	23,802	12.29

Table 5F. Grade 8 Sample Characteristics (N = 192,150)

Demographic Category		N-count	% of Total N-count
NRC	NYC	68,865	35.84
	Big 4 Cities	7,263	3.78
	Urban/Suburban	11,624	6.05
	Rural	11,243	5.85
	Average Needs	58,415	30.40
	Low Needs	31,361	16.32
	Charter	3,379	1.76
Ethnicity	Asian	15,644	8.14
	Black	35,491	18.47
	Hispanic	42,138	21.93
	American Indian	1,018	0.53
	Multiracial	1,096	0.57
	Other	329	0.17
	White	96,434	50.19
ELL	No	181,697	94.56
	Yes	10,453	5.44
SWD	No	163,579	85.13
	Yes	28,571	14.87
SUA	No	168,989	87.95
	Yes	23,161	12.05

Classical Data Analysis

Classical data analysis of the NYSTP Grades 3–8 ELA Tests consists of four primary elements. One element is the analysis of item-level statistical information about student performance. It is important to verify that the items and test forms function as intended. Information on item difficulty (p-value) and item-test correlation (point biserial) is examined thoroughly. If any serious error were to occur with an item (i.e., a printing error or potentially correct distractor), item analysis is the stage at which errors should be flagged and evaluated for rectification (suppression, credit, or other acceptable solution). Analyses of test-level data comprise the second element of classical data analysis. These include examination of the raw score statistics (mean and standard deviation) and test reliability measures (Cronbach's alpha and Feldt-Raju coefficient). Assessment of test speededness is another important element of classical analysis. Additionally, classical differential item functioning (DIF) analysis is conducted at this stage. DIF analysis includes computation of standardized mean differences and Mantel-Haenszel statistics for New York State items to identify potential item bias. All classical data analysis results contribute information on the validity and reliability of the tests (also see Sections III, "Validity," and VII, "Reliability and Standard Error of Measurement").

Item 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 to 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 ELA Tests ranged from 0.36 to 0.97. For Grade 3, the item p-values were between 0.52 and 0.97, with a mean of 0.74. For Grade 4, the item p-values were between 0.48 and 0.93, with a mean of 0.72. For Grade 5, the item p-values were between 0.42 and 0.93, with a mean of 0.73. For Grade 6, the item p-values were between 0.37 and 0.97, with a mean of 0.70. For Grade 7, the item p-values were between 0.41 and 0.93, with a mean of 0.75. For Grade 8, the item p-values were between 0.36 and 0.97, with a mean of 0.74. These p-value statistics are also provided in Tables 6A–6F, along with point biserial statistics of the key.

Table 6A. Item Analysis, Grade 3

Item	Item Type	N-count	P-value	% Omit	PbisKey
01	MC	193,386	0.93	0.03	0.49
02	MC	193,360	0.80	0.04	0.43
03	MC	193,345	0.95	0.05	0.44

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

Item	Item Type	N-count	P-value	% Omit	PbisKey
04	MC	193,310	0.92	0.07	0.37
05	MC	193,337	0.80	0.05	0.26
06	MC	193,278	0.75	0.08	0.52
07	MC	193,256	0.68	0.09	0.45
08	MC	193,281	0.81	0.08	0.41
09	MC	193,219	0.68	0.11	0.48
10	MC	193,267	0.77	0.09	0.54
11	MC	193,307	0.85	0.07	0.45
12	MC	193,263	0.69	0.09	0.36
13	MC	193,174	0.72	0.14	0.50
14	MC	193,204	0.61	0.12	0.45
15	MC	193,217	0.66	0.11	0.43
16	MC	192,987	0.60	0.23	0.50
17	MC	192,940	0.64	0.26	0.37
18	MC	192,933	0.88	0.26	0.43
19	MC	192,854	0.62	0.30	0.44
20	MC	192,828	0.55	0.31	0.34
21	MC	192,759	0.71	0.35	0.57
22	MC	192,668	0.87	0.40	0.48
23	MC	192,601	0.56	0.43	0.38
24	MC	192,533	0.84	0.47	0.40
25	MC	192,590	0.69	0.44	0.29
26	MC	192,314	0.86	0.58	0.49
27	MC	192,155	0.77	0.66	0.56
28	MC	191,982	0.69	0.75	0.55
29	MC	191,939	0.82	0.77	0.45
30	MC	191,534	0.65	0.98	0.32
31	MC	193,377	0.97	0.03	0.22
32	MC	193,343	0.95	0.05	0.32

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

Item	Item Type	N-count	P-value	% Omit	PbisKey
33	MC	193,310	0.87	0.07	0.40
34	MC	193,246	0.80	0.10	0.36
35	MC	193,174	0.75	0.14	0.25
36	CR	192,918	0.81	0.27	
37	CR	192,859	0.62	0.30	
38	CR	192,685	0.78	0.39	
39	CR	192,356	0.65	0.56	
40	MC	193,260	0.85	0.09	0.34
41	MC	193,228	0.84	0.11	0.39
42	MC	193,143	0.75	0.15	0.37
43	MC	193,120	0.79	0.16	0.48
44	MC	193,147	0.77	0.15	0.36
45	MC	193,167	0.64	0.14	0.46
46	MC	193,044	0.60	0.20	0.43
47	MC	193,028	0.73	0.21	0.51
48	MC	193,046	0.87	0.20	0.58
49	MC	193,025	0.74	0.21	0.50
50	MC	192,913	0.52	0.27	0.41
51	MC	192,625	0.57	0.42	0.39
52	CR	193,130	0.76	0.16	
53	CR	192,857	0.67	0.30	
54	CR	192,467	0.54	0.50	
55	CR	192,383	0.75	0.54	
56	CR	191,857	0.52	0.82	

Table 6B. Item Analysis, Grade 4

Item	Item Type	N-count	P-value	% Omit	PbisKey
01	MC	190,360	0.80	0.02	0.46
02	MC	190,369	0.93	0.02	0.35
03	MC	190,295	0.73	0.06	0.45
04	MC	190,319	0.80	0.04	0.50
05	MC	190,340	0.81	0.03	0.43
06	MC	190,330	0.92	0.04	0.38
07	MC	190,290	0.72	0.06	0.33
08	MC	190,310	0.80	0.05	0.26
09	MC	190,292	0.67	0.06	0.23
10	MC	190,332	0.80	0.04	0.42
11	MC	190,287	0.62	0.06	0.52
12	MC	190,203	0.51	0.10	0.39
13	MC	190,295	0.78	0.06	0.30
14	MC	190,246	0.61	0.08	0.33
15	MC	190,271	0.58	0.07	0.38
16	MC	190,180	0.57	0.12	0.32
17	MC	190,131	0.72	0.14	0.51
18	MC	190,153	0.65	0.13	0.38
19	MC	190,091	0.63	0.16	0.51
20	MC	190,125	0.71	0.15	0.45
21	MC	189,974	0.85	0.22	0.49
22	MC	189,919	0.77	0.25	0.53
23	MC	189,876	0.73	0.28	0.42
24	MC	189,832	0.86	0.30	0.47
25	MC	189,858	0.78	0.29	0.52
26	MC	189,746	0.56	0.34	0.34
27	MC	189,627	0.85	0.41	0.55
28	MC	189,565	0.50	0.44	0.33

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

Item	Item Type	N-count	P-value	% Omit	PbisKey
29	MC	189,483	0.84	0.48	0.46
30	MC	189,379	0.61	0.54	0.25
31	MC	189,130	0.66	0.67	0.51
32	MC	190,347	0.75	0.03	0.30
33	MC	190,354	0.82	0.03	0.38
34	MC	190,297	0.68	0.06	0.37
35	MC	190,262	0.56	0.07	0.32
36	MC	190,172	0.67	0.12	0.32
37	CR	190,190	0.82	0.11	
38	CR	190,049	0.67	0.19	
39	CR	190,037	0.71	0.19	
40	CR	189,865	0.64	0.28	
41	MC	190,282	0.64	0.06	0.29
42	MC	190,261	0.85	0.07	0.43
43	MC	190,210	0.74	0.10	0.36
44	MC	190,192	0.77	0.11	0.38
45	MC	190,185	0.48	0.11	0.24
46	MC	190,186	0.57	0.11	0.42
47	MC	190,160	0.78	0.13	0.52
48	MC	190,163	0.74	0.13	0.51
49	MC	190,086	0.75	0.17	0.40
50	MC	190,086	0.71	0.17	0.45
51	MC	190,024	0.79	0.20	0.56
52	MC	190,027	0.81	0.20	0.56
53	MC	189,976	0.77	0.22	0.43
54	MC	189,894	0.62	0.27	0.38
55	MC	189,727	0.67	0.35	0.38
56	CR	189,837	0.69	0.30	
57	CR	189,830	0.75	0.30	

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

Item	Item Type	N-count	P-value	% Omit	PbisKey
58	CR	190,130	0.80	0.14	
59	CR	190,045	0.81	0.19	
60	CR	189,870	0.65	0.28	

Table 6C. Item Analysis, Grade 5

Item	Item Type	N-count	P-value	% Omit	PbisKey
01	MC	192,434	0.79	0.01	0.34
02	MC	192,388	0.64	0.03	0.47
03	MC	192,407	0.86	0.02	0.46
04	MC	192,384	0.73	0.04	0.34
05	MC	192,379	0.67	0.04	0.48
06	MC	192,376	0.73	0.04	0.42
07	MC	192,392	0.86	0.03	0.37
08	MC	192,329	0.59	0.06	0.38
09	MC	192,337	0.70	0.06	0.38
10	MC	192,361	0.61	0.05	0.36
11	MC	192,359	0.88	0.05	0.36
12	MC	192,311	0.48	0.07	0.23
13	MC	192,300	0.79	0.08	0.59
14	MC	192,341	0.72	0.06	0.44
15	MC	192,310	0.72	0.07	0.34
16	MC	192,349	0.67	0.05	0.23
17	MC	192,243	0.86	0.11	0.48
18	MC	192,244	0.88	0.11	0.52
19	MC	192,198	0.63	0.13	0.27
20	MC	192,201	0.85	0.13	0.49
21	MC	192,100	0.47	0.18	0.26
22	MC	192,148	0.83	0.16	0.48

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

Item	Item Type	N-count	P-value	% Omit	PbisKey
23	MC	192,085	0.88	0.19	0.40
24	MC	192,091	0.87	0.19	0.37
25	MC	192,047	0.42	0.21	0.30
26	MC	192,044	0.72	0.21	0.40
27	MC	191,987	0.53	0.24	0.21
28	MC	191,723	0.80	0.38	0.37
29	MC	191,657	0.57	0.41	0.31
30	MC	191,619	0.70	0.43	0.29
31	MC	191,551	0.64	0.47	0.47
32	MC	191,452	0.69	0.52	0.38
33	MC	191,302	0.64	0.60	0.48
34	MC	192,399	0.76	0.03	0.29
35	MC	192,388	0.86	0.03	0.43
36	MC	192,366	0.82	0.05	0.30
37	MC	192,332	0.85	0.06	0.38
38	MC	192,144	0.54	0.16	0.33
39	CR	192,189	0.77	0.14	
40	CR	192,081	0.67	0.19	
41	CR	191,817	0.76	0.33	
42	CR	191,899	0.66	0.29	
43	MC	192,350	0.86	0.05	0.40
44	MC	192,318	0.70	0.07	0.41
45	MC	192,302	0.70	0.08	0.35
46	MC	192,307	0.79	0.08	0.37
47	MC	192,299	0.93	0.08	0.45
48	MC	192,315	0.87	0.07	0.42
49	MC	192,311	0.72	0.07	0.46
50	MC	192,250	0.93	0.11	0.44
51	MC	192,197	0.81	0.13	0.22

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

Item	Item Type	N-count	P-value	% Omit	PbisKey
52	MC	192,197	0.70	0.13	0.44
53	MC	192,207	0.73	0.13	0.43
54	MC	192,175	0.79	0.14	0.39
55	MC	192,014	0.61	0.23	0.39
56	CR	192,260	0.89	0.10	
57	CR	191,993	0.65	0.24	
58	CR	192,113	0.63	0.18	
59	CR	191,968	0.69	0.25	
60	CR	191,818	0.57	0.33	

Table 6D. Item Analysis, Grade 6

Item	Item Type	N-count	P-value	% Omit	PbisKey
01	MC	195,477	0.91	0.02	0.39
02	MC	195,493	0.93	0.01	0.36
03	MC	195,395	0.64	0.06	0.32
04	MC	195,390	0.57	0.06	0.36
05	MC	195,451	0.62	0.03	0.39
06	MC	195,461	0.87	0.03	0.43
07	MC	195,463	0.85	0.03	0.46
08	MC	195,415	0.58	0.05	0.37
09	MC	195,418	0.73	0.05	0.39
10	MC	195,386	0.49	0.07	0.25
11	MC	195,454	0.58	0.03	0.38
12	MC	195,272	0.50	0.13	0.10
13	MC	195,437	0.97	0.04	0.28
14	MC	195,385	0.57	0.07	0.44
15	MC	195,441	0.94	0.04	0.34
16	MC	195,358	0.45	0.08	0.37

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

Item	Item Type	N-count	P-value	% Omit	PbisKey
17	MC	195,353	0.55	0.08	0.37
18	MC	195,321	0.81	0.10	0.45
19	MC	195,337	0.88	0.09	0.49
20	MC	195,322	0.69	0.10	0.35
21	MC	195,195	0.43	0.16	0.32
22	MC	195,295	0.77	0.11	0.43
23	MC	195,224	0.67	0.15	0.44
24	MC	195,043	0.70	0.24	0.24
25	MC	195,231	0.54	0.15	0.38
26	MC	195,080	0.40	0.22	0.40
27	MC	195,106	0.50	0.21	0.40
28	MC	195,081	0.49	0.22	0.37
29	MC	194,950	0.89	0.29	0.41
30	MC	194,810	0.50	0.36	0.28
31	MC	194,844	0.76	0.34	0.34
32	MC	194,782	0.67	0.38	0.40
33	MC	194,664	0.56	0.44	0.35
34	MC	195,425	0.53	0.05	0.33
35	MC	195,401	0.82	0.06	0.40
36	MC	195,446	0.95	0.04	0.22
37	MC	195,328	0.80	0.10	0.25
38	MC	195,358	0.85	0.08	0.27
39	CR	195,162	0.71	0.18	
40	CR	194,664	0.82	0.44	
41	CR	194,982	0.75	0.27	
42	CR	194,810	0.68	0.36	
43	MC	195,199	0.55	0.16	0.39
44	MC	195,338	0.74	0.09	0.37
45	MC	195,344	0.89	0.09	0.48

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

Item	Item Type	N-count	P-value	% Omit	PbisKey
46	MC	195,325	0.49	0.10	0.48
47	MC	195,345	0.53	0.09	0.40
48	MC	195,323	0.66	0.10	0.41
49	MC	195,273	0.87	0.12	0.45
50	MC	195,137	0.79	0.19	0.52
51	MC	195,275	0.85	0.12	0.49
52	MC	195,225	0.37	0.15	0.34
53	MC	195,244	0.69	0.14	0.44
54	MC	195,225	0.65	0.15	0.39
55	MC	195,149	0.82	0.19	0.45
56	CR	195,325	0.86	0.10	
57	CR	195,148	0.77	0.19	
58	CR	195,009	0.94	0.26	
59	CR	194,829	0.77	0.35	
60	CR	195,053	0.69	0.24	

Table 6E. Item Analysis, Grade 7

Item	Item Type	N-count	P-value	% Omit	PbisKey
01	MC	193,641	0.80	0.02	0.35
02	MC	193,628	0.84	0.03	0.30
03	MC	193,561	0.61	0.06	0.48
04	MC	193,604	0.84	0.04	0.50
05	MC	193,596	0.75	0.04	0.37
06	MC	193,568	0.91	0.06	0.31
07	MC	193,620	0.82	0.03	0.43
08	MC	193,572	0.83	0.05	0.42
09	MC	193,533	0.67	0.07	0.43
10	MC	193,608	0.87	0.04	0.41
11	MC	193,475	0.63	0.10	0.47

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

Item	Item Type	N-count	P-value	% Omit	PbisKey
12	MC	193,587	0.57	0.05	0.36
13	MC	193,479	0.70	0.10	0.38
14	MC	193,523	0.70	0.08	0.44
15	MC	193,569	0.68	0.06	0.44
16	MC	193,578	0.67	0.05	0.27
17	MC	193,529	0.75	0.08	0.45
18	MC	193,502	0.71	0.09	0.32
19	MC	193,537	0.85	0.07	0.52
20	MC	193,479	0.49	0.10	0.37
21	MC	193,468	0.63	0.11	0.34
22	MC	193,493	0.83	0.10	0.47
23	MC	193,428	0.59	0.13	0.48
24	MC	193,456	0.91	0.11	0.50
25	MC	193,449	0.88	0.12	0.51
26	MC	193,415	0.78	0.14	0.34
27	MC	193,352	0.75	0.17	0.54
28	MC	193,284	0.70	0.20	0.46
29	MC	193,256	0.70	0.22	0.42
30	MC	193,239	0.80	0.23	0.48
31	MC	193,265	0.91	0.21	0.51
32	MC	193,013	0.80	0.34	0.49
33	MC	192,972	0.76	0.36	0.48
34	MC	193,600	0.93	0.04	0.29
35	MC	193,563	0.78	0.06	0.28
36	MC	193,539	0.78	0.07	0.34
37	MC	193,499	0.68	0.09	0.43
38	MC	193,326	0.61	0.18	0.29
39	CR	193,270	0.87	0.21	
40	CR	192,895	0.86	0.40	

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

Item	Item Type	N-count	P-value	% Omit	PbisKey
41	CR	193,085	0.84	0.31	
42	CR	192,819	0.72	0.44	
43	MC	193,339	0.61	0.18	0.24
44	MC	193,503	0.88	0.09	0.32
45	MC	193,491	0.57	0.10	0.25
46	MC	193,487	0.87	0.10	0.31
47	MC	193,497	0.77	0.09	0.41
48	MC	193,443	0.41	0.12	0.20
49	MC	193,455	0.44	0.12	0.32
50	MC	193,432	0.77	0.13	0.50
51	MC	193,430	0.81	0.13	0.40
52	MC	193,352	0.47	0.17	0.28
53	MC	193,404	0.90	0.14	0.50
54	MC	193,370	0.88	0.16	0.49
55	MC	193,314	0.80	0.19	0.50
56	CR	193,361	0.78	0.16	
57	CR	193,207	0.80	0.24	
58	CR	193,353	0.80	0.17	
59	CR	193,006	0.77	0.35	
60	CR	192,765	0.69	0.47	

Table 6F. Item Analysis, Grade 8

Item	Item Type	N-count	P-value	% Omit	PbisKey
01	MC	192,080	0.61	0.04	0.21
02	MC	191,934	0.43	0.11	0.19
03	MC	192,089	0.82	0.03	0.35
04	MC	192,076	0.81	0.04	0.32
05	MC	192,115	0.97	0.02	0.29
06	MC	191,952	0.41	0.10	0.20

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

Item	Item Type	N-count	P-value	% Omit	PbisKey
07	MC	191,963	0.81	0.10	0.43
08	MC	192,081	0.95	0.04	0.35
09	MC	192,033	0.65	0.06	0.22
10	MC	192,011	0.72	0.07	0.42
11	MC	191,965	0.62	0.10	0.45
12	MC	192,010	0.76	0.07	0.50
13	MC	191,998	0.44	0.08	0.27
14	MC	191,970	0.70	0.09	0.47
15	MC	192,002	0.71	0.08	0.40
16	MC	191,926	0.68	0.12	0.45
17	MC	191,952	0.85	0.10	0.41
18	MC	191,879	0.64	0.14	0.39
19	MC	191,961	0.81	0.10	0.48
20	MC	191,866	0.83	0.15	0.35
21	MC	191,872	0.92	0.14	0.39
22	MC	191,870	0.66	0.15	0.23
23	MC	191,778	0.72	0.19	0.30
24	MC	191,734	0.67	0.22	0.46
25	MC	191,730	0.72	0.22	0.41
26	MC	191,729	0.77	0.22	0.50
27	MC	191,621	0.66	0.28	0.41
28	MC	191,884	0.81	0.14	0.23
29	MC	192,022	0.90	0.07	0.25
30	MC	191,848	0.81	0.16	0.33
31	MC	191,998	0.86	0.08	0.35
32	MC	191,900	0.82	0.13	0.28
33	CR	191,797	0.92	0.18	
34	CR	191,585	0.91	0.29	
35	CR	191,536	0.90	0.32	

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

Item	Item Type	N-count	P-value	% Omit	PbisKey
36	CR	191,141	0.74	0.53	
37	MC	191,986	0.85	0.09	0.36
38	MC	191,952	0.84	0.10	0.47
39	MC	191,976	0.82	0.09	0.42
40	MC	191,932	0.81	0.11	0.51
41	MC	191,946	0.71	0.11	0.52
42	MC	191,976	0.72	0.09	0.49
43	MC	191,834	0.72	0.16	0.54
44	MC	191,899	0.57	0.13	0.37
45	MC	191,858	0.64	0.15	0.47
46	MC	191,801	0.57	0.18	0.27
47	MC	191,879	0.36	0.14	0.32
48	MC	191,580	0.71	0.30	0.49
49	MC	191,571	0.69	0.30	0.45
50	CR	191,886	0.84	0.14	
51	CR	191,690	0.85	0.24	
52	CR	191,876	0.90	0.14	
53	CR	191,584	0.84	0.29	
54	CR	191,500	0.72	0.34	

Point-Biserial Correlation Coefficients

Point-biserial (pbis) statistics are used to examine item-test correlations or item discrimination for MC items. In the Tables 6A–6F, point-biserial correlation coefficients were computed for the answer key and reported in the Pbis Key field. The point-biserial correlation is a measure of internal consistency that ranges between +/–1. It indicates a correlation of students' responses to an item relative to their performance on the rest of the test. The criterion for point biserial for the correct answer option used for NYSTP 3–8 ELA Tests was 0.20. The point biserials for the correct answer option that was equal to or greater than 0.20 indicated that students who responded correctly also tended to do well on the overall test. The only items that had a low point biserial were operational item number 12 in the Grade 6 test and operational item number 2 in the Grade 8 test. Point biserials for correct answer options on the tests ranged from 0.10–0.59. For Grade 3, the pbis were between 0.22 and 0.58. For Grade 4, the pbis were between 0.23 and 0.56. For Grade 5, the pbis were between 0.21 and 0.59. For Grade 6, pbis were between 0.10

and 0.52. For Grade 7, the pbis were between 0.20 and 0.54. For Grade 8, the pbis were between 0.19 and 0.54.

Test Statistics and Reliability Coefficients

Test statistics including raw-score mean and raw-score 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 coefficient, were computed for the Grades 3–8 ELA Tests. Both types of reliability estimates are appropriate to use when a test contains both MC and CR items. The calculated Cronbach's alpha reliabilities and Feldt-Raju reliability coefficients both ranged from 0.90–0.92. All reliabilities met or exceeded 0.90, across statistics, which is a good indication that the NYSTP 3–8 ELA Tests are acceptably reliable. High reliability indicates that scores are consistent and not unduly influenced by random errors. For more information on test reliability and standard error of measurement, see Section VII, "Reliability and Standard Error of Measurement."

Table 7. NYSTP ELA 2012 Test Form Statistics and Reliability

Grade	Max RS	RS Mean	RS SD	P-value Mean	Cronbach's Alpha	Feldt-Raju
3	67	48.59	11.83	0.74	0.92	0.92
4	73	52.07	12.02	0.72	0.92	0.92
5	73	52.36	11.99	0.73	0.91	0.92
6	73	51.46	11.57	0.70	0.91	0.91
7	73	54.65	11.74	0.75	0.92	0.92
8	67	50.59	10.07	0.74	0.90	0.90

Speededness

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

Differential Item Functioning

Classical differential item functioning (DIF) was evaluated using two methods. First, the standardized mean difference (SMD) was computed for all items. The SMD statistic (Dorans, Schmitt, and Bleistein, 1992) compares the mean scores of reference and focal groups, after adjusting for ability differences. A moderate amount of significant DIF, for or against the focal group, is represented by an SMD with an absolute value between 0.10 and 0.19, inclusive. A

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

Classical DIF analyses were conducted on subgroups of the Needs/Resource Capacity Category (focal group: High Needs; reference group: Low Needs), gender (focal group: Female; reference group: Male), ethnicity (focal groups: Black, Hispanic, and Asian; reference group: White), and English language learners (focal group: English language learners; reference group: Non-English language learners). The DIF analyses were conducted using all cases from the clean data sets. Table 8 shows the numbers of cases for subgroups.

