New York State Testing Program 2011: English Language Arts, Grades 3–8

Technical Report

CTB/McGraw-Hill
Monterey, California 93940
2011

Copyright

Table of Contents

List of Tables

Section I: Introduction and Overview

Introduction

An overview of the New York State Testing Program (NYSTP) Grades 3–8 English Language Arts (ELA) 2011 Operational (OP) Tests is provided in this report. The report contains information about OP test development and content, item and test statistics, validity and reliability, differential item functioning studies, test administration and scoring, scaling, and student performance.

Test Purpose

The NYSTP is an assessment system designed to measure concepts, processes, and skills taught in schools in New York. 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 age) are the target population for the Grades 3–8 testing program. Nonpublic schools may participate in the testing program, but the participation is not mandatory for them. In 2011, nonpublic schools participated in all grade tests but were not well represented in the testing program. The New York State Education Department (NYSED) made a decision to exclude these schools from the data analyses. Public school students were required to take all State assessments administered at their grade level, except for a very small percentage of students with disabilities who took the New York State Alternate Assessment (NYSAA) for students with severe disabilities. For more detail on this exemption, please refer to the *School Administrator's Manual*, available online at http://www.p12.nysed.gov/osa/manuals/.

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 these 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 among Levels I, II, III, and IV were established during the process of Standard Setting in 2006. In 2010, a change in the test administration window between the 2008–2009 and 2009–2010 school years and a decision to align the proficiency standards with Grade 8 student performance on the NYS 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 academic strengths and weaknesses of their students. These scores can be effectively used by teachers at the classroom level to modify their instructional content and format to best serve their students' specific needs. Detailed information on the properties and use of SPI scores are provided in Section VI, "IRT Scaling and Equating."

Testing Accommodations

In accordance with federal law under the Americans with Disabilities Act and Fairness in Testing, as outlined by the *Standards for Educational and Psychological Testing* (American Education Research Association, American Psychological Association, and National Council on Measurement in Education, 1999), accommodations that do not alter the measurement of any construct being tested are allowed for test takers. The allowance is in accordance with a student's individual education program (IEP) or section 504 Accommodation Plan (504 Plan). School principals are responsible for ensuring that proper accommodations are provided when necessary and that staff providing accommodations are properly trained. Details on testing accommodations can be found in the *School Administrator's Manual*.

Test Transcriptions

For the visually impaired students, large type and braille editions of the test books are provided. The students dictate and/or record their responses; the teachers transcribe student responses to multiple-choice (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 CTB/McGraw-Hill and printed by NYSED, and the braille editions are produced by Braille Publishers, Inc. The lead transcribers are members of the National Braille Association, California Transcribers and Educators of the Visually Handicapped, and the Contra Costa Braille Transcribers, and they have Library of Congress and Nemeth Code [Braille] Certifications.

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 classrooms during May 2011 over a three-day (Grades 3, 4, and 5) or two-day (Grades 6, 7, and 8) period. The tests were printed in black and white and incorporated the concepts of universal design. Details on the administration and scoring of these tests can be found in Section IV, "Test Administration and Scoring."

Test Configuration

The OP test books were administered, in order, on two to three consecutive days, depending on the grade. Table 1 provides information on the number and type of items in each book, as well as testing times. Students were administered a Reading section (Book 1, all grades; Book 2, Grades 6, 7, and 8; Book 3, Grades 3, 4, and 5) and a Listening section (Book 2, all grades). The 2011 *Teacher's Directions* (http://www.p12.nysed.gov/apda/ei/directions/ela3-5-td-11.pdf and http://www.p12.nysed.gov/apda/ei/directions/ela6-8-td-11.pdf) as well as the 2011 *School Administrator's Manual* available online (http://www.p12.nysed.gov/apda/sam/ela/elaei-sam-11.pdf) provide details on security, scheduling, classroom organization and preparation, test materials, and administration.

Table 1. NYSTP ELA 2011 Test Configuration

Grade	Day	Book	Number of Items			Allotted Time (minutes)	
			MC	CR	Total**	Testing	Prep
3	1	1	35	0	35	60	10
	2	2	8	3	11	30	15
	3	3	0	5	5	60	10
	Totals		43	8	51	150	35
4	1	1	43	0	43	70	10
	2	2	8	3	11	30	15
	3	3	0	5	5	60	10
	Totals		51	8	59	160	35
5	1	1	35	0	35	60	10
	2	2	8	3	11	30	15
	3	3	0	5	5	60	10
	Totals		43	8	51	150	35
6	1	1	41	0	41	70	10
	2	2	8	8	16	90	15
	Totals		49	8	57	160	26

(Continued on next page)

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

Grade	Day	Book	Number of Items			Allotted Time (minutes)	
			MC	CR	Total**	Testing	Prep
7			41	0	41	70	10
			8	8	16	90	15
	Totals		49	8	57	160	25
8			41	0	41	70	10
			8	8	16	90	15
	Totals		49	8	57	160	25

**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 use of test items associated with informational passages; Standard 2 is assessed primarily by use of test items associated with literary passages; and Standard 3 is assessed by use of test items associated with a combination of genres. In addition, students are also tested on writing mechanics, which is assessed independent of alignment to the Learning Standards, since writing mechanics is associated with all three Learning Standards. 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 2011 OP tests.

Table 2. NYSTP ELA 2011 Test Blueprint

Grade	Total Points	Writing Mechanics Points	Standard	Target Reading and Listening Points	Selected Reading and Listening Points	Target % of Test (excluding Writing)	Selected % of Test (excluding Writing)
3	57	3	S1	19	17	33.0	30.0
			S2	27	28	47.0	49.0
			S3	11	12	20.0	21.0
4	66	3	S1	24	21	36.0	32.0
			S2	29	30	44.5	45.0
			S3	13	15	19.5	23.0

(Continued on next page)

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

Grade	Total Points	Writing Mechanics Points	Standard	Target Reading and Listening Points	Selected Reading and Listening Points	Target % of Test (excluding Writing)	Selected % of Test (excluding Writing)
5	58	3	S1	25	26	43.0	45.0
			S2	21	19	36.0	33.0
			S3	12	13	21.0	22.0
6	64	3	S1	23	23	36.0	36.0
			S2	28	26	44.5	41.0
			S3	12	15	19.5	23.0
7	64	3	S1	25	27	39.0	42.0
			S2	25	22	39.0	35.0
			S3	14	15	22.0	23.0
8	64	3	S1	25	28	39.0	44.0
			S2	25	21	39.0	33.0
			S3	14	15	22.0	23.0

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 following events: passage review, item review, rangefinding, and test form final-eyes review. These events are described in details 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, CTB/McGraw-Hill facilitated test specifications meetings in Albany, New York, during which committees of state educators, along with NYSED staff, reviewed the standards and performance indicators to make the following determinations:

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

The committees were composed of teachers from around the state 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

The first step in the process of item development for the NYSED-owned items appearing in the 2011 Grades 3–8 ELA Tests was selection of passages to be used. The CTB/McGraw-Hill passage selectors were provided with specifications based on the test design (see Appendix A). After an internal CTB/McGraw-Hill editorial and supervisory review, the passages were submitted to NYSED for their approval and then brought to a formal passage review meeting in Albany, New York. The purpose of the meeting was for committees of New York educators to review and decide whether to approve the passages. CTB/McGraw-Hill and NYSED staff were both present, with CTB/McGraw-Hill staff facilitating. After the committees completed their reviews, NYSED reviewed and approved the committees' decisions regarding the passages.

The content-lead editors at CTB/McGraw-Hill then selected from the approved passages those 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. Writers were trained in the NYSEP and test specifications. This training entailed specific assignments that spelled out the performance indicators and depth-of-knowledge levels to assess for each passage. In addition, 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.

CTB/McGraw-Hill editors and supervisors reviewed the items to verify that they met the specifications and criteria outlined in the writing assignments and, as necessary, revised them. After all revisions from CTB/McGraw-Hill staff had been incorporated, the items were submitted to NYSED staff for their review and approval. CTB/McGraw-Hill incorporated any necessary revisions from NYSED and prepared the items for a formal item review.