Table 8. NYSTP ELA 2012 Classical DIF Sample N-Counts

Grade	Ethnicity				Gender		Needs/Resource Capacity Category		English Language Learner Status	
	Black/African American	Hispanic/ Latino	Asian	White	Female	Male	High	Low	Yes	No
3	30,274	44,302	16,218	93,400	90,012	94,182	101,277	82,917	16,105	168,089
4	31,045	43,302	15,945	92,389	89,407	93,274	99,757	82,924	15,180	167,501
5	31,392	42,536	15,614	94,299	90,061	93,780	98,904	84,937	12,810	171,031
6	32,335	41,929	16,480	96,569	91,470	95,843	98,871	88,442	10,734	176,579
7	32,809	41,179	15,355	97,080	90,864	95,559	97,490	88,933	10,298	176,125
8	33,347	41,259	15,578	96,165	91,519	94,830	97,771	88,578	10,283	176,066

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	6
4	5
5	8
6	9
7	10
8	7

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

Section VI: IRT Scaling and Equating

IRT Models and Rationale for Use

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

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

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

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

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

where

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

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

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

where

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

and

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

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

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

and

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

where

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

and

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

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

Calibration Sample

The cleaned sample data were used for calibration and scaling of NYSTP ELA Tests. It should be noted that the scaling was done on nearly all (96–99%, depending on grade level) of the New York State public school student population in each tested grade and that exclusion of some cases during the data cleaning process had minimal effect on parameter estimation. As shown in Tables 10 through 12, the 2012 OP samples were comparable to 2011 populations in terms of 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	36.27	37.02	35.78	36.56
Big 4 Cities	4.17	4.13	4.20	4.19
Urban/Suburban	7.90	6.37	7.50	6.62
Rural	5.73	5.69	5.87	5.79
Average Needs	29.56	29.49	30.13	29.86
Low Needs	13.87	14.25	14.56	14.46
Charter	2.51	3.04	1.95	2.51
ETHNICITY				
Asian	8.10	8.44	7.90	8.40
Black	18.60	17.63	18.59	18.00
Hispanics	23.29	23.66	22.63	23.34
American Indian	0.56	0.56	0.48	0.54
Multiracial	0.80	1.00	0.70	0.86
White	48.52	48.51	49.57	48.69
Other	0.14	0.21	0.13	0.16
ELL STATUS				
No	90.75	91.50	91.81	91.89
Yes	9.25	8.50	8.19	8.11
DISABILITY				
No	85.80	85.86	84.90	84.84
Yes	14.20	14.14	15.10	15.16
ACCOMMODATIONS				
No	75.21	88.25	75.02	86.99
Yes	24.79	11.75	24.98	13.01

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	34.90	35.62	34.63	35.27
Big 4 Cities	3.97	4.13	3.95	3.96
Urban/Suburban	7.09	6.42	6.99	6.21
Rural	5.81	5.91	5.77	5.83
Average Needs	30.47	30.15	30.86	30.64
Low Needs	15.15	14.70	15.35	15.27
Charter	2.61	3.06	2.43	2.81
ETHNICITY				
Asian	8.36	8.15	7.75	8.47
Black	18.49	18.32	18.84	18.27
Hispanics	21.92	22.91	21.82	22.27
American Indian	0.49	0.51	0.47	0.51
Multiracial	0.66	0.77	0.63	0.72
White	49.97	49.18	50.36	49.58
Other	0.12	0.16	0.13	0.18
ELL STATUS				
No	93.13	93.16	93.97	94.37
Yes	6.87	6.84	6.03	5.63
DISABILITY				
No	84.72	84.45	84.73	84.82
Yes	15.28	15.55	15.27	15.18
ACCOMMODATIONS				
No	75.60	86.79	77.57	87.53
Yes	24.40	13.21	22.43	12.47

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	34.61	34.81	35.30	35.84
Big 4 Cities	3.79	3.92	3.76	3.78
Urban/Suburban	7.06	6.25	6.66	6.05
Rural	5.78	6.04	5.70	5.85
Average Needs	30.83	30.50	30.87	30.40
Low Needs	16.09	16.10	16.30	16.32
Charter	1.84	2.38	1.41	1.76
ETHNICITY				
Asian	7.65	7.96	7.93	8.14
Black	18.87	18.42	18.74	18.47
Hispanics	21.40	21.93	21.05	21.93
American Indian	0.49	0.52	0.49	0.53
Multiracial	0.55	0.70	0.48	0.57
White	50.90	50.31	51.18	50.19
Other	0.13	0.16	0.12	0.17
ELL STATUS				
No	94.61	94.56	94.62	94.56
Yes	5.39	5.44	5.38	5.44
DISABILITY				
No	84.71	84.95	85.16	85.13
Yes	15.29	15.05	14.84	14.87
ACCOMMODATIONS				
No	78.55	87.71	79.14	87.95
Yes	21.45	12.29	20.86	12.05

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 ELA 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 ELA 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	2.784	–2.875	4.801	–0.01	0.931	193,436
4	2.584	–3.347	3.150	–0.00	0.933	190,402
5	2.404	–4.146	3.817	–0.01	0.935	192,453
6	2.210	–3.761	4.699	–0.00	0.941	195,517
7	2.813	–4.360	3.808	–0.01	0.931	193,678
8	2.388	–6.011	4.031	–0.01	0.922	192,150

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. The fact that all items in the NYSTP 2011 ELA Tests demonstrated good model fit further supports the use of the chosen models. Item fit statistics are presented in Tables 14–19.

Table 14. ELA Grade 3 Item Fit Statistics

Item	Model	Chi Square	DF	*Z*-observed	*Z*-critical	Fit OK?
1	3PL	60.86	7	14.40	180.54	Y
2	3PL	122.98	7	31.00	180.54	Y
3	3PL	235.78	7	61.15	180.54	Y
4	3PL	62.34	7	14.79	180.54	Y
5	3PL	34.99	7	7.48	180.54	Y
6	3PL	116.58	7	29.29	180.54	Y

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

Item	Model	Chi Square	DF	Z-observed	Z-critical	Fit OK?
7	3PL	163.89	7	41.93	180.54	Y
8	3PL	58.08	7	13.65	180.54	Y
9	3PL	130.45	7	32.99	180.54	Y
10	3PL	108.90	7	27.23	180.54	Y
11	3PL	202.29	7	52.19	180.54	Y
12	3PL	152.51	7	38.89	180.54	Y
13	3PL	162.23	7	41.49	180.54	Y
14	3PL	258.54	7	67.23	180.54	Y
15	3PL	203.74	7	52.58	180.54	Y
16	3PL	349.79	7	91.62	180.54	Y
17	3PL	167.28	7	42.84	180.54	Y
18	3PL	48.42	7	11.07	180.54	Y
19	3PL	221.94	7	57.45	180.54	Y
20	3PL	257.15	7	66.86	180.54	Y
21	3PL	202.19	7	52.17	180.54	Y
22	3PL	89.77	7	22.12	180.54	Y
23	3PL	167.79	7	42.97	180.54	Y
24	3PL	101.65	7	25.30	180.54	Y
25	3PL	60.39	7	14.27	180.54	Y
26	3PL	81.58	7	19.93	180.54	Y
27	3PL	217.97	7	56.39	180.54	Y
28	3PL	248.69	7	64.59	180.54	Y
29	3PL	452.26	7	119.00	180.54	Y
30	3PL	78.56	7	19.13	180.54	Y
31	3PL	60.08	7	14.19	180.54	Y
32	3PL	39.14	7	8.59	180.54	Y
33	3PL	62.83	7	14.92	180.54	Y
34	3PL	109.66	7	27.44	180.54	Y
35	3PL	70.32	7	16.92	180.54	Y
36	2PPC	147.76	16	23.29	180.54	Y
37	2PPC	391.94	16	66.46	180.54	Y
38	2PPC	313.73	16	52.63	180.54	Y
39	2PPC	369.42	25	48.71	180.54	Y
40	3PL	108.58	7	27.15	180.54	Y
41	3PL	73.11	7	17.67	180.54	Y
42	3PL	93.86	7	23.21	180.54	Y
43	3PL	128.95	7	32.59	180.54	Y
44	3PL	115.08	7	28.88	180.54	Y
45	3PL	146.60	7	37.31	180.54	Y
46	3PL	172.91	7	44.34	180.54	Y
47	3PL	155.52	7	39.69	180.54	Y

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

Item	Model	Chi Square	DF	Z-observed	Z-critical	Fit OK?
48	3PL	131.34	7	33.23	180.54	Y
49	3PL	178.61	7	45.87	180.54	Y
50	3PL	236.62	7	61.37	180.54	Y
51	3PL	205.32	7	53.00	180.54	Y
52	2PPC	284.12	16	47.40	180.54	Y
53	2PPC	326.03	16	54.81	180.54	Y
54	2PPC	653.53	25	88.89	180.54	Y
55	2PPC	370.96	16	62.75	180.54	Y
56	2PPC	388.69	16	65.88	180.54	Y

Table 15. ELA Grade 4 Item Fit Statistics

Item	Model	Chi Square	DF	Z-observed	Z-critical	Fit OK?
1	3PL	125.66	7	31.71	177.71	Y
2	3PL	39.83	7	8.77	177.71	Y
3	3PL	168.47	7	43.16	177.71	Y
4	3PL	160.55	7	41.04	177.71	Y
5	3PL	99.13	7	24.62	177.71	Y
6	3PL	83.19	7	20.36	177.71	Y
7	3PL	94.28	7	23.33	177.71	Y
8	3PL	97.64	7	24.22	177.71	Y
9	3PL	94.27	7	23.32	177.71	Y
10	3PL	69.99	7	16.83	177.71	Y
11	3PL	242.08	7	62.83	177.71	Y
12	3PL	313.25	7	81.85	177.71	Y
13	3PL	97.53	7	24.20	177.71	Y
14	3PL	111.18	7	27.84	177.71	Y
15	3PL	667.41	7	176.50	177.71	Y
16	3PL	119.52	7	30.07	177.71	Y
17	3PL	108.84	7	27.22	177.71	Y
18	3PL	111.23	7	27.86	177.71	Y
19	3PL	163.03	7	41.70	177.71	Y
20	3PL	145.75	7	37.08	177.71	Y
21	3PL	91.35	7	22.54	177.71	Y
22	3PL	149.28	7	38.03	177.71	Y
23	3PL	110.79	7	27.74	177.71	Y
24	3PL	89.84	7	22.14	177.71	Y
25	3PL	120.82	7	30.42	177.71	Y
26	3PL	161.15	7	41.20	177.71	Y
27	3PL	106.72	7	26.65	177.71	Y
28	3PL	545.70	7	143.97	177.71	Y
29	3PL	154.35	7	39.38	177.71	Y

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

Item	Model	Chi Square	DF	Z-observed	Z-critical	Fit OK?
30	3PL	48.54	7	11.10	177.71	Y
31	3PL	178.33	7	45.79	177.71	Y
32	3PL	46.13	7	10.46	177.71	Y
33	3PL	66.20	7	15.82	177.71	Y
34	3PL	144.23	7	36.68	177.71	Y
35	3PL	147.89	7	37.65	177.71	Y
36	3PL	134.12	7	33.97	177.71	Y
37	2PP	268.16	16	44.58	177.71	Y
38	2PP	291.73	16	48.74	177.71	Y
39	2PP	937.52	16	162.90	177.71	Y
40	2PP	492.78	34	55.64	177.71	Y
41	3PL	136.84	7	34.70	177.71	Y
42	3PL	66.83	7	15.99	177.71	Y
43	3PL	85.32	7	20.93	177.71	Y
44	3PL	139.66	7	35.45	177.71	Y
45	3PL	214.26	7	55.39	177.71	Y
46	3PL	112.45	7	28.18	177.71	Y
47	3PL	121.56	7	30.62	177.71	Y
48	3PL	135.15	7	34.25	177.71	Y
49	3PL	98.37	7	24.42	177.71	Y
50	3PL	149.32	7	38.04	177.71	Y
51	3PL	136.54	7	34.62	177.71	Y
52	3PL	124.56	7	31.42	177.71	Y
53	3PL	122.88	7	30.97	177.71	Y
54	3PL	94.22	7	23.31	177.71	Y
55	3PL	121.09	7	30.49	177.71	Y
56	2PP	281.33	16	46.90	177.71	Y
57	2PP	274.27	16	45.66	177.71	Y
58	2PP	672.64	16	116.08	177.71	Y
59	2PP	287.45	16	47.99	177.71	Y
60	2PP	592.50	34	67.73	177.71	Y

Table 16. ELA Grade 5 Item Fit Statistics

Item	Model	Chi Square	DF	Z-observed	Z-critical	Fit OK?
1	3PL	150.97	7	38.48	179.62	Y
2	3PL	122.52	7	30.87	179.62	Y
3	3PL	94.47	7	23.38	179.62	Y
4	3PL	320.51	7	83.79	179.62	Y
5	3PL	171.57	7	43.98	179.62	Y
6	3PL	95.03	7	23.53	179.62	Y
7	3PL	66.73	7	15.96	179.62	Y

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

Item	Model	Chi Square	DF	Z-observed	Z-critical	Fit OK?
8	3PL	148.22	7	37.74	179.62	Y
9	3PL	109.90	7	27.50	179.62	Y
10	3PL	113.78	7	28.54	179.62	Y
11	3PL	40.38	7	8.92	179.62	Y
12	3PL	38.32	7	8.37	179.62	Y
13	3PL	148.09	7	37.71	179.62	Y
14	3PL	165.59	7	42.39	179.62	Y
15	3PL	110.81	7	27.74	179.62	Y
16	3PL	294.62	7	76.87	179.62	Y
17	3PL	70.54	7	16.98	179.62	Y
18	3PL	115.42	7	28.98	179.62	Y
19	3PL	52.30	7	12.11	179.62	Y
20	3PL	116.17	7	29.18	179.62	Y
21	3PL	232.83	7	60.36	179.62	Y
22	3PL	85.98	7	21.11	179.62	Y
23	3PL	58.44	7	13.75	179.62	Y
24	3PL	62.04	7	14.71	179.62	Y
25	3PL	88.54	7	21.79	179.62	Y
26	3PL	107.79	7	26.94	179.62	Y
27	3PL	107.05	7	26.74	179.62	Y
28	3PL	251.68	7	65.39	179.62	Y
29	3PL	109.41	7	27.37	179.62	Y
30	3PL	95.74	7	23.72	179.62	Y
31	3PL	225.95	7	58.52	179.62	Y
32	3PL	100.33	7	24.94	179.62	Y
33	3PL	209.87	7	54.22	179.62	Y
34	3PL	75.13	7	18.21	179.62	Y
35	3PL	104.28	7	26.00	179.62	Y
36	3PL	61.12	7	14.47	179.62	Y
37	3PL	48.87	7	11.19	179.62	Y
38	3PL	108.33	7	27.08	179.62	Y
39	2PP	234.86	16	38.69	179.62	Y
40	2PP	460.38	16	78.56	179.62	Y
41	2PP	317.52	16	53.30	179.62	Y
42	2PP	812.47	34	94.40	179.62	Y
43	3PL	73.89	7	17.88	179.62	Y
44	3PL	128.50	7	32.47	179.62	Y
45	3PL	65.69	7	15.69	179.62	Y
46	3PL	69.91	7	16.81	179.62	Y
47	3PL	108.22	7	27.05	179.62	Y
48	3PL	129.66	7	32.78	179.62	Y
49	3PL	123.76	7	31.20	179.62	Y

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

Item	Model	Chi Square	DF	*Z*-observed	*Z*-critical	Fit OK?
50	3PL	95.46	7	23.64	179.62	Y
51	3PL	195.18	7	50.29	179.62	Y
52	3PL	126.50	7	31.94	179.62	Y
53	3PL	99.22	7	24.65	179.62	Y
54	3PL	75.27	7	18.25	179.62	Y
55	3PL	82.48	7	20.17	179.62	Y
56	2PP	159.18	16	25.31	179.62	Y
57	2PP	198.25	16	32.22	179.62	Y
58	2PP	281.93	16	47.01	179.62	Y
59	2PP	317.58	16	53.31	179.62	Y
60	2PP	820.38	34	95.36	179.62	Y

Table 17. ELA Grade 6 Item Fit Statistics

Item	Model	Chi Square	DF	*Z*-observed	*Z*-critical	Fit OK?
1	3PL	62.05	7	14.71	182.48	Y
2	3PL	59.70	7	14.08	182.48	Y
3	3PL	127.66	7	32.25	182.48	Y
4	3PL	95.49	7	23.65	182.48	Y
5	3PL	100.78	7	25.06	182.48	Y
6	3PL	83.35	7	20.41	182.48	Y
7	3PL	89.02	7	21.92	182.48	Y
8	3PL	131.63	7	33.31	182.48	Y
9	3PL	137.94	7	35.00	182.48	Y
10	3PL	146.55	7	37.30	182.48	Y
11	3PL	209.95	7	54.24	182.48	Y
12	3PL	36.17	7	7.80	182.48	Y
13	3PL	74.94	7	18.16	182.48	Y
14	3PL	183.15	7	47.08	182.48	Y
15	3PL	34.09	7	7.24	182.48	Y
16	3PL	178.60	7	45.86	182.48	Y
17	3PL	110.29	7	27.61	182.48	Y
18	3PL	82.44	7	20.16	182.48	Y
19	3PL	105.76	7	26.40	182.48	Y
20	3PL	120.81	7	30.42	182.48	Y
21	3PL	209.53	7	54.13	182.48	Y
22	3PL	106.71	7	26.65	182.48	Y
23	3PL	200.21	7	51.64	182.48	Y
24	3PL	90.36	7	22.28	182.48	Y
25	3PL	232.50	7	60.27	182.48	Y
26	3PL	358.64	7	93.98	182.48	Y
27	3PL	177.67	7	45.61	182.48	Y

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

Item	Model	Chi Square	DF	Z-observed	Z-critical	Fit OK?
28	3PL	168.77	7	43.24	182.48	Y
29	3PL	77.35	7	18.80	182.48	Y
30	3PL	89.13	7	21.95	182.48	Y
31	3PL	124.49	7	31.40	182.48	Y
32	3PL	102.83	7	25.61	182.48	Y
33	3PL	99.00	7	24.59	182.48	Y
34	3PL	121.51	7	30.60	182.48	Y
35	3PL	147.82	7	37.64	182.48	Y
36	3PL	53.84	7	12.52	182.48	Y
37	3PL	52.83	7	12.25	182.48	Y
38	3PL	152.30	7	38.83	182.48	Y
39	2PP	326.40	16	54.87	182.48	Y
40	2PP	503.50	16	86.18	182.48	Y
41	2PP	298.48	16	49.94	182.48	Y
42	2PP	512.14	34	57.98	182.48	Y
43	3PL	171.98	7	44.09	182.48	Y
44	3PL	56.68	7	13.28	182.48	Y
45	3PL	83.91	7	20.55	182.48	Y
46	3PL	340.54	7	89.14	182.48	Y
47	3PL	216.69	7	56.04	182.48	Y
48	3PL	142.13	7	36.11	182.48	Y
49	3PL	93.48	7	23.11	182.48	Y
50	3PL	133.55	7	33.82	182.48	Y
51	3PL	159.43	7	40.74	182.48	Y
52	3PL	229.77	7	59.54	182.48	Y
53	3PL	145.48	7	37.01	182.48	Y
54	3PL	169.38	7	43.40	182.48	Y
55	3PL	66.73	7	15.96	182.48	Y
56	2PP	279.48	16	46.58	182.48	Y
57	2PP	238.65	16	39.36	182.48	Y
58	2PP	108.93	16	16.43	182.48	Y
59	2PP	216.64	16	35.47	182.48	Y
60	2PP	617.25	34	70.73	182.48	Y

Table 18. ELA Grade 7 Item Fit Statistics

Item	Model	Chi Square	DF	Z-observed	Z-critical	Fit OK?
1	3PL	100.94	7	25.11	180.77	Y
2	3PL	75.87	7	18.41	180.77	Y
3	3PL	206.57	7	53.34	180.77	Y
4	3PL	148.02	7	37.69	180.77	Y
5	3PL	98.82	7	24.54	180.77	Y

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

Item	Model	Chi Square	DF	Z-observed	Z-critical	Fit OK?
6	3PL	36.45	7	7.87	180.77	Y
7	3PL	100.47	7	24.98	180.77	Y
8	3PL	119.29	7	30.01	180.77	Y
9	3PL	160.04	7	40.90	180.77	Y
10	3PL	65.56	7	15.65	180.77	Y
11	3PL	233.54	7	60.54	180.77	Y
12	3PL	168.12	7	43.06	180.77	Y
13	3PL	118.12	7	29.70	180.77	Y
14	3PL	97.08	7	24.07	180.77	Y
15	3PL	181.80	7	46.72	180.77	Y
16	3PL	57.61	7	13.53	180.77	Y
17	3PL	159.78	7	40.83	180.77	Y
18	3PL	324.95	7	84.98	180.77	Y
19	3PL	83.90	7	20.55	180.77	Y
20	3PL	201.82	7	52.07	180.77	Y
21	3PL	130.67	7	33.05	180.77	Y
22	3PL	109.73	7	27.46	180.77	Y
23	3PL	242.58	7	62.96	180.77	Y
24	3PL	88.26	7	21.72	180.77	Y
25	3PL	77.95	7	18.96	180.77	Y
26	3PL	179.87	7	46.20	180.77	Y
27	3PL	95.62	7	23.68	180.77	Y
28	3PL	150.13	7	38.25	180.77	Y
29	3PL	168.60	7	43.19	180.77	Y
30	3PL	99.13	7	24.62	180.77	Y
31	3PL	82.37	7	20.14	180.77	Y
32	3PL	225.16	7	58.31	180.77	Y
33	3PL	122.59	7	30.89	180.77	Y
34	3PL	34.26	7	7.29	180.77	Y
35	3PL	164.56	7	42.11	180.77	Y
36	3PL	116.59	7	29.29	180.77	Y
37	3PL	108.74	7	27.19	180.77	Y
38	3PL	134.84	7	34.17	180.77	Y
39	2PP	148.12	16	23.36	180.77	Y
40	2PP	169.83	16	27.19	180.77	Y
41	2PP	160.40	16	25.53	180.77	Y
42	2PP	645.00	34	74.09	180.77	Y
43	3PL	53.91	7	12.54	180.77	Y
44	3PL	59.58	7	14.05	180.77	Y
45	3PL	51.89	7	12.00	180.77	Y
46	3PL	150.74	7	38.42	180.77	Y
47	3PL	145.36	7	36.98	180.77	Y

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

Item	Model	Chi Square	DF	Z-observed	Z-critical	Fit OK?
48	3PL	67.51	7	16.17	180.77	Y
49	3PL	385.49	7	101.16	180.77	Y
50	3PL	152.82	7	38.97	180.77	Y
51	3PL	91.41	7	22.56	180.77	Y
52	3PL	186.73	7	48.03	180.77	Y
53	3PL	83.81	7	20.53	180.77	Y
54	3PL	75.83	7	18.40	180.77	Y
55	3PL	99.89	7	24.83	180.77	Y
56	2PP	185.40	16	29.95	180.77	Y
57	2PP	245.29	16	40.53	180.77	Y
58	2PP	318.03	16	53.39	180.77	Y
59	2PP	237.80	16	39.21	180.77	Y
60	2PP	661.84	34	76.14	180.77	Y

Table 19. ELA Grade 8 Item Fit Statistics

Item	Model	Chi Square	DF	Z-observed	Z-critical	Fit OK?
1	3PL	157.15	7	40.13	179.34	Y
2	3PL	109.91	7	27.50	179.34	Y
3	3PL	96.70	7	23.97	179.34	Y
4	3PL	291.52	7	76.04	179.34	Y
5	3PL	41.43	7	9.20	179.34	Y
6	3PL	128.60	7	32.50	179.34	Y
7	3PL	197.12	7	50.81	179.34	Y
8	3PL	24.70	7	4.73	179.34	Y
9	3PL	530.53	7	139.92	179.34	Y
10	3PL	116.51	7	29.27	179.34	Y
11	3PL	365.13	7	95.71	179.34	Y
12	3PL	240.39	7	62.38	179.34	Y
13	3PL	164.26	7	42.03	179.34	Y
14	3PL	211.02	7	54.53	179.34	Y
15	3PL	182.49	7	46.90	179.34	Y
16	3PL	186.18	7	47.89	179.34	Y
17	3PL	118.69	7	29.85	179.34	Y
18	3PL	171.71	7	44.02	179.34	Y
19	3PL	166.38	7	42.60	179.34	Y
20	3PL	88.50	7	21.78	179.34	Y
21	3PL	81.84	7	20.00	179.34	Y
22	3PL	104.29	7	26.00	179.34	Y
23	3PL	118.92	7	29.91	179.34	Y
24	3PL	190.65	7	49.08	179.34	Y

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

Item	Model	Chi Square	DF	Z-observed	Z-critical	Fit OK?
25	3PL	138.41	7	35.12	179.34	Y
26	3PL	209.09	7	54.01	179.34	Y
27	3PL	205.61	7	53.08	179.34	Y
28	3PL	292.94	7	76.42	179.34	Y
29	3PL	90.77	7	22.39	179.34	Y
30	3PL	82.18	7	20.09	179.34	Y
31	3PL	54.22	7	12.62	179.34	Y
32	3PL	156.22	7	39.88	179.34	Y
33	2PP	138.51	16	21.66	179.34	Y
34	2PP	108.04	16	16.27	179.34	Y
35	2PP	199.02	16	32.35	179.34	Y
36	2PP	764.66	34	88.61	179.34	Y
37	3PL	171.78	7	44.04	179.34	Y
38	3PL	97.40	7	24.16	179.34	Y
39	3PL	128.70	7	32.52	179.34	Y
40	3PL	168.41	7	43.14	179.34	Y
41	3PL	253.59	7	65.90	179.34	Y
42	3PL	230.71	7	59.79	179.34	Y
43	3PL	227.45	7	58.92	179.34	Y
44	3PL	175.20	7	44.95	179.34	Y
45	3PL	351.66	7	92.11	179.34	Y
46	3PL	187.32	7	48.19	179.34	Y
47	3PL	565.84	7	149.36	179.34	Y
48	3PL	313.47	7	81.91	179.34	Y
49	3PL	219.94	7	56.91	179.34	Y
50	2PP	261.58	16	43.41	179.34	Y
51	2PP	447.27	16	76.24	179.34	Y
52	2PP	383.16	16	64.90	179.34	Y
53	2PP	309.18	16	51.83	179.34	Y
54	2PP	741.87	34	85.84	179.34	Y

Local Independence

In using IRT models, one of the assumptions made is that the items are locally independent, that a student's response on one item is not dependent upon his or her response to another item. In other words, when a student's ability is accounted for, his or her response to each item is statistically independent.

One way to measure the statistical independence of items within a test is via the Q_3 statistic (Yen, 1984). This statistic was obtained by correlating differences between students' observed

and expected responses for pairs of items after taking into account overall test performance. The Q_3 statistic for binary items was computed as

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

and

$$Q_{3\,jj'} = r\left(d_j, d_{j'}\right).$$

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. When item pairs are flagged by Q_3, the content of the flagged items is examined to identify possible sources of the local dependence. The primary concern about locally dependent items is that they contribute less psychometric information about examinee proficiency than do locally independent items and they inflate score reliability estimates.

The Q_3 statistics were examined on all ELA Tests and no items were found to be significant in terms of local dependency in Grades 4, 5 and 8. In Grade 3, two pairs of items were found to be significant in local dependency: items 6 and 10 (Q_3 = 0.250) and item 6 and 45 (Q_3 = 0.231). In Grade 6, one pair of items was found to be significant in local dependency: items 59 and 60 (Q_3 = 0.206). In Grade 7, one pair of items was found to be significant in local dependency: items 59 and 60 (Q_3 = 0.222). The magnitudes of these statistics were not sufficient to warrant any concern. Anchor items were excluded from Q_3 computation.

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 grade 3 to grade 6 and fifteen percent of the students from the field test samples for grades 7 and 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 using the 2011 field test data were compared with the item parameters calibrated using the 2012 OP data. The equating of 2012 OP data to the NY 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. The final M1 and M2 were obtained after the final anchor set is determined. Table 20 presents the 2012 OP transformation constants for NYSTP Grades 3–8 ELA Tests.

Table 20. NYSTP ELA 2012 Final Transformation Constants

Grade	*M1*	*M2*
3	18.68	664.27
4	26.87	674.38
5	16.69	669.68
6	14.74	663.04
7	16.12	665.12
8	18.59	657.69

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 were used to evaluate the appropriateness of using an item in the 3PL or 2PPC model. Poor-fit items were 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 five iterations:

    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 (M1and M2) with the remaining anchor items (save as v1-v5 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 two items being removed from the anchor set for ELA grades 4 and 5 and one item being removed from the anchor set for grade 6 to grade 8.

Item Parameters

The OP test item parameters were estimated by the software MULTILOG (Thissen, 1991) and are presented in Table 21 through Table 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. 2012 Operational Item Parameter Estimates, Grade 3

Item	Max Pts	a-par/alpha	b-par/gamma1	c-par/gamma2	gamma3
01	1	0.069	628.893	0.064	
02	1	0.047	647.740	0.276	
03	1	0.070	624.041	0.060	
04	1	0.038	618.920	0.012	
05	1	0.019	618.207	0.021	
06	1	0.055	651.162	0.160	
07	1	0.042	655.903	0.160	
08	1	0.035	637.032	0.081	
09	1	0.051	657.167	0.174	
10	1	0.063	650.682	0.175	
11	1	0.041	632.402	0.014	
12	1	0.030	653.133	0.187	

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

Item	Max Pts	a-par/alpha	b-par/gamma1	c-par/gamma2	gamma3
13	1	0.057	656.136	0.217	
14	1	0.061	666.230	0.257	
15	1	0.058	664.420	0.305	
16	1	0.054	662.570	0.130	
17	1	0.030	657.431	0.147	
18	1	0.050	637.519	0.269	
19	1	0.048	663.747	0.215	
20	1	0.047	674.038	0.290	
21	1	0.058	652.564	0.064	
22	1	0.056	638.977	0.199	
23	1	0.034	666.540	0.149	
24	1	0.035	633.012	0.093	
25	1	0.021	649.074	0.166	
26	1	0.061	643.969	0.279	
27	1	0.088	654.306	0.252	
28	1	0.074	658.252	0.189	
29	1	0.038	635.399	0.007	
30	1	0.026	658.299	0.199	
31	1	0.035	592.821	0.022	
32	1	0.039	611.417	0.038	
33	1	0.043	637.755	0.295	
34	1	0.027	629.834	0.008	
35	1	0.017	625.149	0.027	
36	2	0.046	28.010	29.668	
37	2	0.040	25.124	26.738	
38	2	0.055	33.562	35.788	
39	3	0.055	33.817	35.875	37.510
40	1	0.028	622.311	0.009	

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

Item	Max Pts	a-par/alpha	b-par/gamma1	c-par/gamma2	gamma3
41	1	0.033	629.536	0.015	
42	1	0.029	641.280	0.082	
43	1	0.055	650.343	0.269	
44	1	0.027	634.607	0.028	
45	1	0.049	660.994	0.190	
46	1	0.044	664.345	0.192	
47	1	0.053	653.274	0.165	
48	1	0.086	643.003	0.183	
49	1	0.058	654.267	0.215	
50	1	0.050	671.517	0.196	
51	1	0.043	669.182	0.224	
52	2	0.053	33.021	34.741	
53	2	0.057	35.196	38.132	
54	3	0.054	33.654	35.549	36.731
55	2	0.064	41.559	41.069	
56	2	0.051	33.258	34.447	

Table 22. 2012 Operational Item Parameter Estimates, Grade 4

Item	Max Pts	a-par/alpha	b-par/gamma1	c-par/ gamma2	gamma3	gamma4
01	1	0.033	648.167	0.195		
02	1	0.027	606.970	0.028		
03	1	0.043	666.593	0.333		
04	1	0.045	654.977	0.287		
05	1	0.027	640.313	0.127		
06	1	0.030	616.780	0.116		
07	1	0.016	638.693	0.024		
08	1	0.013	605.992	0.011		
09	1	0.012	659.283	0.239		
10	1	0.025	637.911	0.089		

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

Item	Max Pts	a-par/alpha	b-par/gamma1	c-par/ gamma2	gamma3	gamma4
11	1	0.044	670.850	0.149		
12	1	0.036	687.739	0.218		
13	1	0.018	647.313	0.301		
14	1	0.015	655.724	0.008		
15	1	0.017	662.964	0.003		
16	1	0.021	681.370	0.218		
17	1	0.043	662.584	0.212		
18	1	0.020	657.819	0.062		
19	1	0.036	666.896	0.099		
20	1	0.031	659.259	0.184		
21	1	0.047	648.350	0.311		
22	1	0.040	653.498	0.146		
23	1	0.028	655.958	0.198		
24	1	0.038	639.799	0.213		
25	1	0.041	651.861	0.164		
26	1	0.025	683.739	0.230		
27	1	0.053	646.365	0.205		
28	1	0.053	694.175	0.308		
29	1	0.030	636.225	0.099		
30	1	0.012	667.493	0.158		
31	1	0.042	667.951	0.174		
32	1	0.016	641.030	0.183		
33	1	0.026	644.151	0.286		
34	1	0.026	670.105	0.291		
35	1	0.025	688.174	0.278		
36	1	0.021	668.034	0.271		
37	2	0.022	12.366	13.676		
38	2	0.029	17.085	20.211		

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

Item	Max Pts	a-par/alpha	b-par/gamma1	c-par/ gamma2	gamma3	gamma4
39	2	0.024	12.643	16.236		
40	4	0.038	22.347	23.497	25.189	27.471
41	1	0.019	674.665	0.283		
42	1	0.031	637.869	0.194		
43	1	0.027	664.292	0.354		
44	1	0.024	648.488	0.229		
45	1	0.033	705.894	0.336		
46	1	0.025	673.064	0.102		
47	1	0.041	652.772	0.169		
48	1	0.037	656.562	0.156		
49	1	0.023	645.201	0.099		
50	1	0.031	659.767	0.190		
51	1	0.053	654.961	0.204		
52	1	0.057	652.383	0.212		
53	1	0.031	655.024	0.258		
54	1	0.022	665.212	0.118		
55	1	0.022	661.383	0.158		
56	2	0.029	17.795	19.897		
57	2	0.045	28.035	29.725		
58	2	0.032	18.055	20.620		
59	2	0.038	23.079	24.461		
60	4	0.041	24.059	25.546	27.036	29.085

Table 23. 2012 Operational Item Parameter Estimates, Grade 5

Item	Max Pts	a-par/ alpha	b-par/ gamma1	c-par/ gamma2	gamma3	gamma4
01	1	0.062	667.117	0.538		
02	1	0.048	663.816	0.103		

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

Item	Max Pts	a-par/ alpha	b-par/ gamma1	c-par/ gamma2	gamma3	gamma4
03	1	0.061	649.390	0.265		
04	1	0.026	644.916	0.006		
05	1	0.059	664.394	0.195		
06	1	0.046	659.318	0.221		
07	1	0.037	635.496	0.018		
08	1	0.039	669.820	0.182		
09	1	0.035	658.361	0.156		
10	1	0.034	667.729	0.184		
11	1	0.040	637.832	0.192		
12	1	0.017	677.542	0.061		
13	1	0.085	656.231	0.138		
14	1	0.046	658.907	0.169		
15	1	0.033	657.819	0.220		
16	1	0.017	648.353	0.071		
17	1	0.069	649.958	0.264		
18	1	0.077	646.670	0.153		
19	1	0.018	653.181	0.018		
20	1	0.059	646.299	0.112		
21	1	0.043	688.193	0.281		
22	1	0.058	651.226	0.197		
23	1	0.045	636.950	0.054		
24	1	0.038	635.885	0.045		
25	1	0.028	684.074	0.101		
26	1	0.033	650.905	0.014		
27	1	0.013	664.562	0.010		
28	1	0.031	640.650	0.011		
29	1	0.026	667.985	0.103		
30	1	0.023	652.215	0.130		
31	1	0.046	662.655	0.085		

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

Item	Max Pts	a-par/ alpha	b-par/ gamma1	c-par/ gamma2	gamma3	gamma4
32	1	0.030	652.239	0.021		
33	1	0.064	668.070	0.214		
34	1	0.022	638.082	0.013		
35	1	0.070	654.836	0.434		
36	1	0.037	655.854	0.449		
37	1	0.044	646.002	0.266		
38	1	0.024	666.857	0.029		
39	2	0.053	33.128	35.070		
40	2	0.059	37.935	39.435		
41	2	0.063	40.096	41.837		
42	4	0.088	54.480	56.198	58.156	60.396
43	1	0.042	637.463	0.010		
44	1	0.034	653.661	0.022		
45	1	0.032	657.633	0.164		
46	1	0.038	651.660	0.241		
47	1	0.066	635.754	0.037		
48	1	0.056	646.713	0.279		
49	1	0.048	658.680	0.147		
50	1	0.061	635.435	0.042		
51	1	0.018	619.701	0.011		
52	1	0.045	660.374	0.168		
53	1	0.045	658.765	0.191		
54	1	0.035	644.804	0.050		
55	1	0.036	665.300	0.116		
56	2	0.063	39.322	40.434		
57	2	0.048	30.726	31.882		
58	2	0.056	35.621	37.566		
59	2	0.069	43.898	45.734		
60	4	0.076	48.671	49.868	51.394	52.763

Table 24. 2012 Operational Item Parameter Estimates, Grade 6

Item	Max Pts	a-par/ alpha	b-par/ gamma1	c-par/ gamma2	gamma3	gamma4
01	1	0.057	634.360	0.148		
02	1	0.060	632.574	0.227		
03	1	0.030	656.216	0.143		
04	1	0.038	663.295	0.148		
05	1	0.043	660.021	0.160		
06	1	0.058	637.642	0.064		
07	1	0.075	647.033	0.297		
08	1	0.036	660.527	0.100		
09	1	0.039	647.628	0.087		
10	1	0.047	680.730	0.324		
11	1	0.063	668.469	0.288		
12	1	0.008	722.939	0.271		
13	1	0.063	619.138	0.022		
14	1	0.060	664.903	0.181		
15	1	0.055	626.222	0.018		
16	1	0.058	673.476	0.186		
17	1	0.035	661.289	0.060		
18	1	0.054	643.786	0.096		
19	1	0.088	644.347	0.234		
20	1	0.030	645.985	0.021		
21	1	0.054	677.193	0.210		
22	1	0.049	647.541	0.147		
23	1	0.057	658.881	0.208		
24	1	0.023	654.184	0.274		
25	1	0.057	669.757	0.239		
26	1	0.070	674.502	0.147		
27	1	0.056	669.445	0.177		
28	1	0.053	671.700	0.192		
29	1	0.060	638.282	0.174		

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

Item	Max Pts	a-par/ alpha	b-par/ gamma1	c-par/ gamma2	gamma3	gamma4
30	1	0.030	673.166	0.187		
31	1	0.031	638.592	0.008		
32	1	0.051	660.011	0.258		
33	1	0.039	666.075	0.189		
34	1	0.033	667.098	0.132		
35	1	0.043	638.227	0.009		
36	1	0.036	609.339	0.027		
37	1	0.023	627.605	0.081		
38	1	0.028	622.780	0.010		
39	2	0.054	34.281	34.960		
40	2	0.059	37.544	37.460		
41	2	0.048	28.907	31.385		
42	4	0.081	49.910	51.469	53.163	54.965
43	1	0.045	665.074	0.145		
44	1	0.038	648.137	0.137		
45	1	0.088	643.077	0.241		
46	1	0.066	666.912	0.105		
47	1	0.054	667.997	0.184		
48	1	0.048	657.640	0.181		
49	1	0.084	648.308	0.381		
50	1	0.082	651.433	0.214		
51	1	0.071	644.122	0.138		
52	1	0.058	678.227	0.161		
53	1	0.051	655.213	0.150		
54	1	0.065	665.409	0.341		
55	1	0.060	646.542	0.184		
56	2	0.086	54.667	54.884		
57	2	0.082	52.801	53.587		
58	2	0.088	54.991	55.619		

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

Item	Max Pts	a-par/ alpha	b-par/ gamma1	c-par/ gamma2	gamma3	gamma4
59	2	0.052	32.392	33.543		
60	4	0.072	44.483	46.021	47.430	49.023

Table 25. 2012 Operational Item Parameter Estimates, Grade 7

Item	Max Pts	a-par/ alpha	b-par/ gamma1	c-par/ gamma2	gamma3	gamma4
01	1	0.042	651.389	0.354		
02	1	0.028	625.483	0.012		
03	1	0.062	663.630	0.153		
04	1	0.069	647.155	0.217		
05	1	0.032	643.292	0.034		
06	1	0.035	621.754	0.050		
07	1	0.050	646.584	0.216		
08	1	0.051	646.647	0.240		
09	1	0.055	661.458	0.233		
10	1	0.050	640.089	0.225		
11	1	0.060	662.611	0.179		
12	1	0.046	669.581	0.234		
13	1	0.040	655.558	0.194		
14	1	0.044	653.204	0.092		
15	1	0.049	658.596	0.190		
16	1	0.019	643.367	0.015		
17	1	0.059	656.141	0.276		
18	1	0.041	662.446	0.377		
19	1	0.075	646.659	0.193		
20	1	0.047	673.384	0.180		
21	1	0.044	667.469	0.296		
22	1	0.060	647.431	0.220		
23	1	0.064	664.486	0.151		

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

Item	Max Pts	a-par/ alpha	b-par/ gamma1	c-par/ gamma2	gamma3	gamma4
24	1	0.078	639.873	0.163		
25	1	0.079	644.790	0.231		
26	1	0.037	650.607	0.298		
27	1	0.072	654.421	0.156		
28	1	0.063	660.113	0.251		
29	1	0.061	661.467	0.306		
30	1	0.059	649.692	0.200		
31	1	0.103	643.298	0.297		
32	1	0.058	648.747	0.138		
33	1	0.054	650.787	0.119		
34	1	0.035	616.842	0.038		
35	1	0.022	628.552	0.012		
36	1	0.040	653.758	0.363		
37	1	0.048	658.821	0.186		
38	1	0.031	667.876	0.269		
39	2	0.055	32.957	35.007		
40	2	0.056	34.999	35.717		
41	2	0.056	34.274	35.721		
42	4	0.080	48.793	50.502	52.339	54.188
43	1	0.017	650.066	0.034		
44	1	0.034	625.862	0.068		
45	1	0.018	657.190	0.036		
46	1	0.029	623.487	0.010		
47	1	0.037	642.605	0.011		
48	1	0.019	692.970	0.164		
49	1	0.055	679.274	0.221		
50	1	0.064	652.859	0.185		
51	1	0.038	638.838	0.020		

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

Item	Max Pts	a-par/ alpha	b-par/ gamma1	c-par/ gamma2	gamma3	gamma4
52	1	0.020	669.766	0.011		
53	1	0.089	644.562	0.297		
54	1	0.070	642.315	0.194		
55	1	0.062	649.349	0.163		
56	2	0.066	40.659	43.391		
57	2	0.079	49.863	51.352		
58	2	0.073	45.000	48.003		
59	2	0.072	44.219	47.301		
60	4	0.078	48.375	49.501	51.193	52.835

Table 26. 2012 Operational Item Parameter Estimates, Grade 8

Item	Max Pts	a-par/ alpha	b-par/ gamma1	c-par/ gamma2	gamma3	gamma4
01	1	0.019	669.834	0.327		
02	1	0.027	692.136	0.289		
03	1	0.029	622.141	0.023		
04	1	0.023	617.275	0.008		
05	1	0.048	608.020	0.328		
06	1	0.026	691.578	0.244		
07	1	0.048	641.285	0.288		
08	1	0.049	609.557	0.048		
09	1	0.011	624.390	0.008		
10	1	0.039	645.960	0.193		
11	1	0.059	659.082	0.255		
12	1	0.065	647.392	0.255		
13	1	0.018	670.639	0.051		
14	1	0.048	648.362	0.178		
15	1	0.040	649.276	0.240		
16	1	0.041	648.569	0.138		

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

Item	Max Pts	a-par/ alpha	b-par/ gamma1	c-par/ gamma2	gamma3	gamma4
17	1	0.041	630.254	0.200		
18	1	0.045	658.583	0.292		
19	1	0.052	639.197	0.206		
20	1	0.028	619.797	0.015		
21	1	0.045	618.730	0.127		
22	1	0.014	632.790	0.054		
23	1	0.021	630.016	0.047		
24	1	0.046	650.760	0.173		
25	1	0.033	639.985	0.083		
26	1	0.066	647.048	0.277		
27	1	0.049	656.844	0.288		
28	1	0.015	597.606	0.012		
29	1	0.032	631.258	0.580		
30	1	0.026	620.811	0.020		
31	1	0.032	621.316	0.127		
32	1	0.020	610.257	0.009		
33	2	0.067	37.849	41.723		
34	2	0.074	43.338	46.404		
35	2	0.060	35.814	37.499		
36	4	0.067	39.885	41.160	43.460	44.906
37	1	0.037	632.269	0.293		
38	1	0.052	635.528	0.207		
39	1	0.041	634.463	0.182		
40	1	0.060	640.471	0.191		
41	1	0.062	649.082	0.181		
42	1	0.053	648.068	0.191		
43	1	0.076	650.378	0.228		
44	1	0.034	659.639	0.172		

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

Item	Max Pts	a-par/ alpha	b-par/ gamma1	c-par/ gamma2	gamma3	gamma4
45	1	0.065	656.889	0.254		
46	1	0.021	662.092	0.186		
47	1	0.053	677.347	0.172		
48	1	0.046	644.835	0.102		
49	1	0.046	649.831	0.185		
50	2	0.037	23.063	23.061		
51	2	0.042	26.013	26.280		
52	2	0.051	30.958	31.791		
53	2	0.081	49.345	51.821		
54	4	0.065	38.661	39.807	42.131	43.583

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 and CSEM curves are presented in Figure 1 though 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.