Item Review

As was done for the specifications and passage review meetings, the item review committees were composed of New York State educators selected for their content and grade-level expertise. Each committee was composed of approximately ten participants per grade band (i.e., Grades 3/4, 5/6, and 7/8). The committee members were provided with the test items, the New York State Learning Standards, and the test specifications, and they considered the following elements as they reviewed the test items:

- the accuracy and grade-level appropriateness of the items
- the mapping of the items to the assigned performance indicators
- the accompanying exemplary responses (CR items)
- the appropriateness of the correct response and distractors (MC items)
- the conciseness, preciseness, clarity, and reading load of the items
- the existence of any ethnic, gender, regional, or other possible bias evident in the items

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

Materials Development

Following item review, CTB/McGraw-Hill staff assembled the approved passages and items into field test (FT) forms and submitted the FT forms to NYSED for their review and approval. The FT forms were administered to students across New York State, using either the State Sampling Matrix (from 2005 to 2009) to ensure appropriate sampling of students or a census sample (in 2010). In addition, CTB/McGraw-Hill, in conjunction with NYSED test specialists, developed a combined *Teacher's Directions and School Administrator's Manual* 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 fifth year of the NYSTP Grades 3–8 English Language Arts OP Tests were administered in May 2011. The test items were selected from the pool of items primarily

field-tested in 2007, 2008, and 2009, using the data from those FT forms. The pool also included items owned by CTB/McGraw-Hill. These items consisted mostly of *TerraNova*™ items but also included items field-tested in New York State in 2010. Using this extended pool, CTB/McGraw-Hill made preliminary selections for each grade. The selections were reviewed for alignment with the test design, blueprint, and research guidelines for item selection (Appendix 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 CTB/McGraw-Hill and NYSED and reviewed by psychometricians at CTB/McGraw-Hill and at NYSED. Final approval of the selected items was given by NYSED. Two criteria governed the item selection process. The first of these was to meet the content specifications provided by NYSED. Second, within the limits set by these requirements, developers selected items with the best psychometric characteristics from the FT item pool.

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

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

CTB/McGraw-Hill editors traveled to Albany, New York, in September 2010 to finalize item selection and test creation with the NYSED staff (including content and research experts). NYSED discussed the content and data of the proposed selections, explored alternate selections for consideration, determined the final item selections, and ordered those items (assigned positions) in the OP test books. The final test forms were approved by the final eyes committee that consisted of approximately 12 participants across all grade levels. After the approval by NYSED, the tests were produced and administered in May 2011.

In addition to the test books, CTB/McGraw-Hill and NYSED produced a *School Administrator's Manual*, as well as two *Teacher's Directions*, one for Grades 3, 4, and 5 and one for Grades 6, 7, and 8, so that the tests were administered in a standardized

fashion across the state. These documents are located at the following web site: http://www.p12.nysed.gov/apda/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 NYS Regents ELA examinations led to changes in the proficiency cut scores. The results were reviewed by the NYS Technical Advisory Group and were approved by the Board of Regents in July 2010. For each grade level, there are four proficiency levels. Three cut points demarcate the performance standards needed to demonstrate each ascending level of proficiency.

Section III: Validity

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

The American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) *Standards for Educational and Psychological Testing* (1999) 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 0.90–0.92, and for all subgroups the reliability coefficient was equal to or greater than 0.86. Overall, high internal consistency of the New York State 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 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 except for item 30 in Grade 3 test, item 1 in Grade 4 test, and item 6 in Grade 8 test, all other items on the 2011 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 New York State 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 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 New York State 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 New York State Grades 3−8 ELA Tests exhibited first principal components accounting for more than 10% 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)

Grade	Initial Eigenvalues			
	Component	Total	% of Variance	Cumulative %
3	**1**	**10.22**	**20.04**	**20.04**
	2	1.92	3.77	23.81
	3	1.28	2.51	26.31
	4	1.10	2.15	28.47
	5	1.05	2.06	30.53
4	**1**	**10.82**	**18.34**	**18.34**
	2	1.59	2.70	21.04
	3	1.46	2.48	23.52
	4	1.23	2.09	25.61
	5	1.14	1.93	27.54
	6	1.04	1.77	29.31
	7	1.00	1.70	31.01

(Continued on next page)

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

Grade	Initial Eigenvalues			
	Component	Total	% of Variance	Cumulative %
5	**1**	**9.73**	**19.09**	**19.09**
	2	1.50	2.93	22.02
	3	1.21	2.38	24.40
	4	1.15	2.25	26.64
	5	1.04	2.05	28.69
	6	1.03	2.02	30.71
	7	1.02	1.99	32.70
6	**1**	**11.06**	**19.41**	**19.41**
	2	1.63	2.85	22.26
	3	1.51	2.65	24.91
	4	1.17	2.05	26.96
	5	1.06	1.85	28.81
	6	1.02	1.79	30.60
	7	1.02	1.78	32.38
7	**1**	**11.43**	**20.05**	**20.05**
	2	1.67	2.93	22.98
	3	1.46	2.57	25.54
	4	1.25	2.20	27.74
	5	1.07	1.87	29.61
	6	1.05	1.85	31.46
	7	1.01	1.77	33.22
8	**1**	**11.65**	**20.43**	**20.43**
	2	1.69	2.97	23.40
	3	1.62	2.84	26.24
	4	1.10	1.92	28.16
	5	1.05	1.84	30.00
	6	1.03	1.80	31.80

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

Minimization of Bias

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

Four procedures were used to eliminate bias and minimize differential item functioning (DIF) in the New York State 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. Three methods were employed to evaluate the amount of DIF in all test items: standardized mean difference, Mantel-Haenszel (see Section V "Operational Test Data Collection and Classical Analysis"), and Linn-Harnisch (see Section VI, "IRT Scaling and Equating"). 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* (2011) located at http://www.p12.nysed.gov/apda/ei/ssom/ssom-11.pdf.

Test Administration

NYSTP Grades 3–8 ELA Tests were administered at the classroom level during May 2011. The testing window for Grades 3–8 was May 3–6. The makeup test administration window for Grades 3–8 was May 4–11. 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, districtwide, or schoolwide scoring (please refer to the next subsection, "Scoring Models," for more detail). Administrators were responsible for the oversight of scoring operations, including the preparation of the test site, the security of test books, and the 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 districtwide 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 schoolwide scoring, oversight was provided by the principal; otherwise, titles for the schoolwide model were the same as those for the districtwide model. The general title "scoring committee members" included scorers at every site.

Scoring Models

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

1. Regional scoring—The 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.

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

Scoring of Constructed-Response Items

The scoring of CR items was based primarily on the scoring guides, which were created by CTB/McGraw-Hill handscoring and content development specialists during rangefinding sessions. In 2010, the CTB/McGraw-Hill ELA handscoring team was composed of six team leaders, each representing one grade. Team leaders were selected on the basis of their handscoring experiences along with their educational and professional backgrounds.

Sets of actual FT student responses were reviewed and discussed openly, and consensus scores were agreed upon. Scoring guides were developed based on rangefinding decisions. Audio files were created to further explain each section of the scoring guides. Trainers used these materials to train scoring committee members on the criteria for scoring CR items. Scoring Leader Handbooks were also distributed to outline the responsibilities of the scoring roles. CTB/McGraw-Hill handscoring staff also conducted training sessions in New York City to better equip the teachers and administrators with enhanced knowledge of scoring principles and criteria.

Scoring was conducted with pen and pencil scoring as opposed to electronic scoring, and each scoring committee member evaluated actual student papers instead of electronically scanned papers. All scoring committee members were trained by previously trained and approved trainers along with guidance from scoring guides and a CD containing the audio files that highlighted important elements of the scoring guides. Each test book was

scored by three separate scoring committee members, who scored three distinct sections of the test book. After test books were completed, the table facilitator or 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 and audio files to train scoring committee members on the criteria for scoring CR items. Part of the training process was the administration of a consistency assurance set (CAS) that provided the State's scoring sites with information regarding strengths and weaknesses of their scorers. This tool allowed trainers to retrain their scorers, if necessary. The CAS also acknowledged those scorers who had grasped all aspects of the content area being scored and were well prepared to score student responses.

Quality Control Process

Test books were randomly distributed throughout each scoring room so that books from each region, district, school, or class were evenly dispersed. Teams were divided into groups of three to ensure that a variety of scorers graded each book. If a scorer and facilitator could not reach a decision on a paper after reviewing the scoring guides and audio files, they called the New York State ELA Helpline. This call center was established to help teachers and administrators during OP scoring. The helpline staff consisted of trained CTB/McGraw-Hill handscoring personnel who answered questions by phone or fax. When a member of the staff was unable to resolve an issue, they deferred to NYSED for a scoring decision. A quality check was also performed on each completed box of scored tests to certify that all questions were scored and that the scoring committee members darkened each score on the answer document appropriately. The log of calls received by the scoring helpline was delivered to NYSED after the scoring window. To affirm that all schools across the state adhered to scoring guidelines and policies, approximately 5% of the schools' results are audited each year by an outside vendor.