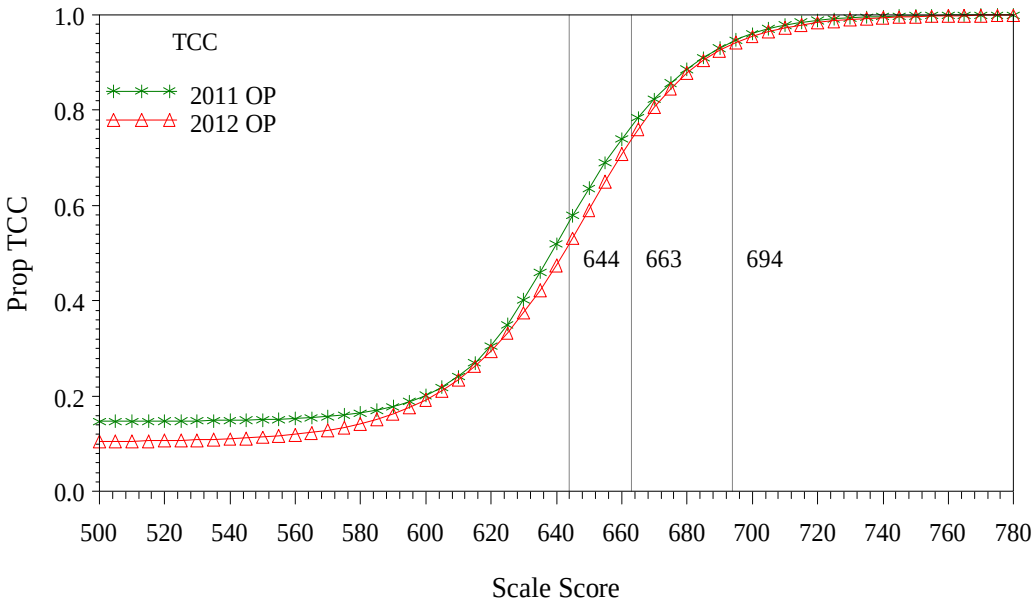

Figure 1. Grade 3 2011 and 2012 OP TCCs

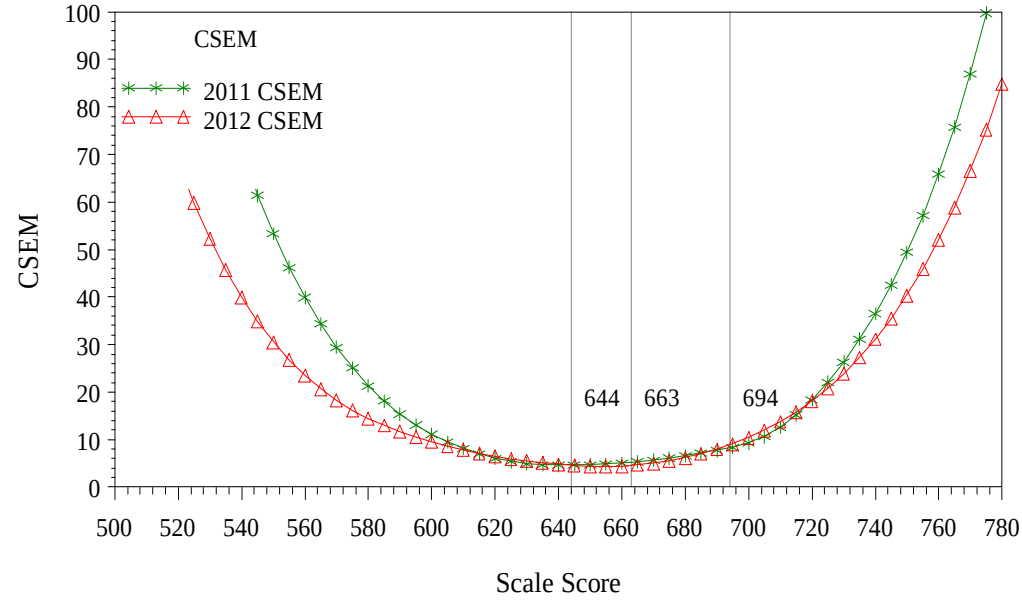

Figure 2. Grade 3 2011 and 2012 CSEM Curves

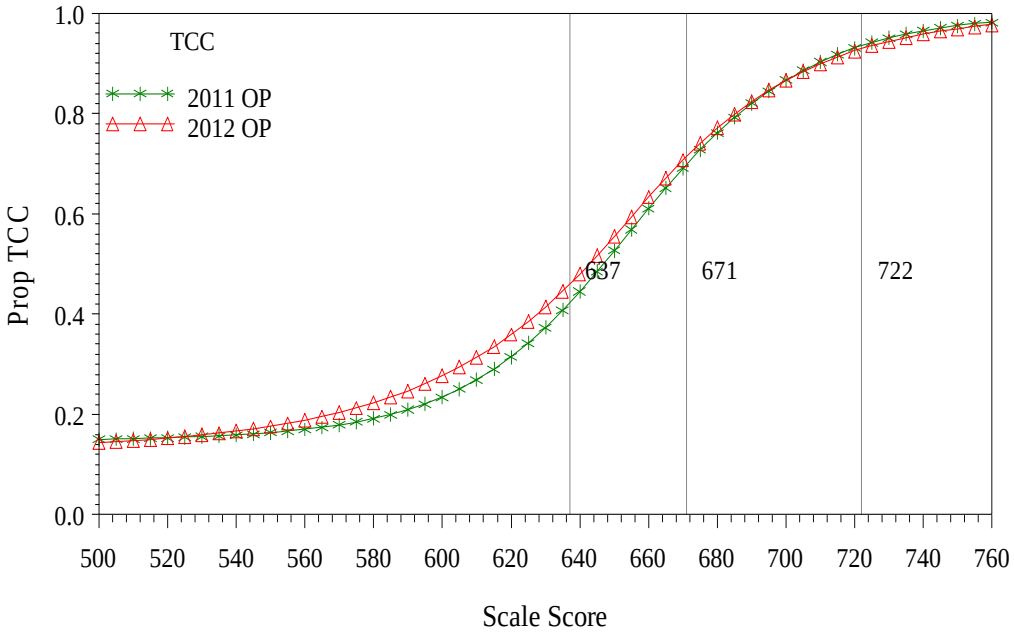

Figure 3. Grade 4 2011 and 2012 OP TCCs

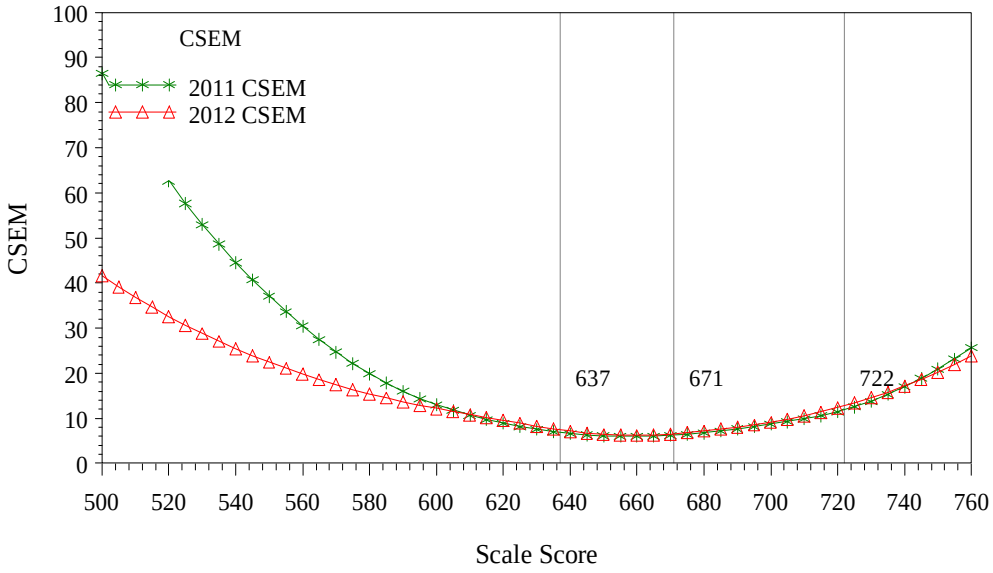

Figure 4. Grade 4 2011 and 2012 CSEM Curves

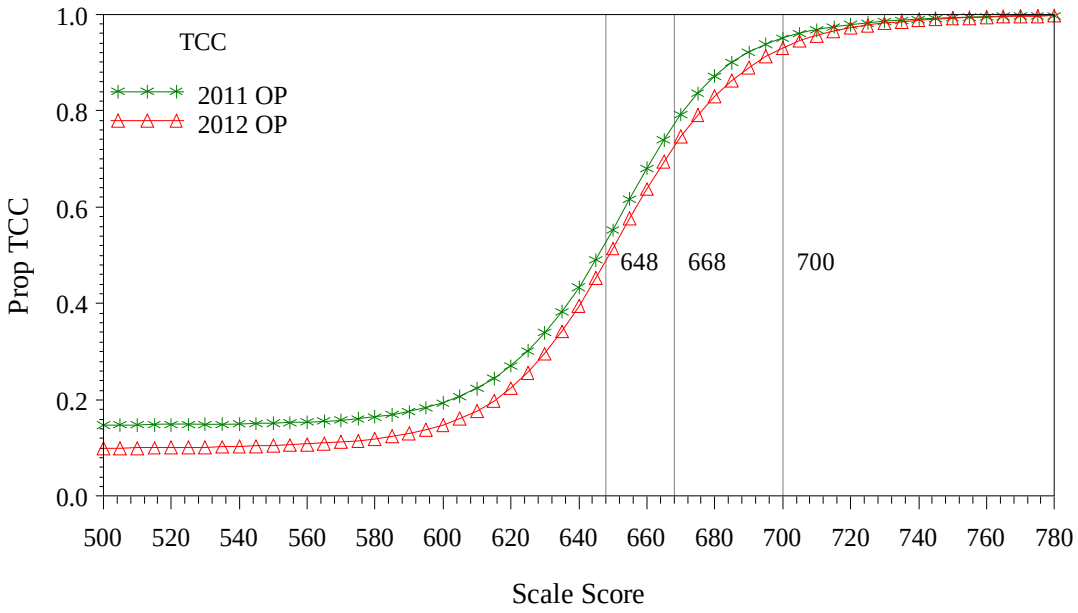

Figure 5. Grade 5 2011 and 2012 OP TCCs

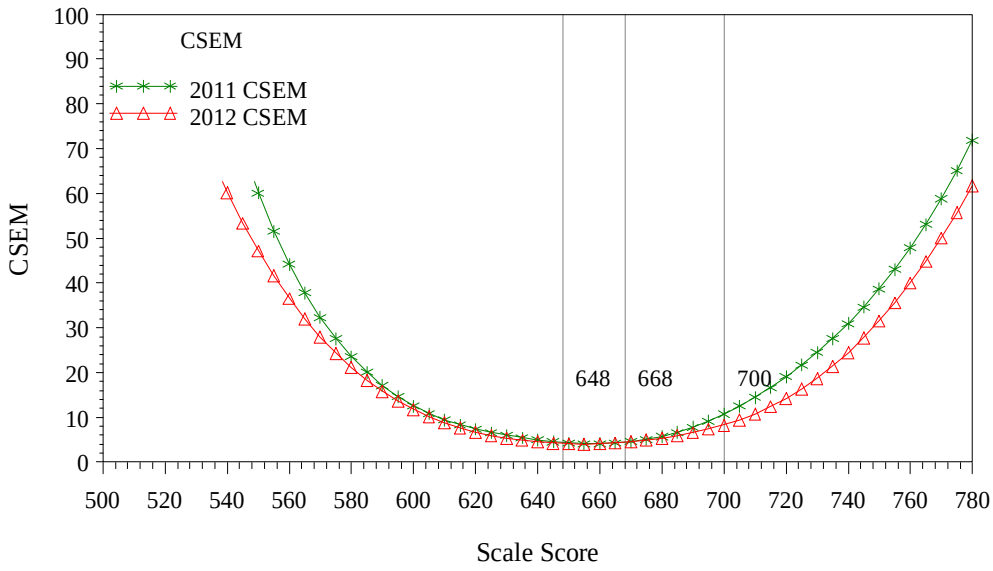

Figure 6. Grade 5 2011 and 2012 CSEM Curves

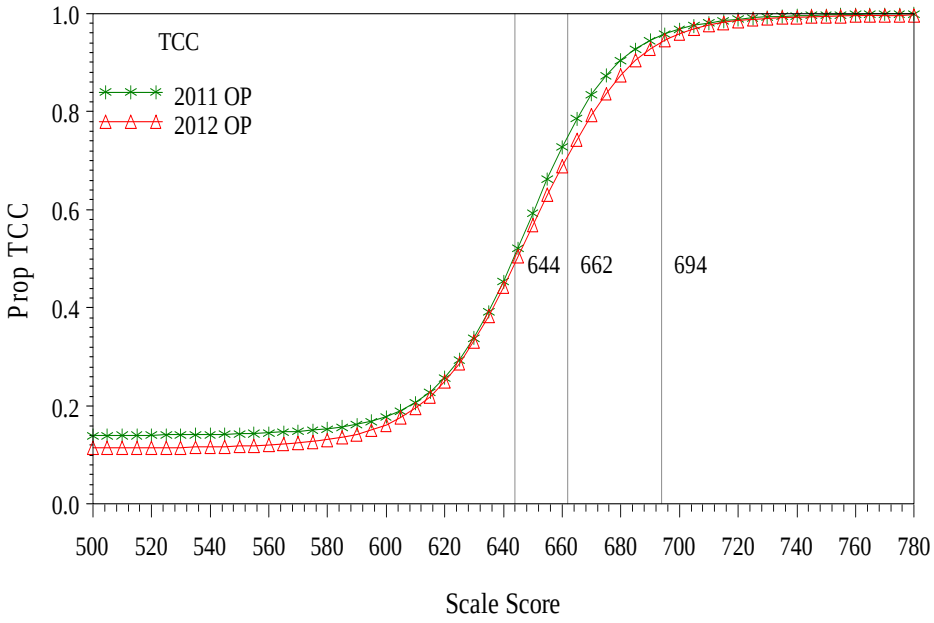

Figure 7. Grade 6 2011 and 2012 OP TCCs

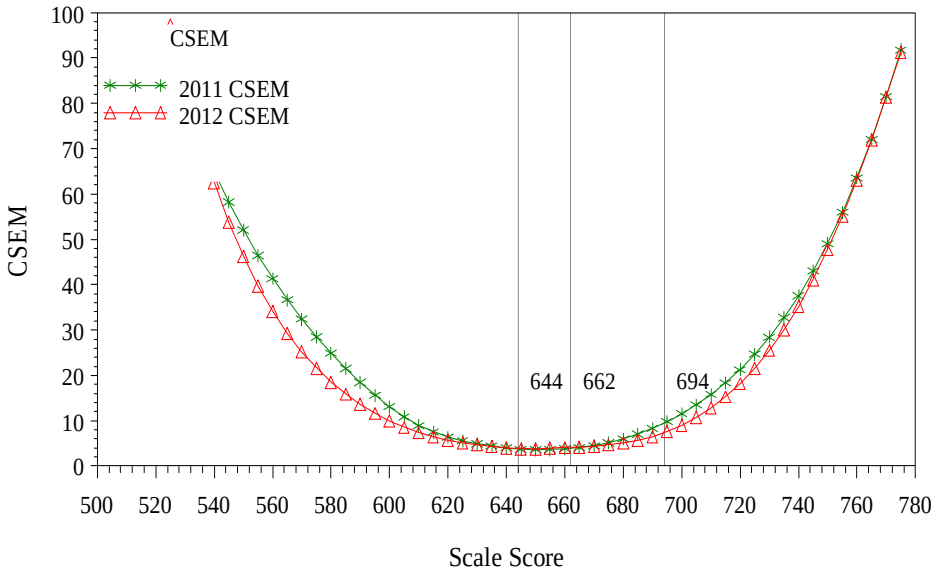

Figure 8. Grade 6 2011 and 2012 CSEM Curves

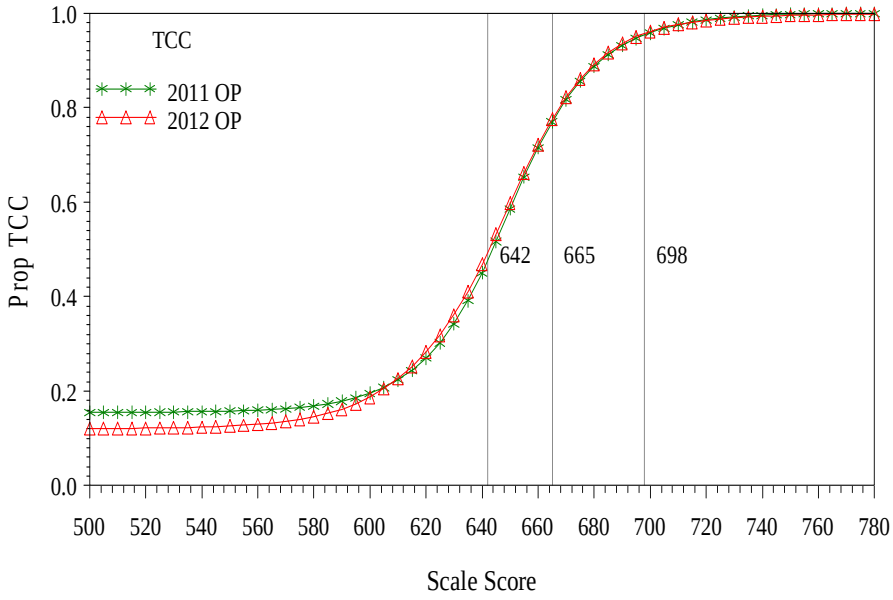

Figure 9. Grade 7 2011 and 2012 OP TCCs

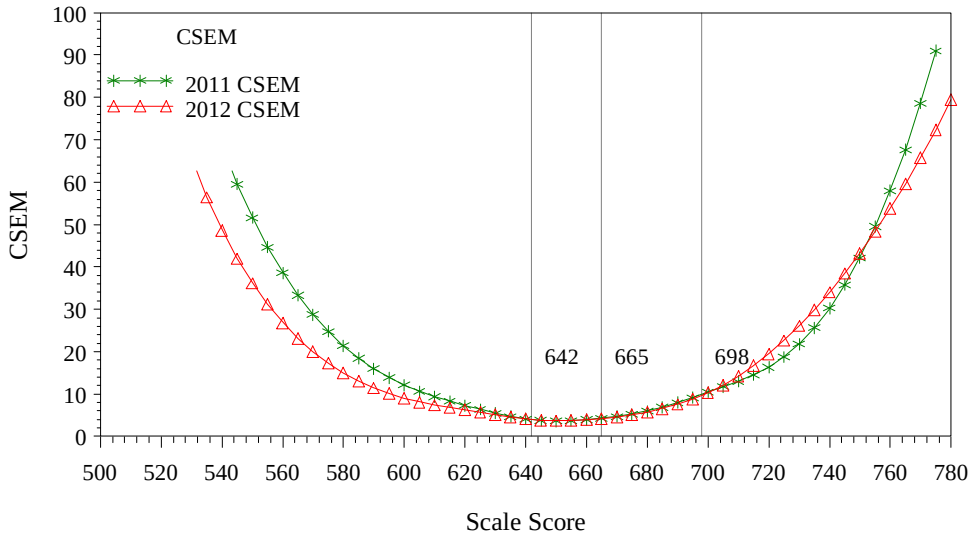

Figure 10. Grade 7 2011 and 2012 CSEM Curves

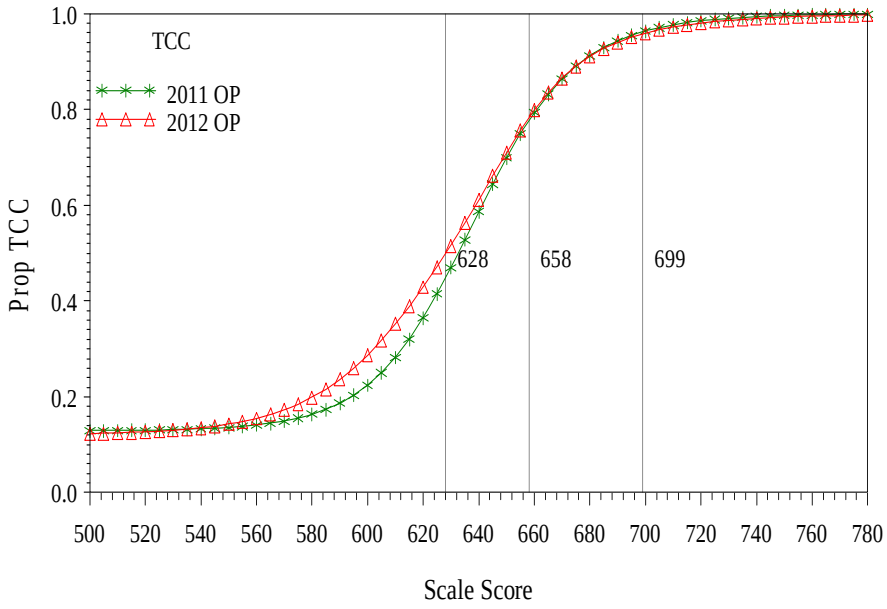

Figure 11. Grade 8 2011 and 2012 OP TCCs

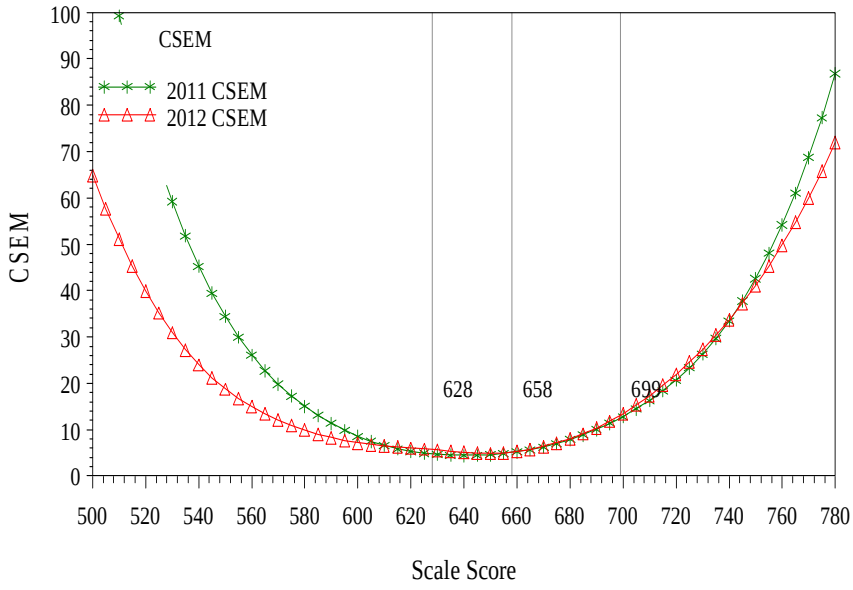

Figure 12. Grade 8 2011 and 2012 CSEM Curves

As seen in Figures 1–12, the 2012 TCCs for all grades were found to be roughly similar to the 2011 TCCs, indicating that the 2012 form were at the same difficulty level as 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 test they took.

Scoring Procedure

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

The final item parameters in the scale score metric were used to produce raw score-to-scale score conversion tables for the Grades 3–8 ELA Tests. An inverse TCC method was employed using 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 NYSTP ELA Tests have a maximum raw score higher than 30 points. In the inverse TCC method, a student's trait estimate is taken to be the trait value that has an expected raw score equal to the student's observed raw score. It was found that for tests containing all MC items, the inverse of the TCC is an excellent first-order approximation to the number of 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.

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.

Raw Score-to-Scale Score and SEM Conversion Tables

The scale score (SS) is the basic score for the NYSTP ELA Tests. It is used to derive other scores that describe test performance, such as the four performance levels and standards-based performance index scores (SPIs). Number correct raw score-to-scale score conversion tables are presented in this section. Note that the lowest and highest obtainable scale scores for each grade were the same as in 2006 (baseline year).

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

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

where

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

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

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

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

Weighted Raw Score	Scale Score	Standard Error
0	475	216
1	475	216
2	475	216
3	475	216
4	475	216
5	475	216
6	475	216
7	515	78
8	559	24
9	575	16
10	584	13
11	591	11
12	596	10
13	600	10
14	604	9
15	608	8
16	611	8
17	613	7
18	616	7

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

Weighted Raw Score	Scale Score	Standard Error
19	618	7
20	620	6
21	623	6
22	624	6
23	626	6
24	628	6
25	630	6
26	631	5
27	633	5
28	635	5
29	636	5
30	637	5
31	639	5
32	640	5
33	642	5
34	643	5
35	644	5
36	645	5
37	647	4
38	648	4
39	649	4
40	651	4
41	652	4
42	653	4
43	654	4
44	656	4
45	657	4
46	658	4
47	659	4
48	661	4

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

Weighted Raw Score	Scale Score	Standard Error
49	662	5
50	664	5
51	665	5
52	667	5
53	668	5
54	670	5
55	672	5
56	674	5
57	676	6
58	678	6
59	680	6
60	683	7
61	686	7
62	690	8
63	695	9
64	700	10
65	708	13
66	722	19
67	780	85

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

Weighted Raw Score	Scale Score	Standard Error
0	430	98
1	430	98
2	430	98
3	430	98
4	430	98
5	430	98
6	430	98
7	430	98

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

Weighted Raw Score	Scale Score	Standard Error
8	430	98
9	430	98
10	469	61
11	515	35
12	537	26
13	552	22
14	562	19
15	571	17
16	578	16
17	584	15
18	590	14
19	595	13
20	599	12
21	603	12
22	607	11
23	610	11
24	613	10
25	616	10
26	619	10
27	622	9
28	625	9
29	627	9
30	629	8
31	632	8
32	634	8
33	636	8
34	638	7
35	640	7
36	642	7
37	644	7

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

Weighted Raw Score	Scale Score	Standard Error
38	645	7
39	647	7
40	649	6
41	651	6
42	653	6
43	654	6
44	656	6
45	658	6
46	660	6
47	661	6
48	663	6
49	665	6
50	667	6
51	669	7
52	671	7
53	673	7
54	675	7
55	677	7
56	679	7
57	682	7
58	684	8
59	687	8
60	690	8
61	692	8
62	696	9
63	699	9
64	703	10
65	707	10
66	712	11
67	717	12

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

Weighted Raw Score	Scale Score	Standard Error
68	723	13
69	731	15
70	740	17
71	754	22
72	775	31
73	775	31

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

Weighted Raw Score	Scale Score	Standard Error
0	495	165
1	495	165
2	495	165
3	495	165
4	495	165
5	495	165
6	495	165
7	495	165
8	563	34
9	584	19
10	594	14
11	601	11
12	606	10
13	610	9
14	614	8
15	616	7
16	619	7
17	621	7
18	624	6
19	625	6
20	627	6

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

Weighted Raw Score	Scale Score	Standard Error
21	629	5
22	631	5
23	632	5
24	634	5
25	635	5
26	636	5
27	638	5
28	639	5
29	640	5
30	641	4
31	643	4
32	644	4
33	645	4
34	646	4
35	647	4
36	648	4
37	649	4
38	651	4
39	652	4
40	653	4
41	654	4
42	655	4
43	656	4
44	657	4
45	658	4
46	659	4
47	661	4
48	662	4
49	663	4
50	664	4

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

Weighted Raw Score	Scale Score	Standard Error
51	665	4
52	667	4
53	668	4
54	669	5
55	671	5
56	672	5
57	674	5
58	675	5
59	677	5
60	679	5
61	681	5
62	683	6
63	685	6
64	687	6
65	690	7
66	693	7
67	696	8
68	700	8
69	705	9
70	712	11
71	721	15
72	738	23
73	795	83

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

Weighted Raw Score	Scale Score	Standard Error
0	480	330
1	480	330
2	480	330
3	480	330

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

Weighted Raw Score	Scale Score	Standard Error
4	480	330
5	480	330
6	480	330
7	480	330
8	480	330
9	567	28
10	586	15
11	595	12
12	601	10
13	605	9
14	609	8
15	612	7
16	615	6
17	617	6
18	619	6
19	621	6
20	623	5
21	625	5
22	627	5
23	628	5
24	630	5
25	631	5
26	632	5
27	634	4
28	635	4
29	636	4
30	637	4
31	639	4
32	640	4
33	641	4

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

Weighted Raw Score	Scale Score	Standard Error
34	642	4
35	643	4
36	644	4
37	645	4
38	646	4
39	647	4
40	648	4
41	649	4
42	651	4
43	652	4
44	653	4
45	654	4
46	655	4
47	656	4
48	657	4
49	658	4
50	660	4
51	661	4
52	662	4
53	663	4
54	665	4
55	666	4
56	667	4
57	669	4
58	670	4
59	672	5
60	673	5
61	675	5
62	677	5
63	678	5

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

Weighted Raw Score	Scale Score	Standard Error
64	680	5
65	683	5
66	685	6
67	688	6
68	691	7
69	695	8
70	700	9
71	707	12
72	722	19
73	785	112

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

Weighted Raw Score	Scale Score	Standard Error
0	470	386
1	470	386
2	470	386
3	470	386
4	470	386
5	470	386
6	470	386
7	470	386
8	470	386
9	536	55
10	571	19
11	584	13
12	591	11
13	597	10
14	601	9
15	605	8
16	608	8

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

Weighted Raw Score	Scale Score	Standard Error
17	611	7
18	614	7
19	616	7
20	619	6
21	621	6
22	623	6
23	625	6
24	626	6
25	628	5
26	629	5
27	631	5
28	632	5
29	634	5
30	635	5
31	636	5
32	637	4
33	639	4
34	640	4
35	641	4
36	642	4
37	643	4
38	644	4
39	645	4
40	646	4
41	647	4
42	648	4
43	649	4
44	650	4
45	651	4
46	652	4

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

Weighted Raw Score	Scale Score	Standard Error
47	654	4
48	655	4
49	656	4
50	657	4
51	658	4
52	659	4
53	660	4
54	662	4
55	663	4
56	664	4
57	665	4
58	667	4
59	668	4
60	670	5
61	672	5
62	673	5
63	675	5
64	677	5
65	680	6
66	682	6
67	685	7
68	689	7
69	693	8
70	699	10
71	707	13
72	723	21
73	790	95

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

Weighted Raw Score	Scale Score	Standard Error
0	430	255
1	430	255
2	430	255
3	430	255
4	430	255
5	430	255
6	430	255
7	430	255
8	509	52
9	538	25
10	556	16
11	566	13
12	573	11
13	578	10
14	583	9
15	587	9
16	591	8
17	594	8
18	597	7
19	599	7
20	602	7
21	604	7
22	607	7
23	609	7
24	611	6
25	613	6
26	615	6
27	617	6
28	619	6
29	620	6

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

Weighted Raw Score	Scale Score	Standard Error
30	622	6
31	624	6
32	626	6
33	627	6
34	629	6
35	631	6
36	632	6
37	634	5
38	635	5
39	637	5
40	639	5
41	640	5
42	642	5
43	643	5
44	645	5
45	646	5
46	648	5
47	649	5
48	651	5
49	652	5
50	654	5
51	655	5
52	657	5
53	659	5
54	661	5
55	663	6
56	665	6
57	667	6
58	670	6
59	673	7

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

Weighted Raw Score	Scale Score	Standard Error
60	676	7
61	680	8
62	684	9
63	690	10
64	697	12
65	708	17
66	728	26
67	790	86

Standard Performance Index

The standard performance index (SPI) reported for each objective measured by the Grades 3–8 ELA Tests is an estimate of the percentage of a related set of appropriate items that the student could be expected to answer correctly. An SPI of 75 on an objective measured by a test means, for example, that the student could be expected to respond correctly to 75 out of 100 items that could be considered appropriate measures of that objective. Stated another way, an SPI of 75 indicates that the student would have a 75% chance of responding correctly to any item chosen at random from the hypothetical pool of all possible items that may be used to measure that objective.

Because objectives on all achievement tests are measured by relatively small numbers of items, 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 item response theory (IRT) and Bayesian methodology. In non-technical terms, the procedure takes into consideration the number of items related to the objective that the student answered correctly, the difficulty level of those items, as well as the student's performance on the rest of the test in which the objective is found. This use of additional information increases the accuracy of the SPI. Details on the SPI derivation procedure are provided in Appendix F.

For the 2012 Grades 3–8 ELA Tests, the performance on objectives was tied to the Level III cut score by computing the SPI target ranges. The expected SPI cuts were computed for the scale scores that are 1 standard error above and 1 standard error below the Level III cut. Table 36 presents the SPI target ranges. The objectives in this table are denoted as follows: 1—Information and Understanding, 2—Literary Response and Expression, and 3—Critical Analysis and Evaluation.

Table 33. SPI Target Ranges

Grade	Objective	# Items	Total Points	Level III Cut SPI Target Range
3	1	23	25	66–77
	2	22	28	74–82
	3	11	14	63–75
4	1	20	27	65–74
	2	28	34	70–79
	3	12	12	60–72
5	1	23	26	71–78
	2	24	28	69–78
	3	13	19	64–73
6	1	22	26	70–78
	2	28	31	66–75
	3	10	16	62–71
7	1	21	25	75–81
	2	26	27	76–84
	3	13	21	69–77
8	1	24	30	81–88
	2	19	19	68–77
	3	11	18	68–78

The SPI is most meaningful in terms of its description of the student's level of skills and knowledge measured by a given objective. The SPI increases the instructional value of test results by breaking down the information provided by the test into smaller, more manageable units. A total test score for a student in Grade 3 who scores below the average on the ELA Test does not provide sufficient information of what specific type of problem the student may be having. On the other hand, this kind of information may be provided by the SPI. For example, evidence that the student has attained an acceptable level of knowledge in the content strand of Information and Understanding but has a low level of knowledge in Literary Response and Expression provides the teacher with a good indication of what type of educational assistance might be most valuable to improve student achievement. Instruction focused on the identified needs of students has the best chance of helping those students increase their skills in the areas measured by the test. SPI reports provide students, parents, and educators the opportunity to identify and target specific areas within the broader content domain for improving student academic performance.

It should be noted that the current 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 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 (RS) and standard error of measurement (SEM), as well as the results from a study of performance level classification accuracy and consistency. The data set for these studies includes all tested New York State public and charter school students who received valid scores. A study of inter-rater reliability was conducted by 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 ELA Tests, we calculated two types of reliability statistics: Cronbach's alpha (Cronbach, 1951) and Feldt-Raju coefficient (Qualls, 1995). These two measures are appropriate for assessment of a test's internal consistency when a single test is administered to a group of examinees on one occasion. The reliability of the test is then estimated by considering how well the items that reflect the same construct yield similar results (or how consistent the results are for different items for the same construct measured by the test). Both Cronbach's alpha and Feldt-Raju coefficient measures are appropriate for tests of multiple-item formats (MC and CR items).

Reliability for Total Test

Overall test reliability is a very good indication of each test's internal consistency. Included in Table 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 ELA Tests.

Table 34. ELA 3–8 Tests Reliability and Standard Error of Measurement

Grade	N-count	# Items	# RS Points	Cronbach's Alpha	SEM of Cronbach	Feldt-Raju Coefficient	SEM of Feldt-Raju
3	197,993	56	67	0.92	3.23	0.92	3.15
4	194,344	60	73	0.92	3.40	0.92	3.34
5	196,623	60	73	0.91	3.47	0.92	3.35
6	199,540	60	73	0.91	3.46	0.91	3.37
7	197,638	60	73	0.92	3.28	0.92	3.19
8	198,294	54	67	0.90	3.12	0.90	3.04

All the coefficients for total test reliability were in the range 0.90–0.92, which indicates high internal consistency. As expected, the lowest reliabilities were found for the shortest test (i.e., Grade 8), and the highest reliabilities were associated with the longer tests (Grades 4 and 7).

Reliability of MC Items

In addition to overall test reliability, Cronbach's alpha and Feldt-Raju coefficient were computed separately for MC and CR item sets. It is important to recognize that reliability is directly affected by test length; therefore, reliability estimates for tests by item type will always be lower than reliability estimates for the overall test form. Table 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	197,993	47	0.90	2.60	0.90	2.57
4	194,344	51	0.90	2.87	0.90	2.86
5	196,623	51	0.88	2.82	0.88	2.81
6	199,540	51	0.88	2.90	0.88	2.88
7	197,638	51	0.90	2.77	0.90	2.75
8	198,294	45	0.87	2.66	0.87	2.64

Reliability of CR Items

Reliability coefficients were also computed for the subsets of CR items. 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	197,993	9	20	0.79	1.81	0.79	1.78
4	194,344	9	22	0.76	1.73	0.77	1.68
5	196,623	9	22	0.81	1.85	0.82	1.77
6	199,540	9	22	0.79	1.76	0.80	1.69
7	197,638	9	22	0.81	1.62	0.83	1.54
8	198,294	9	22	0.77	1.54	0.80	1.45

Note: Results should be interpreted with caution because the number of items is low.

Test Reliability for NCLB Reporting Categories

In this section, reliability coefficients that were estimated for the population and NCLB reporting subgroups are presented. The reporting categories include the following: gender, ethnicity, Needs/Resource Capacity Category (NRC), English language learners (ELL), all students with disabilities (SWD), all students using test accommodations (SUA), students with disabilities using accommodations falling under 504 Plan (SWD/SUA), and English language learners using accommodations specific to their ELL status (ELL/SUA). Accommodations available to students under the 504 Plan are the following: Flexibility in Scheduling/Timing, Flexibility in Setting, Method of Presentation (excluding braille), Method of Response, Braille and Large Type, and others. Accommodations available to English language learners are Time Extension, Separate Location, Third Reading of Listening Selection, and Bilingual Dictionaries and Glossaries.

As shown in Tables 37A–37F, the estimated reliabilities for subgroups were close in magnitude to the test reliability estimates of the population. Cronbach's alpha reliability coefficients were all greater than 0.80. Feldt-Raju reliability coefficients, which tend to be larger than the Cronbach's alpha estimates for the same group, were all larger than 0.80 too. All other test reliability alpha statistics were in the 0.84–0.93 range, indicating 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	197,993	0.92	3.23	0.92	3.15
Gender	Female	96,905	0.92	3.18	0.92	3.10
	Male	101,088	0.92	3.27	0.93	3.19
Ethnicity	Asian	16,463	0.92	3.02	0.92	2.94
	Black	35,676	0.91	3.42	0.92	3.34
	Hispanic	47,112	0.92	3.40	0.92	3.33
	American Indian	1,081	0.91	3.32	0.92	3.24
	Multiracial	2,010	0.92	3.19	0.92	3.09
	Other	403	0.91	3.21	0.92	3.13
	White	95,248	0.91	3.09	0.91	3.00
NRC	New York City	71,611	0.92	3.31	0.92	3.23
	Big 4 Cities	8,022	0.93	3.53	0.93	3.44
	High Needs Urban/Suburban	15,592	0.92	3.38	0.92	3.30
	High Needs Rural	11,018	0.91	3.31	0.92	3.22
	Average Needs	57,844	0.91	3.15	0.91	3.07
	Low Needs	27,939	0.89	2.93	0.90	2.86
	Charter	5,967	0.89	3.25	0.89	3.19
SWD	All Codes	28,201	0.92	3.59	0.93	3.51
SUA	All Codes	27,303	0.92	3.57	0.92	3.49
ELL	ELL=Y	17,112	0.91	3.63	0.91	3.55
SWD/SUA	SUA=504 plan codes	14,459	0.92	3.62	0.92	3.55
ELL/SUA	SUA=ELL codes	7,380	0.91	3.63	0.91	3.55

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	194,344	0.92	3.40	0.92	3.34
Gender	Female	95,398	0.91	3.34	0.91	3.28
	Male	98,946	0.92	3.45	0.92	3.39
Ethnicity	Asian	16,144	0.92	3.17	0.92	3.12
	Black	35,445	0.91	3.59	0.91	3.54
	Hispanic	45,518	0.91	3.58	0.91	3.52
	American Indian	1,035	0.91	3.52	0.92	3.47
	Multiracial	1,696	0.91	3.38	0.92	3.31
	Other	316	0.92	3.28	0.92	3.21
	White	94,190	0.91	3.26	0.91	3.20
NRC	New York City	69,619	0.92	3.48	0.92	3.42
	Big 4 Cities	7,992	0.92	3.74	0.92	3.67
	High Needs Urban/Suburban	15,193	0.92	3.57	0.92	3.51
	High Needs Rural	11,026	0.91	3.46	0.92	3.40
	Average Needs	57,535	0.91	3.32	0.91	3.26
	Low Needs	28,193	0.89	3.10	0.89	3.06
	Charter	4,786	0.89	3.37	0.89	3.34
SWD	All Codes	29,610	0.91	3.81	0.91	3.74
SUA	All Codes	23,132	0.91	3.79	0.92	3.71
ELL	ELL=Y	15,929	0.89	3.84	0.90	3.76
SWD/SUA	SUA=504 plan codes	13,764	0.91	3.84	0.91	3.77
ELL/SUA	SUA=ELL codes	4,819	0.89	3.84	0.89	3.77

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	196,623	0.91	3.47	0.92	3.35
Gender	Female	96,487	0.91	3.42	0.91	3.30
	Male	100,136	0.92	3.50	0.92	3.39
Ethnicity	Asian	15,777	0.91	3.23	0.92	3.11
	Black	36,584	0.91	3.65	0.91	3.55
	Hispanic	45,248	0.91	3.61	0.91	3.51
	American Indian	992	0.90	3.62	0.91	3.50
	Multiracial	1,541	0.91	3.47	0.92	3.34
	Other	317	0.90	3.43	0.91	3.30
	White	96,164	0.90	3.33	0.91	3.22
NRC	New York City	68,552	0.91	3.51	0.92	3.40
	Big 4 Cities	7,965	0.92	3.79	0.93	3.66
	High Needs Urban/Suburban	14,975	0.91	3.63	0.92	3.53
	High Needs Rural	11,387	0.91	3.55	0.91	3.43
	Average Needs	59,075	0.90	3.40	0.91	3.29
	Low Needs	28,658	0.89	3.14	0.89	3.05
	Charter	6,011	0.89	3.58	0.89	3.50
SWD	All Codes	30,703	0.91	3.85	0.91	3.75
SUA	All Codes	25,237	0.91	3.83	0.92	3.72
ELL	ELL=Y	13,537	0.89	3.88	0.90	3.80
SWD/SUA	SUA=504 plan codes	16,158	0.91	3.87	0.91	3.77
ELL/SUA	SUA=ELL codes	4,001	0.89	3.90	0.90	3.81

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	199,540	0.91	3.46	0.91	3.37
Gender	Female	97,552	0.90	3.38	0.91	3.30
	Male	101,988	0.91	3.51	0.91	3.43
Ethnicity	Asian	16,679	0.91	3.26	0.92	3.17
	Black	37,110	0.89	3.64	0.89	3.56
	Hispanic	44,565	0.90	3.62	0.90	3.54
	American Indian	1,004	0.90	3.57	0.91	3.48
	Multiracial	1,448	0.90	3.40	0.91	3.32
	Other	347	0.93	3.40	0.93	3.27
	White	98,387	0.89	3.32	0.90	3.24
NRC	New York City	68,954	0.91	3.54	0.91	3.46
	Big 4 Cities	7,758	0.90	3.76	0.91	3.66
	High Needs Urban/Suburban	14,741	0.90	3.61	0.91	3.53
	High Needs Rural	11,423	0.90	3.54	0.90	3.45
	Average Needs	60,856	0.89	3.37	0.90	3.30
	Low Needs	30,289	0.88	3.15	0.88	3.09
	Charter	5,519	0.87	3.56	0.87	3.51
SWD	All Codes	30,429	0.88	3.85	0.89	3.76
SUA	All Codes	22,594	0.89	3.83	0.89	3.74
ELL	ELL=Y	11,307	0.85	3.91	0.86	3.80
SWD/SUA	SUA=504 plan codes	14,552	0.88	3.87	0.88	3.77
ELL/SUA	SUA=ELL codes	3,383	0.85	3.92	0.86	3.80

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	197,638	0.92	3.28	0.92	3.19
Gender	Female	96,364	0.91	3.18	0.91	3.11
	Male	101,274	0.92	3.34	0.92	3.26
Ethnicity	Asian	15,534	0.92	3.03	0.93	2.94
	Black	37,089	0.91	3.50	0.91	3.43
	Hispanic	43,499	0.91	3.47	0.92	3.40
	American Indian	1,011	0.91	3.45	0.91	3.38
	Multiracial	1,379	0.91	3.26	0.91	3.18
	Other	324	0.92	3.30	0.92	3.21
	White	98,802	0.91	3.12	0.91	3.04
NRC	New York City	67,421	0.92	3.37	0.92	3.29
	Big 4 Cities	7,598	0.92	3.67	0.92	3.58
	High Needs Urban/Suburban	14,572	0.92	3.48	0.92	3.39
	High Needs Rural	11,705	0.91	3.37	0.92	3.28
	Average Needs	60,119	0.91	3.19	0.91	3.11
	Low Needs	31,553	0.88	2.91	0.89	2.85
	Charter	4,670	0.87	3.35	0.87	3.30
SWD	All Codes	29,844	0.91	3.77	0.91	3.69
SUA	All Codes	22,581	0.91	3.74	0.92	3.65
ELL	ELL = Y	10,802	0.89	3.87	0.90	3.76
SWD/SUA	SUA=504 plan codes	15,223	0.91	3.78	0.91	3.70
ELL/SUA	SUA=ELL codes	2,964	0.89	3.86	0.90	3.76

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	198,294	0.90	3.12	0.90	3.04
Gender	Female	97,523	0.89	3.04	0.90	2.97
	Male	100,771	0.90	3.18	0.90	3.11
Ethnicity	Asian	15,838	0.91	2.90	0.92	2.82
	Black	36,955	0.88	3.34	0.88	3.28
	Hispanic	43,156	0.89	3.32	0.89	3.25
	American Indian	1,029	0.90	3.25	0.90	3.18
	Multiracial	1,143	0.90	3.06	0.90	2.97
	Other	333	0.92	3.21	0.92	3.10
	White	99,840	0.88	2.95	0.89	2.88
NRC	New York City	68,901	0.90	3.25	0.90	3.18
	Big 4 Cities	7,278	0.90	3.42	0.91	3.35
	High Needs Urban/Suburban	1,4208	0.89	3.29	0.90	3.22
NRC	High Needs Rural	1,1581	0.89	3.16	0.90	3.09
	Average Needs	60,742	0.88	3.01	0.89	2.94
	Low Needs	32,134	0.86	2.76	0.87	2.71
	Charter	3,450	0.84	3.22	0.84	3.17
SWD	All Codes	29,640	0.88	3.57	0.88	3.50
SUA	All Codes	19,301	0.89	3.53	0.89	3.46
ELL	ELL = Y	10,702	0.85	3.69	0.86	3.61
SWD/SUA	SUA=504 plan codes	13,779	0.88	3.57	0.88	3.50
ELL/SUA	SUA=ELL codes	1,794	0.86	3.69	0.87	3.61

Standard Error of Measurement

The SEM, as computed from Cronbach's alpha and the Feldt-Raju reliability statistics, are presented in Table 37. The SEMs ranged 3.04–3.47, which is reasonable and small. In other words, the error of measurement from the observed test score ranged from approximately ± 3 to ± 3.5 raw score points. The SEMs are directly related to reliability: the higher the reliability, the lower the standard error. As discussed, the reliability of these tests is relatively high, so it was expected that the SEMs would be very low.

The SEMs for subpopulations, as computed from Cronbach's alpha and the Feldt-Raju reliability statistics, are presented in Tables 37A–37F. The SEMs associated with all reliability estimates for all subpopulations are in the range 2.71–3.92, which is acceptably close to those for the entire population. This narrow range indicates that across the Grades 3–8 ELA Tests, all students' test scores are reasonably reliable with minimal error.

Performance Level Classification Consistency and Accuracy

This subsection describes the analyses conducted to estimate performance level classification consistency and accuracy for the Grades 3–8 ELA Tests. In other words, this provides statistical information on the classification of students into the four performance categories. Classification consistency refers to the estimated degree of agreement between examinees' performance classification from two independent administrations of the same test (or from two parallel forms of the test). Because obtaining test scores from two independent administrations of New York State tests was not feasible due to item release after each administration, a psychometric model was used to obtain the estimated classification consistency indices using test scores from a single administration. Classification accuracy can be defined as the agreement between the actual classifications using observed cut scores and true classifications based on known true cut scores (Livingston and Lewis, 1995).

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

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

Classification consistency and accuracy were estimated using the IRT procedure suggested by Lee, Hanson, and Brennan (2002) and Wang, Kolen, and Harris (2000). Appendix H includes a description of the calculations and procedure based on the paper by Lee et al. (2002).

Consistency

The results for classifying students into four performance levels are separated from results based solely on the Level III cut. Tables 39 and 40 include case counts (N-count), classification consistency (Agreement), classification inconsistency (Inconsistency), and Cohen's kappa (Kappa). Consistency indicates the rate that a second administration would yield the same performance category designation (or a different designation for the inconsistency rate). The agreement index is a sum of the diagonal element in the contingency table. The inconsistency index is equal to the "1 – agreement index." Kappa is a measure of agreement corrected for chance.

Table 38 depicts the consistency study results based on the range of performance levels for all grades. Overall, between 76 and 79% of students were estimated to be classified consistently to one of the four performance categories. The coefficient kappa, which indicates the consistency of the placement in the absence of chance, ranged 0.66–0.68.

Table 38. Decision Consistency (All Cuts)

Grade	N-count	Agreement	Inconsistency	Kappa
3	193,436	0.76	0.24	0.66
4	190,402	0.78	0.22	0.66
5	192,453	0.77	0.23	0.66
6	195,517	0.79	0.21	0.67
7	193,678	0.79	0.21	0.68
8	192,150	0.79	0.21	0.66

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 89 to 90% of the classifications of individual students are estimated to remain stable with a second administration. Kappa coefficients for classification consistency based on one cut ranged 0.77–0.80.

Table 39. Decision Consistency (Level III Cut)

Grade	N-count	Agreement	Inconsistency	Kappa
3	193,436	0.90	0.10	0.80
4	190,402	0.89	0.11	0.79
5	192,453	0.89	0.11	0.78
6	195,517	0.89	0.11	0.79
7	193,678	0.90	0.10	0.80
8	192,150	0.89	0.11	0.77

Accuracy

The results of classification accuracy are presented in Table 40. Included in the table are case counts (N-count) and classification accuracy (Accuracy) for all performance levels (All Cuts) and for the Level III cut score, as well as "false positive" and "false negative" rates for both scenarios. It is always the case that the accuracy of the Level III cut score exceeds the accuracy referring to the entire set of cut scores, because there are only two categories in which the true variable can be located, not four. The accuracy rates indicate that the categorization of a student's observed performance is in agreement with the location of his or her true ability 82 to 85% of the time across all performance levels and 92–93% of the time in regard 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	193,436	**0.82**	0.13	0.05	**0.93**	0.04	0.03
4	190,402	**0.84**	0.12	0.05	**0.92**	0.06	0.02
5	192,453	**0.84**	0.11	0.05	**0.92**	0.05	0.03
6	195,517	**0.84**	0.13	0.04	**0.92**	0.06	0.02
7	193,678	**0.85**	0.09	0.07	**0.93**	0.03	0.04
8	192,150	**0.85**	0.11	0.04	**0.92**	0.06	0.02

Section VIII: Summary of Operational Test Results

This section summarizes the distribution of OP scale score results on the NYSTP 2012 Grades 3–8 ELA Tests. These include the scale score means, standard deviations, percentiles, and performance level distributions for each grade's population and specific subgroups. Gender, ethnic identification, Needs/Resource Capacity Category (NRC), English language learners (ELL), students with disabilities (SWD), and students using test accommodations (SUA) variables were used to calculate the results of subgroups required for federal reporting and test equity purposes. Especially, 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 H.

Scale Score Distribution Summary

Scale score distribution summary tables are presented and discussed in Tables 41–47. In Table 41, scale score statistics for total populations of students from public and charter schools are presented. In Tables 42–47, scale score statistics are presented for selected subgroups in each grade level. Some general observations: Females outperformed Males; Asian and White ethnicities outperformed their peers from other ethnic groups; students from Low Needs and Average Needs districts (as identified by NRC) outperformed students from other districts (New York City, Big 4 Cities, Urban/Suburban, Rural, and Charter); and students with ELL, SWD, and/or SUA achieved below the State aggregate (All Students) in every percentile. This pattern of achievement was consistent across all grades.

Table 41. ELA Grades 3–8 Scale Score Distribution Summary

Grade	N-count	SS Mean	SS Std Dev	10th %tile	25th %tile	50th %tile	75th %tile	90th %tile
3	197,993	664.46	21.11	639	653	665	678	690
4	194,344	674.54	29.97	640	658	677	692	707
5	196,623	669.93	18.29	647	659	671	681	693
6	199,540	663.18	16.23	643	653	663	673	683
7	197,638	665.35	18.44	644	655	665	675	685
8	198,294	658.09	21.16	634	646	659	670	684

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 664.46 with a standard deviation of 21.11. By gender subgroup, Females outperformed Males, and the difference was around 5 scale score points. Asian, Multiracial, and White students' scale score means exceeded the average scale score, as did students from Low Needs and Average Needs districts. Among all ethnic groups, the Asian

ethnic group had the highest average scale score mean (671.89). Students from the Big 4 Cities achieved a lower scale score mean than their peers from schools with other NRC designations and about two-thirds of a standard deviation below the population mean. The SWD, SUA, and ELL subgroups scored, on average, approximately four-fifth of one standard deviation below the mean scale score for the population. The SWD/SUA subgroup, which had a scale score mean about 23 scale score units below the population mean, was the lowest performing group analyzed. At the 50th percentile, the following groups exceeded the population score of 665: Female (667), Asian (672), Multiracial (667), White (670), Average Needs districts (668), and Low Needs districts (674).

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	197,993	664.46	21.11	639	653	665	678	690
Gender	Female	96,905	666.93	20.50	643	656	667	678	690
	Male	101,088	662.10	21.41	636	651	664	676	686
Ethnicity	Asian	16,463	671.89	21.82	647	661	672	683	695
	Black	35,676	657.16	19.88	633	647	658	670	680
	Hispanic	47,112	657.71	20.34	633	647	659	670	680
	American Indian	1,081	659.99	20.43	636	649	661	672	683
	Multiracial	2,010	665.51	20.51	640	653	667	678	690
	Other	403	665.66	22.47	642	654	667	678	686
	White	95,248	669.27	20.00	645	658	670	680	690
NRC	New York City	71,611	661.59	21.41	636	651	662	674	686
	Big 4 Cities	8,022	650.85	22.91	623	639	653	665	676
	High Needs Urban/ Suburban	15,592	658.09	20.55	633	647	659	670	680
	High Needs Rural	11,018	660.82	19.56	637	651	662	674	683
	Average Needs	57,844	667.44	19.64	644	657	668	678	690
	Low Needs	27,939	674.64	19.12	653	664	674	686	695
	Charter	5,967	664.15	16.27	644	654	665	674	683
SWD	All Codes	28,201	644.47	22.73	616	631	647	659	670

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
SUA	All Codes	27,303	648.23	21.72	620	636	651	662	674
ELL	ELL=Y	17,112	646.12	20.37	620	636	649	659	668
SWD/SUA	SUA=504 plan codes	14,459	641.69	22.03	616	628	644	656	667
ELL/SUA	SUA=ELL codes	7,380	647.39	19.58	623	637	651	661	668

Grade 4

Scale score statistics and N-counts of demographic groups for Grade 4 are presented in Table 43. The Grade 4 population (All Students) mean was 674.54, with a standard deviation of 29.97. By gender subgroup, Females outperformed Males, and the difference was around 7 scale score points. Asian, Multiracial, and White students' scale score means exceeded the average scale score, as did students from Low Needs, average Needs districts and charter schools. Among all ethnic groups, the Asian ethnic group had the highest average scale score mean (685.55). Students from the Big 4 Cities achieved a lower scale score mean than their peers from schools with other NRC designations and about two-thirds of a standard deviation below the population mean. The SWD/SUA subgroup had a scale score mean nearly 33 scale score units below the population mean and was at or below the scale score of any given percentile for any other subgroup. At the 50th percentile, the following groups exceeded the population score of 675: Female (679), Asian (687), White (682), Average Needs districts (679), Low Needs districts (690), and charter (677).

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	194,344	674.54	29.97	640	658	677	692	707
Gender	Female	95,398	678.22	28.76	644	661	679	696	712
	Male	98,946	670.99	30.67	634	654	673	690	707
Ethnicity	Asian	16,144	685.55	31.09	651	669	687	703	723
	Black	35,445	664.16	28.43	632	649	665	682	696
	Hispanic	45,518	664.90	28.84	632	651	667	682	696
	American Indian	1,035	667.26	28.85	632	651	669	687	699
	Multiracial	1,696	675.05	29.85	640	658	677	694	712
	Other	316	680.60	32.30	644	663	683	699	717
	White	94,190	681.27	28.29	649	667	682	699	712

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
NRC	New York City	69,619	670.62	30.33	636	654	671	690	707
	Big 4 Cities	7,992	653.57	32.87	616	638	656	675	690
	High Needs Urban/ Suburban	15,193	664.98	29.88	629	649	667	684	699
	High Needs Rural	11,026	669.85	29.12	634	654	673	687	703
	Average Needs	57,535	678.65	27.76	645	663	679	696	712
	Low Needs	28,193	688.50	26.06	658	673	690	703	717
	Charter	4,786	676.14	23.86	647	661	677	692	707
SWD	All Codes	29,610	644.94	33.01	607	627	649	665	682
SUA	All Codes	23,132	649.32	32.46	610	634	653	671	684
ELL	ELL=Y	15,929	646.16	30.26	610	634	651	665	677
SWD/ SUA	SUA=504 plan codes	13,764	641.48	33.07	603	625	645	663	677
ELL/SUA	SUA=ELL codes	4,819	647.80	30.15	613	636	653	667	679

Grade 5

Scale score summary statistics for Grade 5 students are in Table 44. Overall, the scale score mean was 669.93, with a standard deviation of 18.29. The difference between mean scale scores by gender groups was about 4 scale score units. Female, Asian, Multiracial, and White students' scale score means exceeded the population mean scale score, as did students from Low Needs and Average Needs districts. The students from the Big 4 Cities scored below their peers from schools with other NRC designations and about two-thirds of a standard deviation below the population mean. The SWD, SUA, and ELL subgroups scored approximately one standard deviation below the mean scale score for the population. The SWD/SUA subgroup, which had a scale score mean nearly 20 scale score units below the population mean, was the lowest performing group analyzed. At the 50th percentile, the following groups exceeded the population score of 670: Female (672), Asian (677), Multiracial (671), White (675), Average Needs (672) and Low Needs districts (679).

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	196,623	669.93	18.29	647	659	671	681	693
Gender	Female	96,487	671.98	18.06	649	661	672	683	693
	Male	100,136	667.95	18.30	645	657	669	679	690
Ethnicity	Asian	15,777	676.81	19.58	654	667	677	687	700
	Black	36,584	663.06	17.18	641	653	664	674	683
	Hispanic	45,248	664.50	17.18	643	655	665	675	685
	American Indian	992	665.38	17.77	645	655	664	677	687
	Multiracial	1,541	670.08	18.47	647	658	671	683	693
	Other	317	672.20	16.81	651	662	674	685	693
	White	96,164	674.00	17.44	653	664	675	685	693
NRC	New York City	68,552	668.04	18.48	646	657	668	679	690
	Big 4 Cities	7,965	657.50	19.42	634	645	658	671	681
	High Needs Urban/ Suburban	14,975	664.04	17.49	643	654	665	675	685
	High Needs Rural	11,387	666.73	17.01	645	657	668	677	687
	Average Needs	59,075	671.91	17.09	651	662	672	683	693
	Low Needs	28,658	679.01	16.66	659	669	679	690	696
	Charter	6,011	665.81	15.76	647	656	667	675	685
SWD	All Codes	30,703	651.94	17.77	631	641	653	663	672
SUA	All Codes	25,237	654.13	17.72	632	644	655	665	675
ELL	ELL=Y	13,537	650.51	16.61	629	641	652	662	669
SWD/ SUA	SUA=504 plan codes	16,158	650.33	17.48	629	640	652	662	671
ELL/SUA	SUA=ELL codes	4,001	650.41	16.27	631	640	652	662	669

Grade 6

Scale score summary statistics for Grade 6 students are in Table 45. The scale score mean was 663.18, with a standard deviation of 16.23. Female, Asian, Multiracial, and White students' scale score means exceeded the population mean scale score, as did students from Low Needs and Average Needs districts. The students from the Big 4 Cities scored below their peers from

schools with other NRC designations and about a half of a standard deviation below the population mean. The SWD and SUA subgroups scored about one standard deviation below the mean scale score for the population. The ELL subgroup, which had a scale score mean more than 21 scale score units below the population mean, was the lowest performing group analyzed. At the 50th percentile, the following groups exceeded the population score of 663: Female (666), Asian (668), Multiracial (669), White (669), Average Needs districts (666), and Low Needs districts (672).

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	199,540	663.18	16.23	643	653	663	673	683
Gender	Female	97,552	665.29	15.83	646	655	666	675	685
	Male	101,988	661.16	16.35	641	652	662	672	680
Ethnicity	Asian	16,679	667.99	17.12	647	658	669	678	688
	Black	37,110	656.89	14.89	640	648	657	666	675
	Hispanic	44,565	657.07	15.17	639	648	657	667	675
	American Indian	1,004	659.00	15.24	641	649	660	669	678
	Multiracial	1,448	665.57	15.72	646	655	666	675	685
	Other	347	665.57	19.42	642	654	665	678	688
	White	98,387	667.51	15.31	648	658	669	677	685
NRC	New York City	68,954	659.61	16.21	640	649	660	670	678
	Big 4 Cities	7,758	653.37	16.36	634	644	654	665	673
	High Needs Urban/ Suburban	14,741	658.42	15.42	640	649	658	669	677
	High Needs Rural	11,423	661.19	15.30	643	653	662	672	678
	Average Needs	60,856	666.02	15.21	647	657	666	675	685
	Low Needs	30,289	671.69	14.58	654	663	672	680	688
	Charter	5,519	660.47	12.94	644	652	661	669	677
SWD	All Codes	30,429	647.14	15.10	630	639	647	656	665
SUA	All Codes	22,594	648.63	15.40	631	640	649	658	666
ELL	ELL=Y	11,307	641.86	14.03	625	634	643	651	657
SWD/ SUA	SUA=504 plan codes	14,552	646.29	15.05	628	637	647	656	663
ELL/SUA	SUA=ELL codes	3,383	641.76	14.66	625	635	643	651	657

Grade 7

Scale score statistics and N-counts of demographic groups for Grade 7 are presented in Table 46. The population scale score mean was 665.35 and the population standard deviation was 18.44. By gender subgroup, Females outperformed Males, the difference was about one-fourth of a standard deviation. Female, Asian, Multiracial, and White students' scale score means exceeded the population mean scale score, as did students from Low Needs and Average Needs districts. Among all ethnic groups the Asian ethnic group had the highest average scale score mean (671.45). The students from the Big 4 Cities scored below their peers from schools with other NRC designations and about two-thirds of a standard deviation below the population mean. The SWD and SUA subgroups scored approximately one standard deviation below the mean scale score for the population. The ELL subgroup, which had a scale score mean more than 24 scale score units below the population mean, was the lowest performing group analyzed. At the 50th percentile, the following groups exceeded the population score of 665: Female (668), Asian (672), Multiracial (667), White (670), Average Needs districts (668), and Low Needs districts (673).