Section V: Operational Test Data Collection and Classical Analysis

Data Collection

OP test data were collected in two phases. During phase 1, a sample of approximately 98% of the student test records were received from the data warehouse and delivered to CTB/McGraw-Hill in May 2011. These data were used for all data analysis. Phase 2 involved submitting "straggler files" to CTB/McGraw-Hill in early-June 2011. 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. CTB/McGraw-Hill established a scoring program, EDITCHECKER, to do initial quality assurance on data and identify errors. This program verifies that the data fields are in-range (as defined), that students' identifying information is present, and that the data are acceptable for delivery to CTB/McGraw-Hill research. NYSED and the data repository were provided with the results of the checking, and some edits to the initial data were made; however, CTB/McGraw-Hill 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 out-of-grade students (students whose grade level did not match the test level) and students from nonpublic schools. Other deleted cases included students with no grade-level data and duplicate record cases. A list of the data cleaning procedures conducted by research and accompanying case counts is presented in Tables 4A–4F.

Table 4A. NYSTP ELA Grade 3 Data Cleaning

Exclusion Rule	# Deleted	# Cases Remain
Initial N		198403
Out of grade	97	198306
No grade	1	198305
Duplicate record	0	198305
Non-public schools	3639	194666
Less than 5 items attempted	7	194659
Out-of-range CR scores	0	194659

Table 4B. NYSTP ELA Grade 4 Data Cleaning

Exclusion Rule	# Deleted	# Cases Remain
Initial N		209784
Out of grade	95	209689
No grade	0	209689
Duplicate record	0	209689
Non-public schools	14048	195641
Less than 5 items attempted	6	195635
Out-of-range CR scores	0	195635

Table 4C. NYSTP ELA Grade 5 Data Cleaning

Exclusion Rule	# Deleted	# Cases Remain
Initial N		201929
Out of grade	60	201869
No grade	2	201867
Duplicate record	0	201867
Non-public schools	3220	198647
Less than 5 items attempted	3	198644
Out-of-range CR scores	0	198644

Table 4D. NYSTP ELA Grade 6 Data Cleaning

Exclusion Rule	# Deleted	# Cases Remain
Initial N		208826
Out of grade	163	208663
No grade	1	208662
Duplicate record	0	208662
Non-public schools	12759	195903
Less than 5 items attempted	4	195899
Out-of-range CR scores	0	195899

Table 4E. NYSTP ELA Grade 7 Data Cleaning

Exclusion Rule	# Deleted	# Cases Remain
Initial N		201474
Out of grade	155	201319
No grade	1	201318
Duplicate record	0	201318
Non-public schools	3126	198192
Less than 5 items attempted	2	198190
Out-of-range CR scores	0	198190

Table 4F. NYSTP ELA Grade 8 Data Cleaning

Exclusion Rule	# Deleted	# Cases Remain
Initial N		215035
Out of grade	173	214862
No grade	1	214861
Duplicate record	0	214861
Non-public schools	15554	199307
Less than 5 items attempted	2	199305
Out-of-range CR scores	0	199305

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 code (NRC) is assigned at district level and is an indicator of district and school socioeconomic status. The ethnicity and gender designations are assigned at the student level. Please note that the tables do not include data for gender variable as it was found that the New York State population is fairly evenly split by gender categories.

Table 5A. Grade 3 Sample Characteristics (N = 194434)

Demographic Category		N-count	% of Total N-count
NRC	NYC	70515	36.27
	Big 4 cities	8105	4.17
	Urban/Suburban	15354	7.90
	Rural	11142	5.73
	Average needs	57475	29.56
	Low needs	26972	13.87
	Charter	4871	2.51
Ethnicity	Asian	15765	8.10
	Black	36201	18.60
	Hispanic	45334	23.29
	American Indian	1084	0.56
	Multi-Racial	1550	0.80
	Unknown	271	0.14
	White	94454	48.52
ELL	No	176653	90.75
	Yes	18006	9.25
SWD	No	167009	85.80
	Yes	27650	14.20
SUA	No	146410	75.21
	Yes	48249	24.79

Table 5B. Grade 4 Sample Characteristics (N = 195391)

Demographic Category		N-count	% of Total N-count
NRC	NYC	69906	35.78
	Big 4 cities	8204	4.20
	Urban/Suburban	14663	7.50
	Rural	11476	5.87
	Average needs	58879	30.13
	Low needs	28450	14.56
	Charter	3813	1.95
Ethnicity	Asian	15450	7.90
	Black	36364	18.59
	Hispanic	44276	22.63
	American Indian	948	0.48
	Multi-Racial	1374	0.70
	Unknown	247	0.13
	White	96976	49.57
ELL	No	179622	91.81
	Yes	16013	8.19
SWD	No	166090	84.90
	Yes	29545	15.10
SUA	No	146756	75.02
	Yes	48879	24.98

Table 5C. Grade 5 Sample Characteristics (N = 198644)

Demographic Category		N-count	% of Total N-count
NRC	NYC	69242	34.90
	Big 4 cities	7883	3.97
	Urban/Suburban	14059	7.09
	Rural	11517	5.81
	Average needs	60455	30.47
	Low needs	30049	15.15
	Charter	5185	2.61
Ethnicity	Asian	16599	8.36
	Black	36723	18.49
	Hispanic	43537	21.92
	American Indian	967	0.49
	Multi-Racial	1310	0.66
	Unknown	239	0.12
	White	99269	49.97
ELL	No	185004	93.13
	Yes	13640	6.87
SWD	No	168289	84.72
	Yes	30355	15.28
SUA	No	150178	75.60
	Yes	48466	24.40

Table 5D. Grade 6 Sample Characteristics (N = 195899)

Demographic Category		N-count	% of Total N-count
NRC	NYC	67739	34.63
	Big 4 cities	7734	3.95
	Urban/Suburban	13682	6.99
	Rural	11294	5.77
	Average needs	60364	30.86
	Low needs	30030	15.35
	Charter	4761	2.43
Ethnicity	Asian	15187	7.75
	Black	36912	18.84
	Hispanic	42740	21.82
	American Indian	923	0.47
	Multi-Racial	1235	0.63
	Unknown	246	0.13
	White	98656	50.36
ELL	No	184082	93.97
	Yes	11817	6.03
SWD	No	165992	84.73
	Yes	29907	15.27
SUA	No	151950	77.57
	Yes	43949	22.43

Table 5E. Grade 7 Sample Characteristics (N = 198190)

Demographic Category		N-count	% of Total N-count
NRC	NYC	68448	34.61
	Big 4 cities	7501	3.79
	Urban/Suburban	13964	7.06
	Rural	11428	5.78
	Average needs	60971	30.83
	Low needs	31829	16.09
	Charter	3648	1.84
Ethnicity	Asian	15168	7.65
	Black	37394	18.87
	Hispanic	42405	21.40
	American Indian	978	0.49
	Multi-Racial	1099	0.55
	Unknown	260	0.13
	White	100886	50.90
ELL	No	187509	94.61
	Yes	10681	5.39
SWD	No	167893	84.71
	Yes	30297	15.29
SUA	No	155685	78.55
	Yes	42505	21.45

Table 5F. Grade 8 Sample Characteristics (N = 199305)

Demographic Category		N-count	% of Total N-count
NRC	NYC	70186	35.30
	Big 4 cities	7467	3.76
	Urban/Suburban	13249	6.66
	Rural	11337	5.70
	Average needs	61364	30.87
	Low needs	32406	16.30
	Charter	2795	1.41
Ethnicity	Asian	15805	7.93
	Black	37354	18.74
	Hispanic	41960	21.05
	American Indian	970	0.49
	Multi-Racial	966	0.48
	Unknown	239	0.12
	White	102011	51.18
ELL	No	188587	94.62
	Yes	10718	5.38
SWD	No	169724	85.16
	Yes	29581	14.84
SUA	No	157734	79.14
	Yes	41571	20.86

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 that 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 0.21 to 0.97. For Grade 3, the item p-values were between 0.21 and 0.97 with a mean of 0.74. For Grade 4, the item p-values were between 0.28 and 0.93 with a mean of 0.70. For Grade 5, the item p-values were between 0.35 and 0.94 with a mean of 0.74. For Grade 6, the item p-values were between 0.48 and 0.97 with a mean of 0.73. For Grade 7, the item p-values were between 0.32 and 0.93 with a mean of 0.72. For Grade 8, the item p-values were between 0.26 and 0.96 with a mean of 0.72. 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 Typee	N-count	P-value	% Omit	Pbis Key
1	MC	194545	0.81	0.04	0.43
2	MC	194481	0.93	0.06	0.45
3	MC	194272	0.77	0.14	0.47
4	MC	194241	0.90	0.18	0.48
5	MC	193950	0.41	0.32	0.27
6	MC	193883	0.64	0.37	0.29
7	MC	193702	0.93	0.47	0.46
8	MC	194432	0.90	0.08	0.39
9	MC	194113	0.94	0.12	0.40
10	MC	194317	0.92	0.14	0.45
11	MC	194286	0.88	0.17	0.44
12	MC	194171	0.69	0.21	0.43
13	MC	194219	0.68	0.18	0.37
14	MC	194292	0.75	0.15	0.41
15	MC	194222	0.87	0.18	0.44
16	MC	194061	0.66	0.25	0.34
17	MC	194018	0.89	0.29	0.50
18	MC	193924	0.87	0.34	0.44
19	MC	193854	0.73	0.39	0.39
20	MC	194353	0.81	0.13	0.41
21	MC	194391	0.65	0.11	0.30
22	MC	194190	0.70	0.18	0.42
23	MC	194144	0.33	0.21	0.15
24	MC	194156	0.70	0.20	0.38
25	MC	194063	0.47	0.25	0.30
26	MC	193890	0.43	0.33	0.22
27	MC	193412	0.21	0.54	0.09
28	MC	194161	0.47	0.20	0.33
29	MC	194106	0.76	0.24	0.49
30	MC	193918	0.74	0.31	0.37
31	MC	193643	0.65	0.38	0.36
32	MC	193590	0.67	0.50	0.37
33	MC	193159	0.74	0.72	0.41
34	MC	191652	0.49	1.50	0.25
35	MC	190950	0.74	1.89	0.41
36	MC	194583	0.97	0.03	0.25

(Continued on next page)

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

Item	Item Type	N-count	P-value	% Omit	Pbis Key
37	MC	194446	0.87	0.06	0.32
38	MC	194447	0.95	0.07	0.30
39	MC	194429	0.95	0.08	0.30
40	MC	194370	0.82	0.14	0.43
41	CR	194095	0.74	0.29	
42	CR	194216	0.88	0.23	
43	CR	194024	0.71	0.33	
44	MC	194352	0.81	0.15	0.34
45	MC	194251	0.71	0.17	0.40
46	MC	194097	0.70	0.26	0.42
47	CR	194109	0.91	0.28	
48	CR	194038	0.89	0.32	
49	CR	194105	0.86	0.28	
50	CR	193836	0.69	0.42	
51	CR	193512	0.65	0.59	

Table 6B. Item Analysis, Grade 4

Item	Item Type	N-count	P-value	% Omit	Pbis Key
1	MC	195575	0.82	0.02	0.35
2	MC	195517	0.66	0.04	0.36
3	MC	195347	0.89	0.05	0.52
4	MC	195457	0.80	0.07	0.39
5	MC	195516	0.78	0.04	0.47
6	MC	195423	0.66	0.07	0.36
7	MC	195420	0.72	0.07	0.42
8	MC	195408	0.61	0.06	0.26
9	MC	195382	0.47	0.07	0.26
10	MC	195404	0.70	0.08	0.30
11	MC	195396	0.55	0.08	0.30
12	MC	195439	0.57	0.07	0.35
13	MC	195478	0.62	0.05	0.41
14	MC	195442	0.54	0.07	0.28
15	MC	195386	0.89	0.07	0.38
16	MC	195399	0.76	0.08	0.44

(Continued on next page)

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

Item	Item Type	N-count	P-value	% Omit	Pbis Key
17	MC	195468	0.92	0.07	0.36
18	MC	195425	0.56	0.07	0.50
19	MC	195311	0.52	0.12	0.32
20	MC	195436	0.75	0.08	0.35
21	MC	195465	0.92	0.08	0.43
22	MC	195392	0.81	0.09	0.47
23	MC	195387	0.69	0.09	0.51
24	MC	195335	0.91	0.12	0.41
25	MC	195236	0.75	0.12	0.35
26	MC	195290	0.74	0.14	0.30
27	MC	195327	0.78	0.13	0.49
28	MC	195139	0.87	0.20	0.35
29	MC	195149	0.86	0.22	0.45
30	MC	195081	0.82	0.24	0.49
31	MC	194978	0.82	0.27	0.41
32	MC	194964	0.60	0.31	0.23
33	MC	194919	0.70	0.34	0.47
34	MC	194701	0.77	0.44	0.44
35	MC	194609	0.82	0.47	0.46
36	MC	194451	0.57	0.55	0.43
37	MC	194368	0.75	0.61	0.39
38	MC	194260	0.65	0.66	0.42
39	MC	194242	0.81	0.68	0.41
40	MC	193798	0.79	0.90	0.42
41	MC	193776	0.49	0.91	0.37
42	MC	193566	0.28	1.03	0.17
43	MC	193323	0.36	1.17	0.22
44	MC	195560	0.75	0.03	0.18
45	MC	195553	0.93	0.03	0.33
46	MC	195443	0.77	0.06	0.25
47	MC	195455	0.91	0.07	0.32
48	MC	195287	0.61	0.17	0.24
49	CR	195328	0.70	0.16	
50	CR	195225	0.66	0.21	
51	CR	195079	0.51	0.28	
52	MC	195287	0.63	0.15	0.20

(Continued on next page)

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

Item	Item Type	N-count	P-value	% Omit	Pbis Key
53	MC	195209	0.90	0.18	0.37
54	MC	195074	0.59	0.26	0.20
55	CR	195367	0.71	0.14	
56	CR	194742	0.44	0.46	
57	CR	195130	0.70	0.26	
58	CR	194717	0.63	0.47	
59	CR	194706	0.60	0.47	

Table 6C. Item Analysis, Grade 5

Item	Item Type	N-count	P-value	% Omit	Pbis Key
1	MC	198606	0.94	0.02	0.30
2	MC	198568	0.70	0.02	0.29
3	MC	198486	0.69	0.05	0.45
4	MC	198418	0.52	0.09	0.13
5	MC	198516	0.70	0.05	0.42
6	MC	198485	0.67	0.06	0.28
7	MC	198398	0.74	0.10	0.43
8	MC	198543	0.92	0.04	0.30
9	MC	198417	0.92	0.06	0.40
10	MC	198492	0.89	0.06	0.33
11	MC	198530	0.83	0.04	0.32
12	MC	198501	0.74	0.05	0.40
13	MC	198439	0.75	0.07	0.51
14	MC	198410	0.71	0.09	0.42
15	MC	198449	0.83	0.08	0.47
16	MC	198422	0.74	0.09	0.44
17	MC	198192	0.35	0.20	0.12
18	MC	198329	0.60	0.12	0.31
19	MC	198372	0.73	0.12	0.45
20	MC	198245	0.45	0.18	0.19
21	MC	198419	0.81	0.10	0.32
22	MC	198350	0.68	0.13	0.31

(Continued on next page)

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

Item	Item Type	N-count	P-value	% Omit	Pbis Key
23	MC	198226	0.59	0.18	0.26
24	MC	198222	0.65	0.18	0.29
25	MC	198226	0.71	0.18	0.41
26	MC	197812	0.84	0.39	0.47
27	MC	197822	0.88	0.40	0.49
28	MC	197506	0.75	0.54	0.47
29	MC	197545	0.87	0.53	0.49
30	MC	197375	0.85	0.60	0.47
31	MC	196622	0.65	0.99	0.45
32	MC	196386	0.62	1.10	0.32
33	MC	196039	0.49	1.27	0.27
34	MC	195682	0.59	1.47	0.37
35	MC	195573	0.57	1.54	0.42
36	MC	198578	0.90	0.03	0.16
37	MC	198570	0.84	0.03	0.27
38	MC	198472	0.91	0.05	0.28
39	MC	198519	0.83	0.05	0.28
40	MC	198320	0.87	0.16	0.43
41	CR	198341	0.81	0.15	
42	CR	198353	0.75	0.15	
43	CR	198215	0.58	0.22	
44	MC	198325	0.77	0.15	0.35
45	MC	198253	0.69	0.17	0.37
46	MC	198213	0.88	0.20	0.25
47	CR	198341	0.90	0.15	
48	CR	198344	0.81	0.15	
49	CR	198336	0.81	0.16	
50	CR	198127	0.76	0.26	
51	CR	198072	0.73	0.29	