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	197,638	665.35	18.44	644	655	665	675	685
Gender	Female	96,364	667.82	17.83	647	658	668	677	689
	Male	101,274	663.01	18.71	641	652	664	673	685
Ethnicity	Asian	15,534	671.45	20.45	649	662	672	682	693
	Black	37,089	658.33	16.60	639	649	659	668	677
	Hispanic	43,499	658.82	17.08	639	650	660	670	677
	American Indian	1,011	660.10	16.73	641	651	660	670	677
	Multiracial	1,379	666.90	18.65	646	657	667	677	689
	Other	324	665.10	19.84	646	656	665	675	685
	White	98,802	669.94	17.58	650	660	670	680	689
NRC	New York City	67,421	661.82	18.38	641	652	663	673	682
	Big 4 Cities	7,598	652.65	18.65	629	643	655	664	673
	High Needs Urban/ Suburban	14,572	659.52	17.46	639	649	660	670	680
	High Needs Rural	11,705	663.08	17.19	643	654	663	673	682
	Average Needs	60,119	668.19	17.48	648	659	668	677	689
	Low Needs	31,553	674.48	16.85	656	665	673	682	693
	Charter	4,670	662.73	13.24	647	655	663	672	677

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
SWD	All Codes	29,844	647.05	17.42	626	637	649	658	665
SUA	All Codes	22,581	649.39	17.48	628	640	651	660	668
ELL	ELL=Y	10,802	641.66	18.13	621	634	644	654	660
SWD/SUA	SUA=504 plan codes	15,223	646.76	16.99	626	637	648	658	665
ELL/SUA	SUA=ELL codes	2,964	642.31	17.08	621	634	645	654	660

Grade 8

Scale score statistics and N-counts of demographic groups for Grade 8 are presented in Table 47. The population scale score mean was 658.09 with a standard deviation of 21.16. By gender subgroup, Females outperformed Males, but the difference was less than 6 scale score points. Female, Asian, Multiracial, and White students' scale score means exceeded the population mean scale score, as did students from Low Needs and Average Needs districts. The students from the Big 4 Cities scored below their peers from schools with other NRC designations and about two-thirds of a standard deviation below the population mean. The SWD and SUA subgroups scored approximately one standard deviation below the mean scale score for the population. The ELL subgroup, which had a scale score mean more than 29 scale score units below the population mean, was the lowest performing group analyzed. At the 50th percentile, the following groups exceeded the population score of 658: Female (661), Asian (665), Multiracial (659), White (663), Average Needs districts (663), and Low Needs districts (667).

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	198,294	658.09	21.16	634	646	659	670	684
Gender	Female	97,523	661.19	21.06	637	649	661	673	684
	Male	100,771	655.10	20.83	631	643	655	667	680
Ethnicity	Asian	15,838	664.72	24.24	637	652	665	680	690
	Black	36,955	649.34	18.40	627	639	651	661	670
	Hispanic	43,156	650.20	19.75	626	640	651	663	673
	American Indian	1,029	653.58	20.42	629	642	654	665	676
	Multiracial	1,143	660.77	21.37	635	648	659	673	684

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	Other	333	654.20	24.08	626	642	657	670	680
	White	99,840	663.72	19.92	640	652	663	676	684
NRC	New York City	68,901	653.06	20.86	629	642	654	665	676
	Big 4 Cities	7,278	644.47	21.69	617	632	646	657	670
	High Needs Urban/Suburban	14,208	651.34	19.87	627	640	652	663	673
	High Needs Rural	11,581	656.25	19.12	632	646	657	667	680
	Average Needs	60,742	661.79	19.71	639	651	663	673	684
	Low Needs	32,134	669.00	19.82	648	657	667	680	690
	Charter	3,450	654.64	15.99	635	645	654	663	673
SWD	All Codes	29,640	637.49	19.03	613	627	639	649	659
SUA	All Codes	19,301	639.98	20.08	615	629	642	652	663
ELL	ELL = Y	10,702	628.78	19.53	604	617	631	642	651
SWD/SUA	SUA=504 plan codes	13,779	637.44	19.11	615	626	639	649	659
ELL/SUA	SUA=ELL codes	1,794	628.11	20.01	604	617	631	642	651

Performance Level Distribution Summary

Students are classified as Level I (Below Standards), Level II (Meets Basic Standards), Level III (Meets Proficiency Standards), and Level IV (Exceeds Proficiency Standards). The original proficiency cut scores used to distinguish among Levels I, II, III, and IV established during the process of Standard Setting in 2006 were adjusted after the 2010 OP test administration to reflect a change in the test administration window between the 2008–2009 and 2009–2010 school years and the State's policy decision to align the proficiency standards with Grade 8 student performance on the New York State Regents ELA examination. The theoretical cut scores established in 2010 were used as the cut scores for the 2012 administration.

Table 48 shows the ELA cut scores used for classification of students to the four performance level categories in 2012.

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

Grade	New York State Cut Scores "Operational Cuts"		
	Level		
	II	III	IV
3	644	663	694
4	637	671	722
5	648	668	700
6	644	662	694
7	642	665	698
8	628	658	699

Tables 49–55 show the performance level distribution 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 distributions reflect the same achievement trends in the scale score summary discussion. More Female students were classified in Level III and above categories than Male students. Similarly, more Asian and White students were classified in Level III and above categories than their peers from other ethnic groups. Consistently with the scale score distribution across group pattern, students from Low and Average Needs districts outperformed students from High Needs districts (New York City, Big 4 Cities, Urban/Suburban, and Rural). The Level III and above rates for students in the ELL, SWD, and SUA subgroups were low, compared to the total population of examinees. Across grades, the following subgroups consistently performed above the population average: Asian, White, Average Needs, and Low Needs.

Table 49. ELA Grades 3–8 Test Performance Level Distributions

Grade	N-count	Percentage of NYS Student Population in Performance Level				
		Level I	Level II	Level III	Level IV	Levels III & IV
3	197,993	13.40	30.85	48.96	6.79	55.75
4	194,344	8.95	31.45	54.97	4.63	59.60
5	196,623	10.37	31.73	53.17	4.73	57.90
6	199,540	10.28	33.74	53.46	2.51	55.97
7	197,638	8.01	39.30	48.98	3.71	52.68
8	198,294	6.99	42.30	48.87	1.83	50.70

Grade 3

Performance level distributions and N-counts of demographic groups for Grade 3 are presented in Table 50. Statewide, 55.75% of third-graders were Level III and Level IV. 16.08% of Male students were Level I, as compared to only 10.60% of Female students. The percentage of students in Levels III and IV varied widely by ethnicity and NRC subgroups. About 77% of Low Needs district students and about 70% of Asian students were classified in Levels III and IV; whereas the American Indian, Hispanic, Black, Charter, New York City, and/or Big 4 Cities had a range of 46%–70% of students who were in Level I or Level II. About one-third of students with ELL, SWD, or SUA status were in Level I, and fewer than 1% were in Level IV. The following groups had pass rates (percentage of students in Levels III & IV) above the state average: Female, Asian, Multiracial, White, Average Needs districts, and Low Needs districts.

Table 50. Performance Level Distribution Summary, by Subgroup, Grade 3

Demographic Category (Subgroup)		N-count	Level I %	Level II %	Level III %	Level IV %	Levels III & IV %
State	All Students	197,993	13.40	30.85	48.96	6.79	55.75
Gender	Female	96,905	10.60	29.21	51.84	8.35	60.19
	Male	101,088	16.08	32.43	46.20	5.29	51.49
Ethnicity	Asian	16,463	7.65	21.89	57.19	13.27	70.47
	Black	35,676	20.85	39.53	37.14	2.48	39.62
	Hispanic	47,112	19.96	38.47	38.84	2.73	41.56
	American Indian	1,081	16.56	36.36	43.02	4.07	47.09
	Multiracial	2,010	13.23	29.80	49.50	7.46	56.97
	Other	403	11.17	28.29	53.35	7.20	60.55
	White	95,248	8.33	25.36	57.01	9.31	66.31
NRC	New York City	71,611	16.38	34.54	43.52	5.56	49.08
	Big 4 Cities	8,022	32.72	37.37	28.20	1.71	29.91
	High Needs Urban/ Suburban	15,592	20.25	37.08	39.58	3.09	42.68
	High Needs Rural	11,018	15.85	36.19	44.34	3.63	47.97
	Average Needs	57,844	9.44	28.21	54.81	7.54	62.35
	Low Needs	27,939	4.46	18.31	63.35	13.87	77.23
	Charter	5,967	9.40	36.15	51.06	3.39	54.45
SWD	All Codes	28,201	44.44	36.72	17.90	0.94	18.84
SUA	All Codes	27,303	36.73	39.15	23.06	1.06	24.13
ELL	ELL=Y	17,112	38.46	43.00	18.27	0.26	18.53
SWD/ SUA	SUA=504 plan codes	14,459	49.98	35.80	13.67	0.55	14.23
ELL/SUA	SUA=ELL codes	7,380	35.89	44.34	19.43	0.34	19.77

Grade 4

Performance level distributions and N-counts of demographic groups for Grade 4 are presented in Table 51. Across New York, approximately 60% of fourth-grade students were in Levels III and IV. As was seen in Grade 3, the Low Needs subgroup had the highest percentage of students in Levels III and IV (79.45%), and the SWD/SUA subgroup had the lowest (16.58%). Students in the Black, Hispanic, and American Indian subgroups had percentages classified in Levels III and IV below 50%, which was more than 20% below the other ethnic subgroups. More than twice as many Big 4 Cities students were in Level I than the state population. About a third of the students with ELL, SWD, or SUA status were in Level I (over three times the statewide rate of 8.95%) and fewer than 1% were in Level IV. The following groups had percentages of students classified in Levels III and IV, above the state average: Female, Asian, Multiracial, White, Average Needs districts, and Low Needs districts.

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	194,344	8.95	31.45	54.97	4.63	59.60
Gender	Female	95,398	6.63	29.13	58.39	5.85	64.24
	Male	98,946	11.19	33.68	51.66	3.47	55.13
Ethnicity	Asian	16,144	5.46	20.22	63.94	10.38	74.32
	Black	35,445	13.85	42.87	41.57	1.71	43.28
	Hispanic	45,518	13.16	41.75	43.46	1.63	45.08
	American Indian	1,035	11.88	39.32	46.76	2.03	48.79
	Multiracial	1,696	8.96	30.37	56.66	4.01	60.67
	Other	316	5.70	26.90	60.76	6.65	67.41
	White	94,190	5.65	24.04	64.07	6.24	70.31
NRC	New York City	69,619	10.57	36.91	48.46	4.06	52.52
	Big 4 Cities	7,992	24.74	44.54	29.78	0.94	30.72
	High Needs Urban/ Suburban	15,193	14.33	39.10	44.88	1.69	46.57
	High Needs Rural	11,026	11.14	34.72	51.66	2.49	54.14
	Average Needs	57,535	6.34	26.85	61.87	4.94	66.81
	Low Needs	28,193	2.76	17.80	70.29	9.15	79.45
	Charter	4,786	4.89	33.93	58.02	3.16	61.18
SWD	All Codes	29,610	34.43	45.39	19.79	0.39	20.18
SUA	All Codes	23,132	28.82	46.01	24.76	0.41	25.17
ELL	ELL=Y	15,929	29.66	52.05	18.19	0.09	18.29
SWD/ SUA	SUA=504 plan codes	13,764	38.27	45.15	16.41	0.17	16.58
ELL/SUA	SUA=ELL codes	4,819	27.06	53.23	19.61	0.10	19.71

Grade 5

Performance level distributions and N-counts of demographic groups for Grade 5 are presented in Table 52. About 57.90% of the Grade 5 students were in Levels III and IV. As was seen in Grades 3 and 4, the Low Needs subgroup had the highest percentage of students in Levels III and IV (79.47%). Students in the American Indian, Black, and Hispanic subgroups had rates less than 45% of students classified in Levels III and IV, approximately 20% less than other ethnic subgroups. Over two times as many Big 4 Cities students were in Level I than the State population. About 33–39% of the students with ELL, SWD, or SUA status were in Level I (approximately three times as many as the statewide rate of 10.37%), yet only about 13–22% were in Levels III and IV (combined) and a very low percentage (less than 1%) in Level IV. The following groups had percentages of students classified in Levels III and IV, above the state average: Female, Asian, Multiracial, White, Average Needs districts, and Low Needs districts.

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	196,623	10.37	31.73	53.17	4.73	57.90
Gender	Female	96,487	8.33	29.85	55.80	6.02	61.82
	Male	100,136	12.34	33.54	50.63	3.49	54.12
Ethnicity	Asian	15,777	6.74	19.57	63.19	10.50	73.69
	Black	36,584	16.77	42.59	39.02	1.62	40.64
	Hispanic	45,248	14.64	40.57	42.90	1.89	44.79
	American Indian	992	13.71	42.34	41.03	2.92	43.95
	Multiracial	1,541	10.25	33.87	49.64	6.23	55.87
	Other	317	7.57	29.97	57.41	5.05	62.46
	White	96,164	6.49	25.30	61.91	6.30	68.21
NRC	New York City	68,552	11.95	35.65	47.98	4.41	52.40
	Big 4 Cities	7,965	28.70	41.19	28.83	1.28	30.11
	High Needs Urban/ Suburban	14,975	15.97	40.07	42.20	1.76	43.96
	High Needs Rural	11,387	12.19	36.83	48.72	2.26	50.98
	Average Needs	59,075	7.57	28.86	58.94	4.63	63.57
	Low Needs	28,658	3.44	17.09	69.62	9.85	79.47
	Charter	6,011	11.15	42.09	45.08	1.68	46.76
SWD	All Codes	30,703	37.96	44.24	17.48	0.33	17.80
SUA	All Codes	25,237	33.32	44.68	21.58	0.42	22.00
ELL	ELL=Y	13,537	38.68	48.01	13.25	0.06	13.31
SWD/ SUA	SUA=504 plan codes	16,158	41.52	43.76	14.51	0.21	14.72
ELL/SUA	SUA=ELL codes	4,001	40.16	46.19	13.60	0.05	13.65

Grade 6

Performance level distributions and N-counts of demographic groups for Grade 6 are presented in Table 53. Statewide, 55.97% of Grade 6 students were classified in Levels III and IV. As was seen in other grades, the Low Needs subgroup had the most students classified in these two proficiency levels (78.74%), and the ELL, SWD, and SUA subgroups had the fewest. Students in the American Indian, Black, and Hispanic subgroups had about 38–45% of students classified in Levels III and IV. Students from Low Needs districts outperformed students in all other subgroups, across demographic categories as in the previous grades. The majority of students with ELL, SWD, and/or SUA status were in Level II, but fewer than 1% were in Level IV. The following groups had percentages of students classified in Levels III and IV, above the state average: Female, Asian, Multiracial, White, Average Needs districts, and Low Needs districts.

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	199,540	10.28	33.74	53.46	2.51	55.97
Gender	Female	97,552	7.65	31.58	57.43	3.35	60.77
	Male	101,988	12.81	35.81	49.67	1.71	51.38
Ethnicity	Asian	16,679	7.37	23.66	63.74	5.22	68.96
	Black	37,110	16.04	46.43	36.89	0.63	37.53
	Hispanic	44,565	16.44	44.39	38.49	0.68	39.17
	American Indian	1,004	14.54	40.74	43.73	1.00	44.72
	Multiracial	1,448	7.46	31.15	57.46	3.94	61.40
	Other	347	12.10	28.82	53.03	6.05	59.08
	White	98,387	5.81	25.83	64.80	3.57	68.37
NRC	New York City	68,954	13.98	40.61	43.70	1.72	45.42
	Big 4 Cites	7,758	24.74	44.03	30.83	0.40	31.23
	High Needs Urban/ Suburban	14,741	15.16	41.82	41.99	1.03	43.02
	High Needs Rural	11,423	10.92	38.00	49.58	1.51	51.08
	Average Needs	60,856	6.74	28.62	61.80	2.84	64.63
	Low Needs	30,289	2.90	18.36	73.10	5.64	78.74
	Charter	5,519	9.08	43.98	46.33	0.62	46.95
SWD	All Codes	30,429	37.51	47.44	14.96	0.10	15.06
SUA	All Codes	22,594	33.65	48.00	18.20	0.15	18.35
ELL	ELL=Y	11,307	50.51	44.09	5.40	0.00	5.40
SWD/ SUA	SUA=504 plan codes	14,552	39.73	47.01	13.20	0.05	13.26
ELL/SUA	SUA=ELL codes	3,383	50.69	43.54	5.76	0.00	5.76

Grade 7

Performance level distributions and N-counts of demographic groups for Grade 7 are presented in Table 54. In Grade 7, 52.68% of the students were in Levels III and IV. Over 10% more Female than Male students were classified in these two proficiency levels. Close to 76% of Big 4 Cities students were in Levels I and II. About 76% of Low Needs students were in Levels III and IV. About 4% of ELL students were in Levels III and IV. The ELL, SWD, and SUA subgroups were well below the performance achievement of the general population, with around 84–96% of those students in Levels I and II. The following subgroups had percentages of students in Levels III and IV, above the general population: Female, Asian, Multiracial, White, Average Needs districts, and Low Needs districts.

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	197,638	8.01	39.30	48.98	3.71	52.68
Gender	Female	96,364	5.62	36.17	53.59	4.62	58.21
	Male	101,274	10.29	42.29	44.59	2.84	47.42
Ethnicity	Asian	15,534	6.26	25.38	60.70	7.66	68.36
	Black	37,089	12.58	53.06	33.49	0.87	34.36
	Hispanic	43,499	12.51	51.20	35.26	1.03	36.29
	American Indian	1,011	10.29	49.85	38.58	1.29	39.86
	Multiracial	1,379	6.16	40.46	48.66	4.71	53.37
	Other	324	6.79	41.98	47.22	4.01	51.23
	White	98,802	4.60	30.96	59.10	5.34	64.44
NRC	New York City	67,421	10.55	45.97	41.02	2.46	43.48
	Big 4 Cities	7,598	22.72	53.92	22.57	0.79	23.36
	High Needs Urban/ Suburban	14,572	12.73	48.59	37.28	1.39	38.68
	High Needs Rural	11,705	8.79	44.87	43.96	2.38	46.34
	Average Needs	60,119	5.38	34.55	55.73	4.34	60.07
	Low Needs	31,553	2.11	22.39	67.67	7.83	75.50
	Charter	4,670	4.60	51.86	42.53	1.01	43.53
SWD	All Codes	29,844	32.09	56.12	11.65	0.14	11.79
SUA	All Codes	22,581	27.76	56.19	15.88	0.17	16.05
ELL	ELL=Y	10,802	42.03	53.48	4.48	0.01	4.49
SWD/ SUA	SUA=504 plan codes	15,223	32.80	55.92	11.19	0.09	11.28
ELL/SUA	SUA=ELL codes	2,964	40.89	54.18	4.89	0.03	4.93

Grade 8

Performance level distributions and N-counts of demographic groups for Grade 8 are presented in Table 55. In Grade 8, 50.70% of the students were in Levels III and IV. About 11% more Female students than Male students were in Levels III or IV. Over 60% of American Indian, Black, and Hispanic students were in Levels I and II. Over 74% of Low Needs students were in Levels III and IV, while no ELL students were in Level IV. The ELL, SWD, and SUA subgroups were well below the performance achievement of the general population, with over 84% of those students in Levels I and II. The following subgroups had a higher percentage of students in Levels III and IV than the general population: Female, Asian, Multiracial, White, Average Needs districts, and Low Needs districts.

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	198,294	6.99	42.30	48.87	1.83	50.70
Gender	Female	97,523	5.05	38.62	53.74	2.59	56.33
	Male	100,771	8.87	45.88	44.16	1.09	45.25
Ethnicity	Asian	15,838	6.60	27.76	61.22	4.42	65.64
	Black	36,955	10.72	59.25	29.66	0.36	30.03
	Hispanic	43,156	11.44	54.95	33.12	0.49	33.61
	American Indian	1,029	9.52	50.73	38.58	1.17	39.75
	Multiracial	1,143	5.42	38.93	52.67	2.97	55.64
	Other	333	12.31	42.04	44.74	0.90	45.65
	White	99,840	3.73	32.82	60.91	2.54	63.45
NRC	New York City	68,901	9.70	51.12	38.00	1.18	39.18
	Big 4 Cities	7,278	19.46	55.94	24.10	0.51	24.61
	High Needs Urban/ Suburban	14,208	10.91	52.76	35.72	0.61	36.33
	High Needs Rural	11,581	6.93	45.71	46.35	1.00	47.35
	Average Needs	60,742	4.36	36.47	57.13	2.05	59.17
	Low Needs	32,134	1.96	23.97	69.99	4.08	74.06
	Charter	3,450	4.06	56.49	38.93	0.52	39.45
SWD	All Codes	29,640	27.38	61.63	10.93	0.06	11.00
SUA	All Codes	19,301	24.80	59.17	15.90	0.13	16.03
ELL	ELL=Y	10,702	43.26	53.80	2.93	0.00	2.93
SWD/ SUA	SUA=504 plan codes	13,779	27.95	60.64	11.36	0.05	11.41
ELL/SUA	SUA=ELL codes	1,794	45.37	51.28	3.34	0.00	3.34

Section IX: Longitudinal Comparison of Results

This section provides longitudinal comparison of OP scale score results on the NYSTP 2006–2012 Grades 3–8 ELA Tests. These include the scale score means, standard deviations, and performance level distributions for each grade's public and charter school population. The longitudinal results are presented in Table 56.

Table 56. ELA Grades 3–8 Test Longitudinal Results

Grade	Year	N-Count	Scale Score Mean	Standard Deviation	Percentage of Students in Performance Levels				
					Level I	Level II	Level III	Level IV	Level III & IV
3	2012	197,993	664.46	21.11	13.40	30.85	48.96	6.79	55.75
	2011	196,476	663.47	21.19	11.52	32.50	51.39	4.59	55.98
	2010	196,425	667.90	33.09	13.77	31.47	38.11	16.66	54.77
	2009	198,123	669.97	35.81	4.75	19.37	65.17	10.72	75.89
	2008	195,231	669.00	39.41	5.84	23.92	57.84	12.40	70.24
	2007	198,320	666.99	42.23	8.92	23.89	57.29	9.90	67.20
	2006	185,533	668.79	40.91	8.53	22.47	61.92	7.07	69.00
4	2012	194,344	674.54	29.97	8.95	31.45	54.97	4.63	59.60
	2011	197,040	671.84	28.98	8.20	31.96	57.38	2.46	59.84
	2010	199,254	672.82	29.50	8.34	34.82	50.87	5.97	56.84
	2009	195,634	669.93	34.72	4.28	18.76	69.69	7.27	76.96
	2008	196,367	666.40	39.90	7.34	21.37	62.85	8.44	71.29
	2007	197,306	664.70	39.52	7.79	24.17	59.82	8.22	68.04
	2006	190,847	665.73	40.74	8.92	22.40	59.94	8.74	68.68
5	2012	196,623	669.93	18.29	10.37	31.73	53.17	4.73	57.90
	2011	200,195	667.82	19.47	10.41	31.78	53.39	4.41	57.80
	2010	197,200	672.41	32.09	11.54	35.90	39.71	12.85	52.56
	2009	197,522	675.47	34.58	0.62	17.09	68.72	13.57	82.29
	2008	197,318	667.35	30.89	1.78	20.45	71.83	5.94	77.77
	2007	201,841	665.39	37.98	4.89	26.88	61.37	6.86	68.24
	2006	201,138	662.69	41.17	6.38	26.45	54.86	12.31	67.17
6	2012	199,540	663.18	16.23	10.28	33.74	53.46	2.51	55.97
	2011	198,076	662.62	18.11	11.55	32.55	51.93	3.96	55.89
	2010	197,845	664.48	24.67	11.30	34.44	47.40	6.85	54.26
	2009	197,674	667.31	27.64	0.13	18.87	71.98	9.02	81.00
	2008	199,689	661.45	30.03	1.63	31.20	62.49	4.68	67.17
	2007	204,237	661.47	33.98	2.46	34.22	53.93	9.40	63.32
	2006	204,104	656.52	40.85	7.28	32.24	48.88	11.60	60.48

Table 56. ELA Grades 3–8 Test 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	197,638	665.35	18.44	8.01	39.30	48.98	3.71	52.68
	2011	200,140	663.71	19.60	9.23	39.37	47.82	3.58	51.40
	2010	199,943	667.91	31.29	10.35	39.53	38.94	11.18	50.12
	2009	202,400	667.19	27.06	0.42	19.15	73.51	6.91	80.42
	2008	205,946	662.30	29.29	1.75	27.90	67.79	2.56	70.35
	2007	211,545	654.84	38.23	5.90	36.22	51.91	5.98	57.89
	2006	210,518	652.29	40.95	8.03	35.55	48.66	7.76	56.42
8	2012	198,294	658.09	21.16	6.99	42.30	48.87	1.83	50.70
	2011	201,278	655.28	22.15	7.47	45.50	45.23	1.80	47.03
	2010	204,080	659.07	31.11	8.95	39.98	43.37	7.70	51.06
	2009	207,083	661.09	30.82	1.72	29.66	63.75	4.87	68.62
	2008	207,646	657.26	37.66	4.95	38.53	50.80	5.73	56.53
	2007	213,676	655.39	39.32	6.12	36.75	51.45	5.68	57.13
	2006	212,138	650.14	40.78	9.42	41.20	44.53	4.84	49.38

It should be noted, however, that although the ELA scales were maintained between the 2006 and 2012 administrations and the scale scores from the 2006–2012 administrations can be directly compared, the performance level results between 2006–2009 OP tests 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 ELA grades except Grades 3 and 5 between the 2006 and 2010 test administrations. The Grade 3 mean scale score dropped 1 scale score point in 2010, and the Grade 5 mean scale score dropped 3 scale score points in 2010. In 2012, the mean scale score for all grades increased from over one-half scale score point for Grade 6 to about 3 scale score points for Grades 4 and 8. Moderate gains were observed for Grades 4, 5, 6, and 8 for which total gains were 6 or 9 scale score points between the 2006 and 2012 test administrations. The largest gain in scale score points between the 2006 and 2012 test administrations was noted for Grade 7 (13 scale score points). Grade 3 dropped more than 4 scale score points between the 2006 and 2012 test administrations. A relatively steady yearly gain was noticed for Grade 7 with an overall population mean scale score increase of 16 scale points between 2006 and 2010, and then the mean scale score dropped about 4 scale score points in 2011. For Grades 3 and 4, a slight mean scale score decline (1 to 2 scale score points) was observed between 2006 and 2007, and a small increase (approximately 2 points) was observed for years 2007 and 2008. The following was noted for Grades 3 and 4: a small increase (approximately 2 points) for Grade 3 and a moderate increase (4 points) for Grade 4 between 2008 and 2009; a slight decline (2 points) for Grade 3 and a moderate increase (3 points) for Grade 4 between 2009 and 2010; a moderate decline (approximately 4 points) for Grade 3 and a slight decline (1 point) for Grade 4 between 2010 and 2011; and a slight gain for Grade 3 (1

point) and Grade 4 (2 points) between 2011 and 2012. A relatively steady yearly gain was noticed for Grades 5 and 8, with the overall population mean scale score increases of 13 and 11 scale score points respectively between 2006 and 2009; a slight decline (2–3 scale score points) between 2009 and 2010; a moderate decline (approximately 4 points) between 2010 and 2011; and then a slight gain (2 points) between 2011 and 2012. For Grade 6, an increase of approximately 5 scale score points was observed between 2006 and 2007, no score change was noticed between 2007 and 2008, but a 6 scale score point increase was observed between 2008 and 2009. A moderate mean scale score decline (3 scale score points) was observed between 2009 and 2010, a slight decline (approximately 2 points) between 2010 and 2011, and then a slight gain (0.5 point) between 2011 and 2012.