Table 6D. Item Analysis, Grade 6

Item	Item Type	N-count	P-value	% Omit	Pbis Key
1	MC	195701	0.65	0.10	0.15
2	MC	195837	0.96	0.02	0.30
3	MC	195821	0.97	0.03	0.31
4	MC	195744	0.71	0.06	0.31
5	MC	195746	0.82	0.07	0.52
6	MC	195813	0.94	0.04	0.29
7	MC	195713	0.52	0.08	0.35
8	MC	195669	0.69	0.10	0.44
9	MC	195675	0.63	0.09	0.40
10	MC	195723	0.79	0.07	0.40
11	MC	195641	0.55	0.12	0.31
12	MC	195769	0.86	0.05	0.47
13	MC	195700	0.87	0.09	0.46
14	MC	195744	0.84	0.06	0.49
15	MC	195630	0.56	0.11	0.32
16	MC	195696	0.61	0.08	0.36
17	MC	195741	0.88	0.06	0.42
18	MC	195591	0.81	0.14	0.39
19	MC	195673	0.62	0.08	0.38
20	MC	195663	0.83	0.10	0.48
21	MC	195673	0.78	0.10	0.54
22	MC	195647	0.78	0.11	0.52
23	MC	195641	0.85	0.11	0.45
24	MC	195656	0.76	0.11	0.44
25	MC	195605	0.58	0.13	0.36
26	MC	195587	0.90	0.14	0.40
27	MC	195550	0.57	0.17	0.31
28	MC	195523	0.50	0.17	0.33
29	MC	195569	0.56	0.15	0.36
30	MC	195520	0.84	0.16	0.38
31	MC	195467	0.69	0.18	0.42
32	MC	195207	0.67	0.34	0.40
33	MC	195187	0.65	0.34	0.42

(Continued on next page)

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

Item	Item Type	N-count	P-value	% Omit	Pbis Key
34	MC	195183	0.56	0.34	0.11
35	MC	195136	0.85	0.38	0.38
36	MC	194958	0.73	0.44	0.48
37	MC	194920	0.80	0.46	0.46
38	MC	194883	0.82	0.48	0.38
39	MC	194485	0.58	0.70	0.39
40	MC	194383	0.77	0.74	0.41
41	MC	193889	0.48	1.02	0.28
42	MC	195787	0.91	0.06	0.30
43	MC	195817	0.96	0.04	0.32
44	MC	195729	0.75	0.07	0.39
45	MC	195611	0.55	0.13	0.21
46	MC	195432	0.71	0.23	0.41
47	CR	195217	0.85	0.35	
48	CR	195448	0.88	0.23	
49	CR	194880	0.75	0.52	
50	MC	195674	0.82	0.10	0.45
51	MC	195614	0.58	0.12	0.32
52	MC	195590	0.88	0.15	0.39
53	CR	195608	0.80	0.15	
54	CR	195271	0.71	0.32	
55	CR	194759	0.61	0.58	
56	CR	194663	0.51	0.63	
57	CR	194760	0.69	0.58	

Table 6E. Item Analysis, Grade 7

Item	Item Type	N-count	P-value	% Omit	Pbis Key
1	MC	197995	0.80	0.09	0.47
2	MC	198041	0.68	0.06	0.27
3	MC	197933	0.70	0.11	0.36
4	MC	198041	0.84	0.06	0.51
5	MC	197769	0.58	0.20	0.43
6	MC	197976	0.84	0.08	0.43
7	MC	197985	0.84	0.08	0.40
8	MC	198059	0.88	0.05	0.45
9	MC	197946	0.83	0.11	0.50
10	MC	197849	0.70	0.16	0.26
11	MC	197855	0.53	0.15	0.34
12	MC	197821	0.47	0.18	0.22
13	MC	198014	0.62	0.07	0.42
14	MC	197605	0.52	0.28	0.29
15	MC	197997	0.93	0.08	0.39
16	MC	197952	0.84	0.09	0.52
17	MC	197947	0.86	0.11	0.43
18	MC	197954	0.81	0.11	0.42
19	MC	197919	0.75	0.11	0.51
20	MC	197880	0.85	0.14	0.46
21	MC	197921	0.84	0.12	0.53
22	MC	197938	0.91	0.12	0.42
23	MC	197640	0.60	0.26	0.44
24	MC	197683	0.56	0.23	0.42
25	MC	197624	0.54	0.26	0.33
26	MC	197561	0.54	0.29	0.42
27	MC	197431	0.38	0.34	0.03
28	MC	197398	0.32	0.38	0.23
29	MC	197427	0.67	0.37	0.33
30	MC	197127	0.80	0.52	0.44
31	MC	197061	0.74	0.55	0.36
32	MC	196800	0.73	0.68	0.39
33	MC	196856	0.92	0.65	0.47
34	MC	196696	0.62	0.73	0.25
35	MC	196399	0.72	0.85	0.48
36	MC	196265	0.64	0.93	0.47

(Continued on next page)

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

Item	Item Type	N-count	P-value	% Omit	Pbis Key
37	MC	195878	0.93	1.13	0.43
38	MC	195646	0.84	1.25	0.47
39	MC	195322	0.66	1.41	0.37
40	MC	195226	0.80	1.46	0.49
41	MC	194380	0.63	1.91	0.36
42	MC	198009	0.80	0.09	0.25
43	MC	198032	0.77	0.07	0.33
44	MC	197939	0.51	0.12	0.28
45	MC	197964	0.82	0.09	0.40
46	MC	197787	0.73	0.20	0.21
47	CR	197823	0.84	0.19	
48	CR	197635	0.80	0.28	
49	CR	197346	0.76	0.43	
50	MC	198005	0.89	0.09	0.33
51	MC	197910	0.60	0.12	0.31
52	MC	197841	0.73	0.16	0.32
53	CR	197437	0.79	0.38	
54	CR	195004	0.67	1.61	
55	CR	197416	0.89	0.39	
56	CR	197001	0.74	0.60	
57	CR	196847	0.69	0.68	

Table 6F. Item Analysis, Grade 8

Item	Item Type	N-count	P-value	% Omit	Pbis Key
1	MC	199188	0.68	0.06	0.29
2	MC	199212	0.69	0.04	0.31
3	MC	199132	0.73	0.08	0.42
4	MC	199174	0.86	0.05	0.43
5	MC	199188	0.66	0.05	0.48
6	MC	199168	0.81	0.06	0.47
7	MC	198939	0.59	0.17	0.24
8	MC	199115	0.83	0.08	0.43
9	MC	199087	0.81	0.09	0.47
10	MC	199160	0.85	0.06	0.47
11	MC	199094	0.80	0.09	0.44
12	MC	199192	0.86	0.04	0.42
13	MC	199111	0.78	0.08	0.39
14	MC	199022	0.60	0.13	0.42
15	MC	198987	0.63	0.15	0.38
16	MC	198978	0.63	0.15	0.30
17	MC	199129	0.91	0.08	0.44
18	MC	199013	0.80	0.13	0.40
19	MC	199066	0.84	0.11	0.35
20	MC	199065	0.80	0.11	0.44
21	MC	199073	0.89	0.11	0.38
22	MC	199023	0.93	0.12	0.40
23	MC	199037	0.87	0.13	0.43
24	MC	199009	0.92	0.12	0.34
25	MC	198996	0.69	0.14	0.31
26	MC	198871	0.75	0.20	0.45
27	MC	198962	0.83	0.15	0.47
28	MC	198775	0.49	0.25	0.26
29	MC	198907	0.88	0.19	0.45
30	MC	198864	0.69	0.21	0.38
31	MC	198532	0.67	0.37	0.47
32	MC	198558	0.39	0.35	0.33
33	MC	198453	0.26	0.41	0.13

(Continued on next page)

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

Item	Item Type	N-count	P-value	% Omit	Pbis Key
34	MC	198504	0.83	0.39	0.48
35	MC	198451	0.60	0.41	0.39
36	MC	198153	0.91	0.55	0.48
37	MC	197999	0.69	0.64	0.39
38	MC	197804	0.67	0.71	0.53
39	MC	197611	0.85	0.82	0.40
40	MC	197385	0.58	0.94	0.37
41	MC	196706	0.71	1.30	0.50
42	MC	199184	0.96	0.06	0.35
43	MC	199062	0.59	0.12	0.40
44	MC	198950	0.64	0.17	0.51
45	MC	199049	0.96	0.11	0.21
46	MC	198899	0.89	0.19	0.41
47	CR	198584	0.78	0.36	
48	CR	198454	0.69	0.43	
49	CR	197913	0.65	0.70	
50	MC	199031	0.46	0.13	0.37
51	MC	198851	0.38	0.21	0.22
52	MC	198941	0.55	0.16	0.40
53	CR	198702	0.73	0.30	
54	CR	198504	0.83	0.40	
55	CR	198264	0.74	0.52	
56	CR	196967	0.51	1.17	
57	CR	197293	0.73	1.01	

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 New York State test was 0.15. The point biserials for the correct answer option that was equal to or greater than 0.15 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 item number 27 in Grade 3 test, item number 4 and item number 17 in Grade 5 test, item number 34 in Grade 6 test, item

number 27 in Grade 7 test, and item number 33 in Grade 8 test. Point biserials for correct answer options on the tests ranged 0.03–0.54. For Grade 3, the pbis were between 0.09 and 0.50. For Grade 4, the pbis were between 0.17 and 0.52. For Grade 5, the pbis were between 0.12 and 0.51. For Grade 6, pbis were between 0.11 and 0.54. For Grade 7, the pbis were between 0.03 and 0.53. For Grade 8, the pbis were between 0.13 and 0.53.

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. Calculated Cronbach's alpha reliabilities ranged 0.91–0.92. Feldt-Raju reliability coefficients ranged 0.91–0.93. 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 error (for more information on test reliability and standard error of measurement, see Section VII, "Reliability and Standard Error of Measurement").

Table 7. NYSTP ELA 2011 Test Form Statistics and Reliability

Grade	Max RS	RS Mean	RS SD	P-value Mean	Cronbach's Alpha	Feldt-Raju
3	60	44.66	9.78	0.74	0.91	0.91
4	69	47.43	11.75	0.70	0.91	0.92
5	61	45.28	9.88	0.74	0.91	0.91
6	67	48.83	11.38	0.73	0.92	0.92
7	67	48.66	11.32	0.72	0.92	0.92
8	67	48.26	11.65	0.72	0.92	0.93

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 CTB/McGraw-Hill routinely conducts additional speededness analyses based on actual test data. The general rule of thumb is that omit rates should be less than 5.0%. Tables 6A–6F show the omit rates for items on the Grades 3–8 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 and one method of IRT DIF computation (described in Section VI) were employed because no single method can identify all DIF items on a test (Hambleton, Clauser, Mazer, and Jones, 1993).

Classical DIF analyses were conducted on subgroups of the needs resource category (focal group: High Needs; reference group: Low Needs), gender (focal group: Female; reference group: Male), ethnicity (focal groups: Black, Hispanic, and Asian; reference group: White), 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 number of cases for subgroups.

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

Grade	Ethnicity				Gender		Needs Resource Category		English Language Learner Status	
	Black/African American	Hispanic/ Latino	Asian	White	Female	Male	High	Low	Yes	No
3	36201	45334	15765	94454	95217	99442	105116	84447	18006	176653
4	36364	44276	15450	96976	95752	99883	104249	87329	16013	179622
5	36723	43537	16599	99269	97150	101494	102701	90504	13640	185004
6	36912	42740	15187	98656	95408	100491	100449	90394	11817	184082
7	37394	42405	15168	100886	97181	101009	101341	92800	10681	187509
8	37354	41960	15805	102011	97377	101928	102239	93770	10718	188587

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	10
5	6
6	10
7	11
8	8

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

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

where

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

and

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

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

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

and

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

where

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

and

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

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

Calibration Sample

The calibration sample included response data from both the OP form and the two FT anchor forms, each containing 12 items. The data containing student responses to items included in the FT anchor forms, administered approximately two weeks after the OP test to representative samples of NYS students, were collected and used for a purpose of equating 2011 OP tests to NYS OP scales as described in the "Scaling and Equating" subsection.

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

The cleaned sample data were used for calibration and scaling of New York State 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 2011 OP samples were comparable to 2010 populations in terms of needs resource 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	2010 Grade 3 Population	2011 Grade 3 Sample	2010 Grade 4 Population	2011 Grade 4 Sample
	%	%	%	%
NRC SUBGROUPS				
NYC	36.09	36.27	35.87	35.78
Big 4 cities	4.33	4.17	4.15	4.20
Urban/Suburban	8.07	7.90	8.03	7.50
Rural	5.91	5.73	5.82	5.87
Average needs	29.47	29.56	29.82	30.13
Low needs	14.00	13.87	14.58	14.56
Charter	2.13	2.51	1.73	1.95
ETHNICITY				
Asian	7.69	8.10	8.10	7.90
Black	18.94	18.60	18.96	18.59
Hispanics	22.24	23.29	21.53	22.63
American Indian	0.49	0.56	0.47	0.48
Multi-Racial	0.56	0.80	0.49	0.70
White	50.02	48.52	50.38	49.57
Unknown	0.06	0.14	0.06	0.13
ELL STATUS				
No	90.59	90.75	91.79	91.81
Yes	9.41	9.25	8.21	8.19
DISABILITY				
No	85.77	85.80	85.16	84.90
Yes	14.23	14.20	14.84	15.10
ACCOMMODATIONS				
No	76.01	75.21	75.68	75.02
Yes	23.99	24.79	24.32	24.98

Table 11. Grades 5 and 6 Demographic Statistics

Demographics	2010 Grade 5 Population	2011 Grade 5 Sample	2010 Grade 6 Population	2011 Grade 6 Sample
	%	%	%	%
NRC SUBGROUPS				
NYC	34.78	34.90	34.59	34.63
Big 4 cities	4.04	3.97	3.88	3.95
Urban/Suburban	7.81	7.09	7.53	6.99
Rural	5.91	5.81	5.91	5.77
Average needs	30.09	30.47	31.17	30.86
Low needs	15.08	15.15	15.04	15.35
Charter	2.28	2.61	1.88	2.43
ETHNICITY				
Asian	7.66	8.36	7.51	7.75
Black	19.09	18.49	18.99	18.84
Hispanics	21.23	21.92	20.92	21.82
American Indian	0.47	0.49	0.49	0.47
Multi-Racial	0.43	0.66	0.38	0.63
White	51.07	49.97	51.66	50.36
Unknown	0.05	0.12	0.05	0.13
ELL STATUS				
No	93.38	93.13	94.55	93.97
Yes	6.62	6.87	5.45	6.03
DISABILITY				
No	84.72	84.72	84.58	84.73
Yes	15.28	15.28	15.42	15.27
ACCOMMODATIONS				
No	76.25	75.60	78.31	77.57
Yes	23.75	24.40	21.69	22.43

Table 12. Grades 7 and 8 Demographic Statistics

Demographics	2010 Grade 7 Population	2011 Grade 7 Sample	2010 Grade 8 Population	2011 Grade 8 Sample
	%	%	%	%
NRC SUBGROUPS				
NYC	35.11	34.61	35.39	35.30
Big 4 cities	3.91	3.79	3.74	3.76
Urban/Suburban	7.26	7.06	7.28	6.66
Rural	5.99	5.78	5.98	5.70
Average needs	31.42	30.83	31.26	30.87
Low needs	14.84	16.09	15.19	16.30
Charter	1.47	1.84	1.16	1.41
ETHNICITY				
Asian	7.59	7.65	7.60	7.93
Black	19.00	18.87	18.82	18.74
Hispanics	20.47	21.40	20.53	21.05
American Indian	0.49	0.49	0.45	0.49
Multi-Racial	0.36	0.55	0.29	0.48
White	52.06	50.90	52.26	51.18
Unknown	0.04	0.13	0.05	0.12
ELL STATUS				
No	94.83	94.61	95.03	94.62
Yes	5.17	5.39	4.97	5.38
DISABILITY				
No	84.88	84.71	85.07	85.16
Yes	15.12	15.29	14.93	14.84
ACCOMMODATIONS				
No	79.41	78.55	79.51	79.14
Yes	20.59	21.45	20.49	20.86

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

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

Demographics	2011 Grade 3 FT Anchor Form 1	2011 Grade 3 FT Anchor Form 2	2011 Grade 3 OP Sample	2011 Grade 4 FT Anchor Form 1	2011 Grade 4 FT Anchor Form 2	2011 Grade 4 OP Sample
	%	%	%	%	%	%
NRC SUBGROUPS						
NYC	32.45	32.82	36.27	31.85	31.66	35.78
Big 4 cities	4.18	3.89	4.17	4.04	4.14	4.20
Urban/Suburban	7.43	8.02	7.90	7.16	8.05	7.50
Rural	5.90	5.63	5.73	5.71	5.44	5.87
Average needs	31.11	32.25	29.56	33.01	31.89	30.13
Low needs	16.74	15.80	13.87	16.62	17.10	14.56
Charter	2.18	1.59	2.51	1.62	1.72	1.95