The variability of scale score distribution decreased steadily across years for ELA Grade 6. The scale score standard deviation was around 40 scale score points in 2006 and dropped to around 16 scale score points in 2012. For Grades 3 and 4, the variability of scale score distribution decreased in 2009, 2010, and 2011. The standard deviations for these grades decreased from about 40 scale score points in 2006, 2007, and 2008, to approximately 35 points in 2009, then to 33 and 30 scale score points in 2010, and then to 21 and 29 scale score points in 2012. The standard deviation for Grade 5 decreased from approximately 40 scale score points in 2006 to about 31 scale score points in 2008, then increased to approximately 35 scale score points in 2009; then it decreased to 32 scale score points in 2010 and to 18 scale score points in 2012. The variability of scale score distribution decreased steadily across years for ELA Grades 7 and 8 between 2006 and 2009. The scale score standard deviation was around 40 scale score points for these grades in 2006; dropped to around 30 scale score points in 2009; it increased to approximately 31 scale score points in 2010 and then decreased to 18 and 21 scale score points in 2012.

Appendix A—ELA Passage Specifications

General Guidelines

- Each passage must have a clear beginning, middle, and end.
- Passages may be excerpted from larger works, but internal editing must be avoided. No edits may be made to poems.
- Passages should be age- and grade-appropriate and should contain subject matter of interest to the students being tested.
- Informational passages should span a broad range of topics, including history, science, careers, career training, etc.
- Literary passages should span a variety of genres and should include both classic and contemporary literature.
- Material may be selected from books, magazines (such as *Cricket*, *Cobblestone*, *Odyssey*, *National Geographic World*, and *Sports Illustrated for Kids*), and newspapers.
- Avoid selecting literature that is widely studied. To that end, do not select passages from basals.
- If the accompanying art is not integral to the passage, and if permissions are granted separately, you may choose not to use that art or to use different art.
- Illustration- or photograph-dependent passages should be avoided whenever possible.
- Passages should bring a range of cultural diversity to the tests. They should be written by, as well as about, people of different cultures and races.
- Passages should be suitable for items to be written that test the performance indicators as outlined in the New York State Learning Standards Core Curricula.
- Passages (excluding poetry) should be analyzed for readability. Readability statistics are useful in helping to determine grade-level appropriateness of text prior to presenting the passages for formal committee review. An overview of the readability concept for passages selected for the 2011 OP administration is provided below.

Use of Readability Formulae in New York State Assessments

A variety of readability formulae currently exist that can be used to help determine the readability level of text. The formulae most associated with the K–12 environment are the Dale-Chall, the Fry, and the Spache formulae. Others (such as Flesch-Kincaid) are more associated with general text (such as newspapers and mainstream publications).

Readability formulae provide some useful information about the reading difficulty of a passage or stimulus. However, it should be noted that a readability score is not the most reliable indicator of grade-level appropriateness and, therefore, should not be the sole determinant of whether a particular passage or stimulus should be included in assessment or instructional materials.

Readability formulae are quantitative measures that assess the surface characteristics of text (e.g., the number of letters or syllables in a word, the number of words in a sentence, the number of sentences in a paragraph, the length of the passage). In order to truly measure the readability

of any text, qualitative factors (e.g., density of concepts, organization of text, coherence of ideas, level of student interest, and quality of writing) must also be considered.

One basic drawback to the usability of readability formulae is that not all passage or stimulus formats can be processed. To produce a score, the formulae generally require a minimum of 100 words in a sample (for Flesch Reading Ease and the Flesch-Kincaid, 200-word samples are recommended). This requirement renders the readability formulae essentially unusable for passages such as poems and many functional documents. Another drawback is evident in passages with specialized vocabulary. For example, if a passage contains scientific terminology, the readability score might appear to be above grade-level, even though the terms might be footnoted or explained within the context of the passage.

In light of the drawbacks that exist in the use of readability formulae, rather than relying solely on readability indices, Pearson relies on the expertise of the educators in the State of New York to help determine the suitability of passages and stimuli to be used in statewide assessments. Prospective passages are submitted for review to panels of New York State educators familiar with the abilities of the students to be tested and with the grade-level curricula. The passages are reviewed for readability, appropriateness of content, potential interest level, quality of writing, and other qualitative features that cannot be measured via readability formulae.

Appendix B—Criteria for Item Acceptability

For Multiple-Choice Items:

Check that the content of each item

- is targeted to assess only one objective or skill (unless specifications indicate otherwise)
- deals with material that is important in testing the targeted performance indicator
- uses grade-appropriate content and thinking skills
- is presented at a reading level suitable for the grade level being tested
- has a stem that facilitates answering the question or completing the statement without looking at the answer choices
- has a stem that does **not** present clues to the correct answer choice
- has answer choices that are plausible and attractive to the student who has not mastered the objective or skill
- has mutually exclusive distractors
- has one and only one correct answer choice
- is free of cultural, racial, ethnic, age, gender, disability, regional, or other apparent bias

Check that the format of each item

- is worded in the positive unless it is absolutely necessary to use the negative form
- is free of extraneous words or expressions in both the stem and the answer choices (e.g., the same word or phrase does not begin each answer choice)
- indicates emphasis on key words, such as best, first, least, not, and others, that are important and might be overlooked
- places the interrogative word at the **beginning** of a stem in the form of a question or places the omitted portion of an incomplete statement at the **end** of the statement
- indicates the correct answer choice
- provides the rationale for all distractors
- is conceptually, grammatically, and syntactically consistent—between the stem and answer choices, and among the answer choices
- has answer choices balanced in length or contains two long and two short choices
- clearly identifies the passage or other stimulus material associated with the item
- clearly identifies a need of art, if applicable, and the art is conceptualized and sketched, with important considerations explicated

Also check that

- one item does not present clues to the correct answer choice for any other item
- any item based on a passage is answerable from the information given in the passage and is not dependent on skills related to other content areas
- any item based on a passage is truly passage-dependent; that is, **not** answerable without reference to the passage
- there is a balance of reasonable, non-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 can be scored with the specified rubric or scoring tool; that is, the range of possible correct responses must be wide enough to allow for diversity of responses, but narrow enough so that students who do not clearly show their grasp of the objective or skill being assessed cannot obtain the maximum score
- presented without clues to the correct response
- checked for accuracy and documented against reliable, up-to-date sources (including rubrics)
- free of cultural, racial, ethnic, age, gender, disability, or other apparent bias

Check that the format of each item is

- appropriate for the question being asked and the intended response
- worded clearly and concisely, using simple vocabulary and sentence structure
- precise and unambiguous in its directions for the desired response
- free of extraneous words or expressions
- worded in the positive rather than in the negative form
- conceptually, grammatically, and syntactically consistent
- marked with emphasis on key words, such as best, first, least, and others that are important and might be overlooked
- clearly identified as needing art, if applicable, and the art is conceptualized and sketched, with important considerations explicated

Also check that

- one item does not present clues to the correct response to any other item
- there is balance of reasonable, non-stereotypic representation of economic classes, races, cultures, ages, genders, and persons with disabilities in context and art
- for each set of items related to a reading passage, each item is designed to elicit a unique and independent response
- items designed to assess reading do not depend on prior knowledge of the subject matter used in the prompt/question

Appendix C—Psychometric Guidelines for Operational Item Selection

It is primarily up to the content development department to select items for the 2012 OP test. Research will provide support, as necessary, and will review the final item selection. Research will provide data files with parameters for all FT items eligible for the item pool. The pools of items eligible for 2012 item selection included 2011 FT items and items owned by Pearson and the items field-tested in New York State in 2011. All items for each grade will be on the same (grade-specific) scale.

Here are general guidelines for item selection:

- Satisfy the content specifications in terms of objective coverage and the number and percentage of MC and CR items on the test. An often-used criterion for objective coverage is within 5% of the percentages of score points and items per objective.
- Avoid selecting poor-fitting items, items with too high/low p-values, items with flagged point biserials (The research department will provide a list of such items.).
- Avoid items flagged for local dependency if the flagged items come from different passages. If the flagged items come from the same passage, they are expected to be dependent on each other to some degree and are not a problem.
- Minimize the number of items flagged for DIF (gender, ethnic, and High/Low Needs schools). Flagged items should be reviewed for content again. It needs to be remembered that some items may be flagged for DIF by chance only and their content may not necessarily be biased against any of the analyzed groups. Research will provide DIF information for each item. It is also possible to get "significant" DIF, but 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 third of the section of the 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 the research department with the following item selection information:
    - Percentage of score points per learning standard (target, 2011 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.)
    - ITEMWIN output (including TCCs)
    - Summary file with IRT item parameters for selected items

Appendix D—Factor Analysis Results

As described in Section III, "Validity," a principal component factor analysis was conducted on the Grades 3–8 ELA Tests data. The analyses were conducted for the total population of students and selected subpopulations: English language learners (ELL), students with disabilities (SWD), students using accommodations (SUA), SWD students using disability accommodations (SWD/SUA) and ELL students using ELL-related accommodations (ELL/SUA). Table D1 contains the results of factor analysis on subpopulation data.

Table D1. Factor Analysis Results for ELA Tests (Selected Subpopulations)

		Initial Eigenvalues			
Grade	Subgroups	Component	Total	% of Variance	Cumulative %
3	ELL	**1**	**9.90**	**17.67**	**17.67**
		2	1.60	2.86	20.53
		3	1.27	2.27	22.80
		4	1.14	2.04	24.84
		5	1.13	2.01	26.85
		6	1.07	1.91	28.76
		7	1.04	1.86	30.62
		8	1.03	1.84	32.46
		9	1.00	1.79	34.25
	SWD	**1**	**11.23**	**20.05**	**20.05**
		2	1.74	3.10	23.15
		3	1.39	2.48	25.63
		4	1.16	2.07	27.70
		5	1.10	1.96	29.66
		6	1.04	1.85	31.51
		7	1.01	1.80	33.31
	SUA	**1**	**10.82**	**19.32**	**19.32**
		2	1.73	3.09	22.41
		3	1.38	2.46	24.87
		4	1.20	2.14	27.01
		5	1.11	1.98	28.99
		6	1.05	1.87	30.86

Table D1. Factor Analysis Results for ELA Tests (Selected Subpopulations) (cont.)

		Initial Eigenvalues			
Grade	Subgroups	Component	Total	% of Variance	Cumulative %
3	SUA	7	1.01	1.81	32.67
		8	1.01	1.80	34.47
	SWD/SUA	**1**	**10.60**	**18.93**	**18.93**
		2	1.74	3.11	22.04
		3	1.39	2.48	24.52
		4	1.20	2.14	26.66
		5	1.11	1.99	28.65
		6	1.06	1.89	30.54
		7	1.02	1.82	32.36
		8	1.01	1.80	34.16
	ELL/SUA	**1**	**1.62**	**2.89**	**20.15**
		2	1.25	2.23	22.38
		3	1.20	2.13	24.51
		4	1.14	2.04	26.55
		5	1.08	1.93	28.48
		6	1.07	1.91	30.39
		7	1.04	1.86	32.25
		8	1.03	1.83	34.08
		9	1.01	1.80	35.88
		10	1.62	2.89	20.15
4	ELL	**1**	**8.86**	**14.76**	**14.76**
		2	1.57	2.62	17.38
		3	1.22	2.04	19.42
		4	1.14	1.90	21.32
		5	1.09	1.82	23.14
		6	1.09	1.81	24.95
		7	1.06	1.76	26.71
		8	1.05	1.74	28.45

Table D1. Factor Analysis Results for ELA Tests (Selected Subpopulations) (cont.)

		Initial Eigenvalues			
Grade	Subgroups	Component	Total	% of Variance	Cumulative %
4	ELL	9	1.03	1.71	30.16
		10	1.02	1.71	31.87
		11	1.00	1.67	33.54
	SWD	**1**	**10.21**	**17.02**	**17.02**
		2	1.63	2.71	19.73
		3	1.22	2.03	21.76
		4	1.14	1.90	23.66
		5	1.10	1.83	25.49
		6	1.07	1.78	27.27
		7	1.04	1.73	29.00
		8	1.01	1.69	30.69
		9	1.00	1.67	32.36
	SUA	**1**	**10.05**	**16.75**	**16.75**
		2	1.59	2.66	19.41
		3	1.24	2.07	21.48
		4	1.14	1.90	23.38
		5	1.10	1.83	25.21
		6	1.07	1.78	26.99
		7	1.03	1.72	28.71
		8	1.02	1.70	30.41
		9	1.01	1.68	32.09
	SWD/SUA	**1**	**9.67**	**16.12**	**16.12**
		2	1.63	2.72	18.84
		3	1.24	2.07	20.91
		4	1.15	1.91	22.82
		5	1.11	1.85	24.67
		6	1.07	1.79	26.46
		7	1.05	1.75	28.21

Table D1. Factor Analysis Results for ELA Tests (Selected Subpopulations) (cont.)

		Initial Eigenvalues			
Grade	Subgroups	Component	Total	% of Variance	Cumulative %
4	SWD/SUA	8	1.03	1.71	29.92
		9	1.02	1.69	31.61
	ELL/SUA	**1**	**8.64**	**14.40**	**14.40**
		2	1.60	2.66	17.06
		3	1.28	2.13	19.19
		4	1.15	1.92	21.11
		5	1.13	1.89	23.00
		6	1.11	1.86	24.86
		7	1.10	1.84	26.70
		8	1.10	1.83	28.53
		9	1.07	1.78	30.31
		10	1.06	1.77	32.08
		11	1.05	1.75	33.83
		12	1.03	1.71	35.54
		13	1.02	1.69	37.23
		14	1.01	1.68	38.91
5	ELL	**1**	**8.72**	**14.53**	**14.53**
		2	1.42	2.36	16.89
		3	1.30	2.17	19.06
		4	1.25	2.08	21.14
		5	1.14	1.90	23.04
		6	1.11	1.85	24.89
		7	1.06	1.77	26.66
		8	1.06	1.76	28.42
		9	1.04	1.73	30.15
		10	1.02	1.71	31.86
		11	1.02	1.70	33.56

Table D1. Factor Analysis Results for ELA Tests (Selected Subpopulations) (cont.)

		Initial Eigenvalues			
Grade	Subgroups	Component	Total	% of Variance	Cumulative %
5	SWD	**1**	**9.85**	**16.42**	**16.42**
		2	1.44	2.40	18.82
		3	1.38	2.30	21.12
		4	1.25	2.08	23.20
		5	1.11	1.84	25.04
		6	1.07	1.78	26.82
		7	1.04	1.73	28.55
		8	1.02	1.69	30.24
		9	1.01	1.68	31.92
	SUA	**1**	**9.88**	**16.47**	**16.47**
		2	1.44	2.40	18.87
		3	1.38	2.30	21.17
		4	1.25	2.09	23.26
		5	1.10	1.83	25.09
		6	1.07	1.79	26.88
		7	1.04	1.73	28.61
		8	1.01	1.69	30.30
		9	1.00	1.67	31.97
	SWD/SUA	**1**	**9.50**	**15.84**	**15.84**
		2	1.45	2.42	18.26
		3	1.40	2.33	20.59
		4	1.26	2.10	22.69
		5	1.12	1.86	24.55
		6	1.08	1.80	26.35
		7	1.05	1.75	28.10
		8	1.02	1.70	29.80
		9	1.01	1.69	31.49

Table D1. Factor Analysis Results for ELA Tests (Selected Subpopulations) (cont.)

		Initial Eigenvalues			
Grade	Subgroups	Component	Total	% of Variance	Cumulative %
5	SWD/SUA	10	1.01	1.68	33.17
	ELL/SUA	**1**	**8.74**	**14.57**	**14.57**
		2	1.45	2.41	16.98
		3	1.38	2.30	19.28
		4	1.26	2.10	21.38
		5	1.16	1.93	23.31
		6	1.13	1.88	25.19
		7	1.11	1.84	27.03
		8	1.08	1.80	28.83
		9	1.08	1.80	30.63
		10	1.06	1.77	32.40
		11	1.05	1.76	34.16
		12	1.03	1.71	35.87
		13	1.02	1.70	37.57
		14	1.01	1.68	39.25
6	ELL	**1**	**6.92**	**11.53**	**11.53**
		2	1.57	2.62	14.15
		3	1.24	2.06	16.21
		4	1.18	1.96	18.17
		5	1.15	1.91	20.08
		6	1.12	1.87	21.95
		7	1.11	1.85	23.80
		8	1.09	1.82	25.62
		9	1.07	1.79	27.41
		10	1.06	1.76	29.17
		11	1.05	1.75	30.92
		12	1.04	1.73	32.65
		13	1.03	1.72	34.37

Table D1. Factor Analysis Results for ELA Tests (Selected Subpopulations) (cont.)

		Initial Eigenvalues			
Grade	Subgroups	Component	Total	% of Variance	Cumulative %
6	ELL	14	1.02	1.70	36.07
		15	1.01	1.69	37.76
		16	1.01	1.68	39.44
		17	1.00	1.67	41.11
	SWD	**1**	**8.15**	**13.58**	**13.58**
		2	1.72	2.87	16.45
		3	1.27	2.12	18.57
		4	1.20	2.00	20.57
		5	1.11	1.85	22.42
		6	1.07	1.79	24.21
		7	1.06	1.77	25.98
		8	1.03	1.71	27.69
		9	1.02	1.70	29.39
		10	1.01	1.68	31.07
	SUA	**1**	**8.30**	**13.84**	**13.84**
		2	1.71	2.85	16.69
		3	1.26	2.10	18.79
		4	1.21	2.02	20.81
		5	1.11	1.85	22.66
		6	1.08	1.81	24.47
		7	1.05	1.75	26.22
		8	1.02	1.70	27.92
		9	1.01	1.69	29.61
		10	1.01	1.68	31.29
	SWD/SUA	**1**	**7.81**	**13.01**	**13.01**
		2	1.67	2.79	15.80
		3	1.28	2.13	17.93
		4	1.22	2.03	19.96

Table D1. Factor Analysis Results for ELA Tests (Selected Subpopulations) (cont.)

		Initial Eigenvalues			
Grade	Subgroups	Component	Total	% of Variance	Cumulative %
6	SWD/SUA	5	1.13	1.88	21.84
		6	1.10	1.83	23.67
		7	1.06	1.77	25.44
		8	1.04	1.73	27.17
		9	1.03	1.71	28.88
		10	1.02	1.69	30.57
		11	1.00	1.67	32.24
	ELL/SUA	**1**	**7.04**	**11.73**	**11.73**
		2	1.57	2.62	14.35
		3	1.29	2.15	16.50
		4	1.23	2.05	18.55
		5	1.21	2.02	20.57
		6	1.17	1.95	22.52
		7	1.14	1.90	24.42
		8	1.13	1.89	26.31
		9	1.13	1.88	28.19
		10	1.11	1.85	30.04
		11	1.10	1.83	31.87
		12	1.09	1.82	33.69
		13	1.07	1.78	35.47
		14	1.06	1.77	37.24
		15	1.05	1.75	38.99
		16	1.05	1.74	40.73
		17	1.03	1.72	42.45
		18	1.03	1.71	44.16
		19	1.01	1.69	45.85
7	ELL	**1**	**8.69**	**14.49**	**14.49**
		2	1.67	2.78	17.27

Table D1. Factor Analysis Results for ELA Tests (Selected Subpopulations) (cont.)

		Initial Eigenvalues			
Grade	Subgroups	Component	Total	% of Variance	Cumulative %
7	ELL	3	1.30	2.17	19.44
		4	1.19	1.99	21.43
		5	1.18	1.96	23.39
		6	1.09	1.81	25.20
		7	1.07	1.79	26.99
		8	1.07	1.78	28.77
		9	1.06	1.77	30.54
		10	1.05	1.75	32.29
		11	1.03	1.71	34.00
		12	1.01	1.69	35.69
	SWD	**1**	**9.67**	**16.12**	**16.12**
		2	1.69	2.82	18.94
		3	1.34	2.23	21.17
		4	1.17	1.95	23.12
		5	1.12	1.86	24.98
		6	1.06	1.77	26.75
		7	1.04	1.74	28.49
		8	1.02	1.70	30.19
		9	1.02	1.69	31.88
	SUA	**1**	**9.98**	**16.63**	**16.63**
		2	1.66	2.76	19.39
		3	1.32	2.20	21.59
		4	1.18	1.96	23.55
		5	1.10	1.83	25.38
		6	1.06	1.76	27.14
		7	1.04	1.73	28.87
		8	1.03	1.71	30.58
		9	1.01	1.69	32.27

Table D1. Factor Analysis Results for ELA Tests (Selected Subpopulations) (cont.)

		Initial Eigenvalues			
Grade	Subgroups	Component	Total	% of Variance	Cumulative %
7	SWD/SUA	**1**	**9.48**	**15.80**	**15.80**
		2	1.67	2.79	18.59
		3	1.34	2.23	20.82
		4	1.17	1.95	22.77
		5	1.12	1.86	24.63
		6	1.06	1.77	26.40
		7	1.05	1.75	28.15
		8	1.04	1.73	29.88
		9	1.02	1.70	31.58
		10	1.01	1.68	33.26
	ELL/SUA	**1**	**8.69**	**14.48**	**14.48**
		2	1.70	2.83	17.31
		3	1.39	2.31	19.62
		4	1.25	2.08	21.70
		5	1.19	1.98	23.68
		6	1.17	1.96	25.64
		7	1.15	1.91	27.55
		8	1.11	1.85	29.40
		9	1.11	1.85	31.25
		10	1.09	1.82	33.07
		11	1.07	1.78	34.85
		12	1.06	1.77	36.62
		13	1.04	1.74	38.36
		14	1.03	1.72	40.08
		15	1.03	1.71	41.79
		16	1.02	1.70	43.49
8	ELL	**1**	**6.73**	**12.46**	**12.46**
		2	1.65	3.05	15.51

Table D1. Factor Analysis Results for ELA Tests (Selected Subpopulations) (cont.)

		Initial Eigenvalues			
Grade	Subgroups	Component	Total	% of Variance	Cumulative %
8	ELL	3	1.32	2.45	17.96
		4	1.22	2.26	20.22
		5	1.15	2.12	22.34
		6	1.11	2.05	24.39
		7	1.09	2.02	26.41
		8	1.07	1.98	28.39
		9	1.05	1.94	30.33
		10	1.04	1.93	32.26
		11	1.03	1.91	34.17
		12	1.02	1.89	36.06
		13	1.01	1.87	37.93
		14	1.00	1.86	39.79
	SWD	**1**	**7.59**	**14.05**	**14.05**
		2	1.67	3.10	17.15
		3	1.24	2.30	19.45
		4	1.13	2.09	21.54
		5	1.08	2.01	23.55
		6	1.07	1.98	25.53
		7	1.05	1.94	27.47
		8	1.03	1.91	29.38
		9	1.02	1.89	31.27
		10	1.01	1.87	33.14
	SUA	**1**	**7.90**	**14.63**	**14.63**
		2	1.63	3.02	17.65
		3	1.24	2.29	19.94
		4	1.12	2.07	22.01
		5	1.08	1.99	24.00
		6	1.08	1.99	25.99

Table D1. Factor Analysis Results for ELA Tests (Selected Subpopulations) (cont.)

		Initial Eigenvalues			
Grade	Subgroups	Component	Total	% of Variance	Cumulative %
8	SUA	7	1.05	1.94	27.93
		8	1.04	1.92	29.85
		9	1.01	1.87	31.72
		10	1.01	1.86	33.58
	SWD/SUA	**1**	**7.36**	**13.63**	**13.63**
		2	1.63	3.02	16.65
		3	1.25	2.31	18.96
		4	1.14	2.11	21.07
		5	1.09	2.02	23.09
		6	1.08	2.00	25.09
		7	1.06	1.96	27.05
		8	1.04	1.92	28.97
		9	1.03	1.90	30.87
		10	1.01	1.87	32.74
		11	1.01	1.87	34.61
		12	1.00	1.86	36.47
	ELL/SUA	**1**	**7.04**	**13.05**	**13.05**
		2	1.73	3.21	16.26
		3	1.36	2.51	18.77
		4	1.31	2.42	21.19
		5	1.24	2.30	23.49
		6	1.21	2.24	25.73
		7	1.18	2.18	27.91
		8	1.14	2.12	30.03
		9	1.13	2.08	32.11
		10	1.11	2.06	34.17
		11	1.09	2.02	36.19
		12	1.08	2.00	38.19

Table D1. Factor Analysis Results for ELA Tests (Selected Subpopulations) (cont.)

		Initial Eigenvalues			
Grade	Subgroups	Component	Total	% of Variance	Cumulative %
8	ELL/SUA	13	1.07	1.97	40.16
		14	1.05	1.95	42.11
		15	1.04	1.93	44.04
		16	1.04	1.92	45.96
		17	1.01	1.88	47.84

Appendix E—Items Flagged for DIF

These tables support the DIF information in Section V, "Operational Test Data Collection and Classical Analysis." They include item numbers, focal group, and directions of DIF and DIF statistics. Table E1 shows items flagged by the SMD or Mantel-Haenszel methods. Note that positive values of SMD and Delta in Table E1 indicate DIF in favor of a focal group and negative values of SMD and Delta indicate DIF against a focal group.