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

Demographics	2011 Grade 5 FT Anchor Form 1	2011 Grade 5 FT Anchor Form 2	2011 Grade 5 OP Sample	2011 Grade 6 FT Anchor Form 1	2011 Grade 6 FT Anchor Form 2	2011 Grade 6 OP Sample
	%	%	%	%	%	%
NRC SUBGROUPS						
NYC	31.28	31.72	34.9	30.68	31.86	34.63
Big 4 cities	3.57	3.66	3.97	3.94	3.75	3.95
Urban/Suburban	7.25	7.49	7.09	7.10	7.97	6.99
Rural	5.66	6.12	5.81	5.97	5.93	5.77
Average needs	33.22	32.42	30.47	34.4	32.76	30.86
Low needs	17.02	16.48	15.15	16.05	15.84	15.35
Charter	2.02	2.11	2.61	1.86	1.90	2.43

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

Demographics	2011 Grade 7 FT Anchor Form 1	2011 Grade 7 FT Anchor Form 2	2011 Grade 7 OP Sample	2011 Grade 8 FT Anchor Form 1	2011 Grade 8 FT Anchor Form 2	2011 Grade 8 OP Sample
	%	%	%	%	%	%
NRC SUBGROUPS						
NYC	30.95	31.14	34.61	31.8	31.58	35.30
Big 4 cities	2.90	3.49	3.79	3.67	3.37	3.76
Urban/Suburban	6.70	6.69	7.06	5.75	5.94	6.66
Rural	5.85	5.70	5.78	6.06	7.02	5.70
Average needs	33.39	32.83	30.83	33.52	31.51	30.87
Low needs	18.17	18.1	16.09	17.92	19.16	16.30
Charter	2.04	2.04	1.84	1.28	1.42	1.41

Calibration Process

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

The NYSTP ELA Tests did not incur anything problematic during item calibration. The number of estimation cycles was set to 50 for all grades with convergence criterion of 0.001 for all grades. The maximum value of *a*-parameters was set to 3.4, and the range for *b*-parameters was set to be between -7.5 and 7.5. The maximum *c*-parameter value was set to 0.50. These are default parameters that have been used for calibration of NYS test data since its first administration in 1999. The estimated parameters were in the original theta metric, and all the items were well within the prescribed parameter ranges. A number of items on the OP test are set to the default value of the *c*-parameter. When the PARDUX program encounters difficulty estimating the *c*-parameter (guessing), it assigns a default *c*-parameter value of 0.200. For the Grades 3–8 ELA Tests, all calibration estimation results are reasonable. The summary of calibration results is presented in Table 16.

Table 16. NYSTP ELA 2011 Calibration Results

Grade	Largest *a*-parameter	*b*-parameter/ Gamma Range		# Items with Default *c*-parameter	Theta Mean	Theta Standard Deviation	# Students
3	2.270	-3.549	1.972	10	-0.14	1.155	194659
4	2.631	-2.371	1.584	12	-0.05	1.121	195635
5	2.481	-3.083	2.527	17	0.00	1.133	198644
6	2.440	-2.786	1.828	12	-0.10	1.134	195899
7	2.882	-2.195	2.834	12	-0.03	1.104	198190
8	2.183	-3.566	2.252	8	-0.13	1.149	199305

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 the2PPC 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 CTB/McGraw-Hill's 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.

Items were considered to have poor fit if the value of the obtained Z_{Q_1} was greater than the value of Z_{Q_1} critical. If the obtained Z_{Q_1} was less than Z_{Q_1} critical, the items were rated as having acceptable fit. All items in the NYSTP 2011 ELA Tests for Grades 6, 7 and 8 demonstrated good model fit. Item 30 in Grade 3, item 1 in Grade 4, and item 6 in Grade 5 exhibited poor item-model fit statistics. The fact that so few items were flagged for poor fit across all NYSTP 2011 ELA Tests further supports the use of the chosen models. Item fit statistics are presented in Tables 17–22.

Table 17. ELA Grade 3 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	Z-observed	Z-critical	Fit OK?
1	3PL	1177.16	7	194328	312.74	518.21	Y
2	3PL	67.13	7	194328	16.07	518.21	Y
3	3PL	225.94	7	194328	58.51	518.21	Y
4	3PL	48.61	7	194328	11.12	518.21	Y
5	3PL	293.24	7	194328	76.50	518.21	Y
6	3PL	899.17	7	194328	238.44	518.21	Y
7	3PL	262.65	7	194328	68.33	518.21	Y
8	3PL	330.97	7	194328	86.59	518.21	Y
9	3PL	151.23	7	194328	38.55	518.21	Y
10	3PL	81.12	7	194328	19.81	518.21	Y
11	3PL	46.48	7	194328	10.55	518.21	Y
12	3PL	135.59	7	194328	34.37	518.21	Y
13	3PL	49.94	7	194328	11.48	518.21	Y
14	3PL	77.52	7	194328	18.85	518.21	Y
15	3PL	36.16	7	194328	7.79	518.21	Y
16	3PL	64.87	7	194328	15.47	518.21	Y
17	3PL	44.78	7	194328	10.10	518.21	Y
18	3PL	80.64	7	194328	19.68	518.21	Y
19	3PL	140.57	7	194328	35.70	518.21	Y
20	3PL	108.18	7	194328	27.04	518.21	Y
21	3PL	181.90	7	194328	46.74	518.21	Y
22	3PL	120.33	7	194328	30.29	518.21	Y
23	3PL	64.29	7	194328	15.31	518.21	Y

(Continued on next page)

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

Item	Model	Chi Square	DF	Total N	*Z*-observed	*Z*-critical	Fit OK?
24	3PL	196.40	7	194328	50.62	518.21	Y
25	3PL	154.63	7	194328	39.45	518.21	Y
26	3PL	149.65	7	194328	38.13	518.21	Y
27	3PL	773.01	7	194328	204.73	518.21	Y
28	3PL	273.80	7	194328	71.31	518.21	Y
29	3PL	144.97	7	194328	36.88	518.21	Y
30	3PL	2807.61	7	194328	748.50	518.21	N
31	3PL	71.81	7	194328	17.32	518.21	Y
32	3PL	110.22	7	194328	27.59	518.21	Y
33	3PL	77.89	7	194328	18.95	518.21	Y
34	3PL	294.81	7	194328	76.92	518.21	Y
35	3PL	148.84	7	194328	37.91	518.21	Y
36	3PL	97.66	7	194328	24.23	518.21	Y
37	3PL	79.20	7	194328	19.30	518.21	Y
38	3PL	106.42	7	194328	26.57	518.21	Y
39	3PL	82.28	7	194328	20.12	518.21	Y
40	3PL	134.14	7	194328	33.98	518.21	Y
41	2PPC	826.60	17	194328	138.85	518.21	Y
42	2PPC	377.20	17	194328	61.77	518.21	Y
43	2PPC	930.78	17	194328	156.71	518.21	Y
44	3PL	17.75	7	194328	2.87	518.21	Y
45	3PL	97.14	7	194328	24.09	518.21	Y
46	3PL	269.10	7	194328	70.05	518.21	Y
47	2PPC	271.00	17	194328	43.56	518.21	Y
48	2PPC	394.50	17	194328	64.74	518.21	Y
49	2PPC	805.08	17	194328	135.15	518.21	Y
50	2PPC	574.45	17	194328	95.60	518.21	Y
51	2PPC	1253.08	26	194328	170.16	518.21	Y

Table 18. ELA Grade 4 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	*Z*-observed	*Z*-critical	Fit OK?
1	3PL	2181.99	7	195514	581.29	521.37	N
2	3PL	72.74	7	195514	17.57	521.37	Y
3	3PL	100.97	7	195514	25.11	521.37	Y
4	3PL	64.17	7	195514	15.28	521.37	Y
5	3PL	31.28	7	195514	6.49	521.37	Y
6	3PL	22.90	7	195514	4.25	521.37	Y
7	3PL	84.64	7	195514	20.75	521.37	Y
8	3PL	87.71	7	195514	21.57	521.37	Y
9	3PL	34.86	7	195514	7.45	521.37	Y

(Continued on next page)