Table E1. NYSTP ELA 2012 Classical DIF Item Flags

Grade	Item #	Subgroup	DIF	SMD	Mantel-Haenszel	Delta
3	32	Asian	Against	No Flag	604.55	–2.35
3	32	Hispanic	Against	No Flag	546.15	–1.54
3	33	ELL	Against	-0.1529	2010.09	–2.03
3	33	Asian	Against	No Flag	1325.37	–2.36
3	33	Hispanic	Against	No Flag	1728.58	–1.82
3	36	ELL	Against	-0.1475	No Flag	No Flag
3	37	Female	In favor	0.1045	No Flag	No Flag
3	51	Asian	Against	-0.1099	No Flag	No Flag
3	54	ELL	In Favor	0.1055	No Flag	No Flag
3	54	Asian	In Favor	0.1108	No Flag	No Flag
4	3	Asian	Against	No Flag	1049.80	–1.74
4	3	Hispanic	Against	-0.1316	2670.08	–1.75
4	4	Asian	Against	No Flag	629.32	–1.53
4	7	ELL	Against	-0.1358	No Flag	No Flag
4	40	Female	In Favor	0.1333	No Flag	No Flag
4	60	Asian	In Favor	0.1019	No Flag	No Flag
5	4	ELL	Against	-0.1147	No Flag	No Flag
5	13	ELL	Against	-0.1057	No Flag	No Flag
5	20	ELL	Against	-0.1238	1032.74	–1.60
5	20	Asian	Against	No Flag	942.56	–2.13
5	21	Asian	In Favor	0.1019	No Flag	No Flag
5	38	Black	In Favor	0.1104	No Flag	No Flag
5	50	ELL	Against	No Flag	708.49	–1.57
5	55	Asian	Against	-0.1026	No Flag	No Flag

Table E1. NYSTP ELA 2012 Classical DIF Item Flags (cont.)

Grade	Item #	Subgroup	DIF	SMD	Mantel-Haenszel	Delta
5	60	Female	In Favor	0.1361	No Flag	No Flag
6	38	Asian	Against	No Flag	814.90	–1.57
6	40	ELL	Against	-0.1361	No Flag	No Flag
6	42	Female	In Favor	0.1751	No Flag	No Flag
6	45	Hispanic	Against	No Flag	1,101.52	–1.64
6	51	ELL	Against	-0.1552	1,200.69	–1.87
6	56	ELL	In Favor	0.1148	No Flag	No Flag
6	56	Black	In Favor	0.1081	No Flag	No Flag
6	56	Hispanic	In Favor	0.1040	No Flag	No Flag
6	57	ELL	In Favor	0.1843	No Flag	No Flag
6	57	Asian	In Favor	0.1172	No Flag	No Flag
6	57	Hispanic	In Favor	0.1371	No Flag	No Flag
6	59	ELL	In Favor	0.1289	No Flag	No Flag
6	60	ELL	In Favor	0.1415	No Flag	No Flag
7	2	ELL	Against	-0.1145	No Flag	No Flag
7	2	Asian	Against	No Flag	1,017.06	–1.79
7	23	Asian	Against	-0.1085	1,032.61	–1.59
7	36	Female	Against	-0.1334	6,094.04	–2.24
7	37	Female	Against	-0.1327	4,989.63	–1.87
7	40	ELL	Against	-0.1143	No Flag	No Flag
7	42	Female	In Favor	0.1272	No Flag	No Flag
7	56	ELL	In Favor	0.1066	No Flag	No Flag
7	56	Asian	In Favor	0.1008	No Flag	No Flag
7	57	ELL	In Favor	0.1037	No Flag	No Flag
7	59	ELL	In Favor	0.1217	No Flag	No Flag
7	60	Female	In Favor	0.2165	No Flag	No Flag
8	12	Female	Against	No Flag	3244.08	–1.71
8	15	Hispanic	Against	-0.1029	No Flag	No Flag
8	34	ELL	Against	-0.1381	No Flag	No Flag
8	36	Female	In Favor	0.1482	No Flag	No Flag

Table E1. NYSTP ELA 2012 Classical DIF Item Flags (cont.)

Grade	Item #	Subgroup	DIF	SMD	Mantel-Haenszel	Delta
8	50	ELL	In Favor	0.1393	No Flag	No Flag
8	51	ELL	In Favor	0.2648	No Flag	No Flag
8	51	Asian	In Favor	0.1071	No Flag	No Flag
8	54	Female	In Favor	0.1403	No Flag	No Flag

Appendix F—Derivation of the Generalized SPI Procedure

The standard performance index (SPI) is an estimated true score (estimated proportion of total or maximum points obtained) based on the performance of a given examinee for the items in a given Learning Standard. Assume a *k*-item test composed of *j* standards with a maximum possible raw score of *n*. Also assume that each item contributes to at most one standard, and the k_j items in standard *j* contribute a maximum of n_j points. Define X_j as the observed raw score on standard *j*. The true score is

$$T_j \equiv E(X_j / n_j).$$

It is assumed that there is information available about the examinee in addition to the standard score, and this information provides a prior distribution for T_j. This prior distribution of T_j for a given examinee is assumed to be $\beta(r_j, s_j)$:

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

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

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

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

where

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

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

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

with

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

and

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

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

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

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

Estimation of the Prior Distribution of T_j

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

The 3PL model is

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

where

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

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

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

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 produces (Novick and Jackson, 1974, p. 113)

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

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),

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

where

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

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

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

and

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

Thus,

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

and the parameters of the prior beta distribution for T_j can be expressed in terms of the parameters of the three-parameter IRT and 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 score (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:

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

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

Check on Consistency and Adjustment of Weight Given to Prior Estimate

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

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

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

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

Possible Violations of the Assumptions

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

If the items in the standard do not permit guessing, it is reasonable to assume $\hat{T}_j$, the expected proportion correct of maximum score for a standard, will be greater or equal to zero. If correct

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

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

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

Appendix G—Derivation of Classification Consistency and Accuracy

Classification Consistency

Assume that θ is a single latent trait measured by a test and denote Φ as a latent random variable. When a test X consists of *K* items and its maximum number correct score is *N*, the marginal probability of the number correct (NC) score *x* is

$$P(X = x) = \int P(X = x \mid \Phi = \theta) g(\theta) d\theta, \quad x = 0,1,..., N$$

where

$g(\theta)$ is the density of θ.

In this report, the marginal distribution $P(X = x)$ is denoted as $f(x)$, and the conditional error distribution $P(X = x \mid \Phi = \theta)$ is denoted as $f(x \mid \theta)$. It is assumed that examinees are classified into one of H mutually exclusive categories on the basis of predetermined H-1 observed score cutoffs, C_1, C_2, …, C_{H-1}. Let L_h represent the h^{th} category into which examinees with $C_{h-1} \le X \le C_h$ are classified. $C_0 = 0$ and $C_H =$ the maximum number-correct score. Then, the conditional and marginal probabilities of each category classification are as follows:

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

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

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

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

Consistent classification means that both X_1 and X_2 fall in the same category. The conditional probability of falling in the same category on the two administrations is

$$P(X_1 \in L_h,\ X_2 \in L_h \mid \theta) = \left[\sum_{x_1 = C_{h-1}}^{C_{h-1}} f(x_1 \mid \theta) \right]^2, \qquad h = 1, 2, \ldots, \text{H}.$$

The agreement index P, conditional on theta, is obtained by

$$P(\theta) = \sum_{h=1}^{H} P(X_1 \in L_h,\ X_2 \in L_h \mid \theta).$$

The agreement index (classification consistency) can be computed as

$$P = \int P(\theta) g(\theta) d(\theta).$$

The probability of consistent classification by chance, P_C, is the sum of squared marginal probabilities of each category classification.

$$P_C = \sum_{h=1}^{H} P(X_1 \in L_h) P(X_2 \in L_h) = \sum_{h=1}^{H} \left[P(X_1 \in L_h) \right]^2.$$

Then, the coefficient kappa (Cohen, 1960) is

$$k = \frac{P - P_C}{1 - P_C}.$$

Classification Accuracy

Let Γ_w denote true category. When an examinee has an observed score, $x \in L_h$ (h =1, 2, …, H), and a latent score , $\theta \in \Gamma_w$ (w=1, 2,…, H), an accurate classification is made when $h = w$. The conditional probability of accurate classification is

$$\gamma(\theta) = P(X \in L_w \mid \theta),$$

where

w is the category such that $\theta \in \Gamma_w$.

Appendix H—Scale Score Frequency Distributions

Tables I1–I6 depict the scale score (SS) distributions, by frequency (N-count), percent, cumulative frequency, and cumulative percent. The data in the tables include all public and charter school students with valid scale scores.

Table H1. Grade 3 ELA 2012 SS Frequency Distribution, State

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
475	51	0.03	51	0.03
515	36	0.02	87	0.04
559	79	0.04	166	0.08
575	135	0.07	301	0.15
584	140	0.07	441	0.22
591	240	0.12	681	0.34
596	311	0.16	992	0.50
600	377	0.19	1,369	0.69
604	443	0.22	1,812	0.92
608	496	0.25	2,308	1.17
611	553	0.28	2,861	1.45
613	634	0.32	3,495	1.77
616	742	0.37	4,237	2.14
618	770	0.39	5,007	2.53
620	871	0.44	5,878	2.97
623	916	0.46	6,794	3.43
624	955	0.48	7,749	3.91
626	1,038	0.52	8,787	4.44
628	1,096	0.55	9,883	4.99
630	1,169	0.59	11,052	5.58
631	1,245	0.63	12,297	6.21
633	1,385	0.70	13,682	6.91

Table H1. Grade 3 ELA 2012 SS Frequency Distribution, State (cont.)

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
635	1,460	0.74	15,142	7.65
636	1,550	0.78	16,692	8.43
637	1,633	0.82	18,325	9.26
639	1,918	0.97	20,243	10.22
640	1,988	1.00	22,231	11.23
642	2,085	1.05	24,316	12.28
643	2,211	1.12	26,527	13.40
644	2,430	1.23	28,957	14.63
645	2,661	1.34	31,618	15.97
647	2,784	1.41	34,402	17.38
648	3,060	1.55	37,462	18.92
649	3,274	1.65	40,736	20.57
651	3,416	1.73	44,152	22.30
652	3,680	1.86	47,832	24.16
653	3,974	2.01	51,806	26.17
654	4,260	2.15	56,066	28.32
656	4,480	2.26	60,546	30.58
657	4,795	2.42	65,341	33.00
658	5,188	2.62	70,529	35.62
659	5,417	2.74	75,946	38.36
661	5,688	2.87	81,634	41.23
662	5,983	3.02	87,617	44.25
664	6,383	3.22	94,000	47.48
665	6,709	3.39	100,709	50.86
667	7,090	3.58	107,799	54.45
668	7,406	3.74	115,205	58.19
670	7,646	3.86	122,851	62.05

Table H1. Grade 3 ELA 2012 SS Frequency Distribution, State (cont.)

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
672	7,843	3.96	130,694	66.01
674	8,088	4.08	138,782	70.09
676	8,207	4.15	146,989	74.24
678	8,202	4.14	155,191	78.38
680	8,044	4.06	163,235	82.44
683	7,731	3.90	170,966	86.35
686	7,217	3.65	178,183	89.99
690	6,371	3.22	184,554	93.21
695	5,298	2.68	189,852	95.89
700	3,858	1.95	193,710	97.84
708	2,603	1.31	196,313	99.15
722	1,286	0.65	197,599	99.80
780	394	0.20	197,993	100.00

Table H2. Grade 4 ELA 2012 SS Frequency Distribution, State

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
430	84	0.04	84	0.04
469	56	0.03	140	0.07
515	103	0.05	243	0.13
537	122	0.06	365	0.19
552	205	0.11	570	0.29
562	213	0.11	783	0.40
571	299	0.15	1,082	0.56
578	370	0.19	1,452	0.75
584	413	0.21	1,865	0.96
590	459	0.24	2,324	1.20
595	486	0.25	2,810	1.45

Table H2. Grade 4 ELA 2012 SS Frequency Distribution, State (cont.)

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
599	545	0.28	3,355	1.73
603	618	0.32	3,973	2.04
607	643	0.33	4,616	2.38
610	685	0.35	5,301	2.73
613	816	0.42	6,117	3.15
616	874	0.45	6,991	3.60
619	944	0.49	7,935	4.08
622	1,035	0.53	8,970	4.62
625	1,140	0.59	10,110	5.20
627	1,260	0.65	11,370	5.85
629	1,414	0.73	12,784	6.58
632	1,366	0.70	14,150	7.28
634	1,578	0.81	15,728	8.09
636	1,669	0.86	17,397	8.95
638	1,716	0.88	19,113	9.83
640	1,957	1.01	21,070	10.84
642	2,038	1.05	23,108	11.89
644	2,273	1.17	25,381	13.06
645	2,413	1.24	27,794	14.30
647	2,471	1.27	30,265	15.57
649	2,712	1.40	32,977	16.97
651	2,954	1.52	35,931	18.49
653	3,164	1.63	39,095	20.12
654	3,380	1.74	42,475	21.86
656	3,583	1.84	46,058	23.70
658	3,823	1.97	49,881	25.67
660	4,094	2.11	53,975	27.77

Table H2. Grade 4 ELA 2012 SS Frequency Distribution, State (cont.)

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
661	4,309	2.22	58,284	29.99
663	4,674	2.41	62,958	32.40
665	4,911	2.53	67,869	34.92
667	5,083	2.62	72,952	37.54
669	5,562	2.86	78,514	40.40
671	5,735	2.95	84,249	43.35
673	6,127	3.15	90,376	46.50
675	6,453	3.32	96,829	49.82
677	6,675	3.43	103,504	53.26
679	7,030	3.62	110,534	56.88
682	7,277	3.74	117,811	60.62
684	7,414	3.81	125,225	64.43
687	7,486	3.85	132,711	68.29
690	7,683	3.95	140,394	72.24
692	7,512	3.87	147,906	76.11
696	7,408	3.81	155,314	79.92
699	7,195	3.70	162,509	83.62
703	6,764	3.48	169,273	87.10
707	6,187	3.18	175,460	90.28
712	5,425	2.79	180,885	93.07
717	4,453	2.29	185,338	95.37
723	3,541	1.82	188,879	97.19
731	2,546	1.31	191,425	98.50
740	1,571	0.81	192,996	99.31
754	901	0.46	193,897	99.77
775	447	0.23	194,344	100.00

Table H3. Grade 5 ELA 2012 SS Frequency Distribution, State

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
495	33	0.02	33	0.02
563	27	0.01	60	0.03
584	42	0.02	102	0.05
594	75	0.04	177	0.09
601	111	0.06	288	0.15
606	150	0.08	438	0.22
610	184	0.09	622	0.32
614	235	0.12	857	0.44
616	314	0.16	1,171	0.60
619	379	0.19	1,550	0.79
621	430	0.22	1,980	1.01
624	486	0.25	2,466	1.25
625	494	0.25	2,960	1.51
627	548	0.28	3,508	1.78
629	651	0.33	4,159	2.12
631	675	0.34	4,834	2.46
632	758	0.39	5,592	2.84
634	801	0.41	6,393	3.25
635	849	0.43	7,242	3.68
636	921	0.47	8,163	4.15
638	962	0.49	9,125	4.64
639	1,051	0.53	10,176	5.18
640	1,144	0.58	11,320	5.76
641	1,223	0.62	12,543	6.38
643	1,350	0.69	13,893	7.07
644	1,408	0.72	15,301	7.78
645	1,574	0.80	16,875	8.58

Table H3. Grade 5 ELA 2012 SS Frequency Distribution, State (cont.)

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
646	1,630	0.83	18,505	9.41
647	1,880	0.96	20,385	10.37
648	1,985	1.01	22,370	11.38
649	2,099	1.07	24,469	12.44
651	2,329	1.18	26,798	13.63
652	2,399	1.22	29,197	14.85
653	2,609	1.33	31,806	16.18
654	2,874	1.46	34,680	17.64
655	3,153	1.60	37,833	19.24
656	3,344	1.70	41,177	20.94
657	3,507	1.78	44,684	22.73
658	3,806	1.94	48,490	24.66
659	4,103	2.09	52,593	26.75
661	4,247	2.16	56,840	28.91
662	4,542	2.31	61,382	31.22
663	4,865	2.47	66,247	33.69
664	5,126	2.61	71,373	36.30
665	5,582	2.84	76,955	39.14
667	5,823	2.96	82,778	42.10
668	6,219	3.16	88,997	45.26
669	6,579	3.35	95,576	48.61
671	6,762	3.44	102,338	52.05
672	7,090	3.61	109,428	55.65
674	7,339	3.73	116,767	59.39
675	7,653	3.89	124,420	63.28
677	7,976	4.06	132,396	67.33
679	7,806	3.97	140,202	71.30

Table H3. Grade 5 ELA 2012 SS Frequency Distribution, State (cont.)

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
681	7,930	4.03	148,132	75.34
683	7,762	3.95	155,894	79.29
685	7,479	3.80	163,373	83.09
687	6,986	3.55	170,359	86.64
690	6,526	3.32	176,885	89.96
693	5,654	2.88	182,539	92.84
696	4,778	2.43	187,317	95.27
700	3,677	1.87	190,994	97.14
705	2,625	1.34	193,619	98.47
712	1,665	0.85	195,284	99.32
721	922	0.47	196,206	99.79
738	352	0.18	196,558	99.97
795	65	0.03	196,623	100.00

Table H4. Grade 6 ELA 2012 SS Frequency Distribution, State

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
480	28	0.01	28	0.01
567	27	0.01	55	0.03
586	46	0.02	101	0.05
595	58	0.03	159	0.08
601	94	0.05	253	0.13
605	130	0.07	383	0.19
609	189	0.09	572	0.29
612	226	0.11	798	0.40
615	257	0.13	1,055	0.53
617	299	0.15	1,354	0.68
619	389	0.19	1,743	0.87

Table H4. Grade 6 ELA 2012 SS Frequency Distribution, State (cont.)

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
621	444	0.22	2,187	1.10
623	478	0.24	2,665	1.34
625	534	0.27	3,199	1.60
627	597	0.30	3,796	1.90
628	632	0.32	4,428	2.22
630	719	0.36	5,147	2.58
631	837	0.42	5,984	3.00
632	855	0.43	6,839	3.43
634	1,026	0.51	7,865	3.94
635	1,143	0.57	9,008	4.51
636	1,256	0.63	10,264	5.14
637	1,375	0.69	11,639	5.83
639	1,409	0.71	13,048	6.54
640	1,623	0.81	14,671	7.35
641	1,752	0.88	16,423	8.23
642	1,980	0.99	18,403	9.22
643	2,118	1.06	20,521	10.28
644	2,338	1.17	22,859	11.46
645	2,631	1.32	25,490	12.77
646	2,780	1.39	28,270	14.17
647	2,992	1.50	31,262	15.67
648	3,210	1.61	34,472	17.28
649	3,537	1.77	38,009	19.05
651	3,732	1.87	41,741	20.92
652	3,995	2.00	45,736	22.92
653	4,321	2.17	50,057	25.09
654	4,641	2.33	54,698	27.41

Table H4. Grade 6 ELA 2012 SS Frequency Distribution, State (cont.)

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
655	4,794	2.40	59,492	29.81
656	5,078	2.54	64,570	32.36
657	5,489	2.75	70,059	35.11
658	5,653	2.83	75,712	37.94
660	5,993	3.00	81,705	40.95
661	6,145	3.08	87,850	44.03
662	6,422	3.22	94,272	47.24
663	6,724	3.37	100,996	50.61
665	6,970	3.49	107,966	54.11
666	7,143	3.58	115,109	57.69
667	7,202	3.61	122,311	61.30
669	7,369	3.69	129,680	64.99
670	7,216	3.62	136,896	68.61
672	7,142	3.58	144,038	72.19
673	7,194	3.61	151,232	75.79
675	6,817	3.42	158,049	79.21
677	6,760	3.39	164,809	82.59
678	6,485	3.25	171,294	85.84
680	5,956	2.98	177,250	88.83
683	5,352	2.68	182,602	91.51
685	4,731	2.37	187,333	93.88
688	3,985	2.00	191,318	95.88
691	3,214	1.61	194,532	97.49
695	2,355	1.18	196,887	98.67
700	1,459	0.73	198,346	99.40
707	776	0.39	199,122	99.79
722	345	0.17	199,467	99.96
785	73	0.04	199,540	100.00

Table H5. Grade 7 ELA 2012 SS Frequency Distribution, State

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
470	47	0.02	47	0.02
536	26	0.01	73	0.04
571	53	0.03	126	0.06
584	73	0.04	199	0.10
591	101	0.05	300	0.15
597	140	0.07	440	0.22
601	186	0.09	626	0.32
605	211	0.11	837	0.42
608	264	0.13	1,101	0.56
611	337	0.17	1,438	0.73
614	343	0.17	1,781	0.90
616	368	0.19	2,149	1.09
619	402	0.20	2,551	1.29
621	479	0.24	3,030	1.53
623	520	0.26	3,550	1.80
625	594	0.30	4,144	2.10
626	637	0.32	4,781	2.42
628	690	0.35	5,471	2.77
629	703	0.36	6,174	3.12
631	746	0.38	6,920	3.50
632	861	0.44	7,781	3.94
634	919	0.46	8,700	4.40
635	980	0.50	9,680	4.90
636	1,022	0.52	10,702	5.41
637	1,140	0.58	11,842	5.99
639	1,213	0.61	13,055	6.61
640	1,322	0.67	14,377	7.27

Table H5. Grade 7 ELA 2012 SS Frequency Distribution, State (cont.)

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
641	1,461	0.74	15,838	8.01
642	1,510	0.76	17,348	8.78
643	1,688	0.85	19,036	9.63
644	1,780	0.90	20,816	10.53
645	1,956	0.99	22,772	11.52
646	2,111	1.07	24,883	12.59
647	2,297	1.16	27,180	13.75
648	2,454	1.24	29,634	14.99
649	2,620	1.33	32,254	16.32
650	2,883	1.46	35,137	17.78
651	3,232	1.64	38,369	19.41
652	3,357	1.70	41,726	21.11
654	3,664	1.85	45,390	22.97
655	4,066	2.06	49,456	25.02
656	4,376	2.21	53,832	27.24
657	4,573	2.31	58,405	29.55
658	4,905	2.48	63,310	32.03
659	5,300	2.68	68,610	34.71
660	5,637	2.85	74,247	37.57
662	6,066	3.07	80,313	40.64
663	6,504	3.29	86,817	43.93
664	6,700	3.39	93,517	47.32
665	7,071	3.58	100,588	50.90
667	7,367	3.73	107,955	54.62
668	7,806	3.95	115,761	58.57
670	8,000	4.05	123,761	62.62
672	8,144	4.12	131,905	66.74

Table H5. Grade 7 ELA 2012 SS Frequency Distribution, State (cont.)

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
673	8,438	4.27	140,343	71.01
675	8,252	4.18	148,595	75.19
677	8,348	4.22	156,943	79.41
680	7,951	4.02	164,894	83.43
682	7,445	3.77	172,339	87.20
685	6,890	3.49	179,229	90.69
689	6,056	3.06	185,285	93.75
693	5,027	2.54	190,312	96.29
699	3,668	1.86	193,980	98.15
707	2,313	1.17	196,293	99.32
723	1,088	0.55	197,381	99.87
790	257	0.13	197,638	100.00

Table H6. Grade 8 ELA 2012 SS Frequency Distribution, State

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
430	20	0.01	20	0.01
509	22	0.01	42	0.02
538	23	0.01	65	0.03
556	39	0.02	104	0.05
566	44	0.02	148	0.07
573	77	0.04	225	0.11
578	100	0.05	325	0.16
583	121	0.06	446	0.22
587	154	0.08	600	0.30
591	195	0.10	795	0.40
594	251	0.13	1,046	0.53
597	294	0.15	1,340	0.68

Table H6. Grade 8 ELA 2012 SS Frequency Distribution, State (cont.)

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
599	333	0.17	1,673	0.84
602	376	0.19	2,049	1.03
604	448	0.23	2,497	1.26
607	530	0.27	3,027	1.53
609	575	0.29	3,602	1.82
611	592	0.30	4,194	2.12
613	711	0.36	4,905	2.47
615	756	0.38	5,661	2.85
617	870	0.44	6,531	3.29
619	904	0.46	7,435	3.75
620	1,067	0.54	8,502	4.29
622	1,228	0.62	9,730	4.91
624	1,206	0.61	10,936	5.52
626	1,449	0.73	12,385	6.25
627	1,483	0.75	13,868	6.99
629	1,811	0.91	15,679	7.91
631	1,903	0.96	17,582	8.87
632	2,139	1.08	19,721	9.95
634	2,410	1.22	22,131	11.16
635	2,612	1.32	24,743	12.48
637	2,866	1.45	27,609	13.92
639	3,354	1.69	30,963	15.61
640	3,640	1.84	34,603	17.45
642	3,911	1.97	38,514	19.42
643	4,169	2.10	42,683	21.53
645	4,514	2.28	47,197	23.80
646	4,949	2.50	52,146	26.30

Table H6. Grade 8 ELA 2012 SS Frequency Distribution, State (cont.)

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
648	5,457	2.75	57,603	29.05
649	5,745	2.90	63,348	31.95
651	6,022	3.04	69,370	34.98
652	6,548	3.30	75,918	38.29
654	6,826	3.44	82,744	41.73
655	7,410	3.74	90,154	45.46
657	7,602	3.83	97,756	49.30
659	8,181	4.13	105,937	53.42
661	8,388	4.23	114,325	57.65
663	8,989	4.53	123,314	62.19
665	9,047	4.56	132,361	66.75
667	9,262	4.67	141,623	71.42
670	9,424	4.75	151,047	76.17
673	9,283	4.68	160,330	80.85
676	9,083	4.58	169,413	85.44
680	8,059	4.06	177,472	89.50
684	7,045	3.55	184,517	93.05
690	5,855	2.95	190,372	96.00
697	4,294	2.17	194,666	98.17
708	2,405	1.21	197,071	99.38
728	1,007	0.51	198,078	99.89
790	216	0.11	198,294	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 Screen 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.

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.

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: Springer-Verlag.

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

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

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

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

Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: Lawrence Erlbaum.

Lord, F.M. and M.R. Novick (1968). *Statistical Theories of Mental Test Scores.* Menlo Park, CA: Addison-Wesley.

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

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

Muraki, E. and R.D. Bock (1991). *PARSCALE*: *Parameter Scaling of Rating Data* [Computer program]. Chicago, IL: Scientific Software, Inc.

Novick, M.R. and P.H. Jackson (1974). *Statistical Methods for Educational and Psychological Research.* New York: McGraw-Hill.

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

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

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

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

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

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, IL: MESA Press.

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

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

Yen, W. M. (1984). Obtaining maximum likelihood trait estimates from number correct scores for the three-parameter logistic model. *Journal of Educational Measurement* 21:93–111.

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

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