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

Item	Model	Chi Square	DF	Total N	*Z*-observed	*Z*-critical	Fit OK?
10	3PL	17.84	7	195514	2.90	521.37	Y
11	3PL	116.74	7	195514	29.33	521.37	Y
12	3PL	34.92	7	195514	7.46	521.37	Y
13	3PL	59.55	7	195514	14.05	521.37	Y
14	3PL	296.92	7	195514	77.48	521.37	Y
15	3PL	141.94	7	195514	36.06	521.37	Y
16	3PL	320.02	7	195514	83.66	521.37	Y
17	3PL	97.60	7	195514	24.21	521.37	Y
18	3PL	343.61	7	195514	89.96	521.37	Y
19	3PL	146.04	7	195514	37.16	521.37	Y
20	3PL	66.39	7	195514	15.87	521.37	Y
21	3PL	92.26	7	195514	22.79	521.37	Y
22	3PL	27.16	7	195514	5.39	521.37	Y
23	3PL	131.28	7	195514	33.22	521.37	Y
24	3PL	146.26	7	195514	37.22	521.37	Y
25	3PL	84.49	7	195514	20.71	521.37	Y
26	3PL	67.79	7	195514	16.25	521.37	Y
27	3PL	99.59	7	195514	24.74	521.37	Y
28	3PL	201.28	7	195514	51.92	521.37	Y
29	3PL	142.38	7	195514	36.18	521.37	Y
30	3PL	81.84	7	195514	20.00	521.37	Y
31	3PL	34.43	7	195514	7.33	521.37	Y
32	3PL	25.94	7	195514	5.06	521.37	Y
33	3PL	95.73	7	195514	23.71	521.37	Y
34	3PL	44.10	7	195514	9.92	521.37	Y
35	3PL	40.55	7	195514	8.97	521.37	Y
36	3PL	142.18	7	195514	36.13	521.37	Y
37	3PL	71.38	7	195514	17.21	521.37	Y
38	3PL	150.74	7	195514	38.42	521.37	Y
39	3PL	77.99	7	195514	18.97	521.37	Y
40	3PL	144.51	7	195514	36.75	521.37	Y
41	3PL	97.89	7	195514	24.29	521.37	Y
42	3PL	254.36	7	195514	66.11	521.37	Y
43	3PL	41.94	7	195514	9.34	521.37	Y
44	3PL	88.53	7	195514	21.79	521.37	Y
45	3PL	107.46	7	195514	26.85	521.37	Y
46	3PL	185.11	7	195514	47.60	521.37	Y
47	3PL	100.36	7	195514	24.95	521.37	Y
48	3PL	128.92	7	195514	32.58	521.37	Y
49	2PPC	262.54	17	195514	42.11	521.37	Y
50	2PPC	326.42	17	195514	53.07	521.37	Y

(Continued on next page)

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

Item	Model	Chi Square	DF	Total N	Z-observed	Z-critical	Fit OK?
51	2PPC	409.46	17	195514	67.31	521.37	Y
52	3PL	76.94	7	195514	18.69	521.37	Y
53	3PL	81.76	7	195514	19.98	521.37	Y
54	3PL	76.95	7	195514	18.69	521.37	Y
55	2PPC	361.18	17	195514	59.03	521.37	Y
56	2PPC	251.55	17	195514	40.22	521.37	Y
57	2PPC	1148.70	17	195514	194.09	521.37	Y
58	2PPC	2202.37	17	195514	374.79	521.37	Y
59	2PPC	720.66	35	195514	81.95	521.37	Y

Table 19. ELA Grade 5 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	Z-observed	Z-critical	Fit OK?
1	3PL	108.40	7	198430	27.10	529.15	Y
2	3PL	128.99	7	198430	32.60	529.15	Y
3	3PL	76.00	7	198430	18.44	529.15	Y
4	3PL	81.37	7	198430	19.88	529.15	Y
5	3PL	148.57	7	198430	37.84	529.15	Y
6	3PL	2022.69	7	198430	538.71	529.15	N
7	3PL	46.49	7	198430	10.56	529.15	Y
8	3PL	92.14	7	198430	22.76	529.15	Y
9	3PL	156.63	7	198430	39.99	529.15	Y
10	3PL	88.32	7	198430	21.73	529.15	Y
11	3PL	52.09	7	198430	12.05	529.15	Y
12	3PL	103.03	7	198430	25.66	529.15	Y
13	3PL	155.70	7	198430	39.74	529.15	Y
14	3PL	94.66	7	198430	23.43	529.15	Y
15	3PL	181.86	7	198430	46.73	529.15	Y
16	3PL	42.01	7	198430	9.36	529.15	Y
17	3PL	63.92	7	198430	15.21	529.15	Y
18	3PL	30.28	7	198430	6.22	529.15	Y
19	3PL	77.62	7	198430	18.87	529.15	Y
20	3PL	1255.79	7	198430	333.75	529.15	Y
21	3PL	69.21	7	198430	16.63	529.15	Y
22	3PL	87.66	7	198430	21.56	529.15	Y
23	3PL	166.40	7	198430	42.60	529.15	Y
24	3PL	1981.25	7	198430	527.64	529.15	Y
25	3PL	62.77	7	198430	14.90	529.15	Y
26	3PL	140.44	7	198430	35.66	529.15	Y
27	3PL	157.65	7	198430	40.26	529.15	Y
28	3PL	88.40	7	198430	21.75	529.15	Y

(Continued on next page)

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

Item	Model	Chi Square	DF	Total N	*Z*-observed	*Z*-critical	Fit OK?
29	3PL	129.47	7	198430	32.73	529.15	Y
30	3PL	72.82	7	198430	17.59	529.15	Y
31	3PL	276.39	7	198430	72.00	529.15	Y
32	3PL	37.87	7	198430	8.25	529.15	Y
33	3PL	53.65	7	198430	12.47	529.15	Y
34	3PL	302.07	7	198430	78.86	529.15	Y
35	3PL	238.74	7	198430	61.94	529.15	Y
36	3PL	344.83	7	198430	90.29	529.15	Y
37	3PL	86.54	7	198430	21.26	529.15	Y
38	3PL	42.69	7	198430	9.54	529.15	Y
39	3PL	222.49	7	198430	57.59	529.15	Y
40	3PL	76.22	7	198430	18.50	529.15	Y
41	2PPC	217.94	17	198430	34.46	529.15	Y
42	2PPC	1547.67	17	198430	262.51	529.15	Y
43	2PPC	1822.19	17	198430	309.59	529.15	Y
44	3PL	80.71	7	198430	19.70	529.15	Y
45	3PL	152.01	7	198430	38.75	529.15	Y
46	3PL	78.90	7	198430	19.22	529.15	Y
47	2PPC	251.44	17	198430	40.21	529.15	Y
48	2PPC	648.25	17	198430	108.26	529.15	Y
49	2PPC	250.41	17	198430	40.03	529.15	Y
50	2PPC	335.79	17	198430	54.67	529.15	Y
51	2PPC	694.00	35	198430	78.77	529.15	Y

Table 20. ELA Grade 6 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	*Z*-observed	*Z*-critical	Fit OK?
1	3PL	165.98	7	195515	42.49	521.37	Y
2	3PL	119.86	7	195515	30.16	521.37	Y
3	3PL	111.93	7	195515	28.04	521.37	Y
4	3PL	22.53	7	195515	4.15	521.37	Y
5	3PL	144.06	7	195515	36.63	521.37	Y
6	3PL	451.87	7	195515	118.90	521.37	Y
7	3PL	747.19	7	195515	197.82	521.37	Y
8	3PL	80.58	7	195515	19.67	521.37	Y
9	3PL	20.23	7	195515	3.54	521.37	Y
10	3PL	184.91	7	195515	47.55	521.37	Y
11	3PL	770.42	7	195515	204.03	521.37	Y
12	3PL	95.08	7	195515	23.54	521.37	Y
13	3PL	69.49	7	195515	16.70	521.37	Y
14	3PL	39.09	7	195515	8.58	521.37	Y

(Continued on next page)

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

Item	Model	Chi Square	DF	Total N	*Z*-observed	*Z*-critical	Fit OK?
15	3PL	62.25	7	195515	14.76	521.37	Y
16	3PL	94.32	7	195515	23.34	521.37	Y
17	3PL	21.38	7	195515	3.84	521.37	Y
18	3PL	37.94	7	195515	8.27	521.37	Y
19	3PL	160.20	7	195515	40.94	521.37	Y
20	3PL	73.38	7	195515	17.74	521.37	Y
21	3PL	92.18	7	195515	22.76	521.37	Y
22	3PL	43.85	7	195515	9.85	521.37	Y
23	3PL	38.37	7	195515	8.38	521.37	Y
24	3PL	38.49	7	195515	8.42	521.37	Y
25	3PL	42.50	7	195515	9.49	521.37	Y
26	3PL	115.68	7	195515	29.05	521.37	Y
27	3PL	17.56	7	195515	2.82	521.37	Y
28	3PL	196.48	7	195515	50.64	521.37	Y
29	3PL	121.31	7	195515	30.55	521.37	Y
30	3PL	126.28	7	195515	31.88	521.37	Y
31	3PL	123.06	7	195515	31.02	521.37	Y
32	3PL	293.54	7	195515	76.58	521.37	Y
33	3PL	311.45	7	195515	81.37	521.37	Y
34	3PL	62.39	7	195515	14.80	521.37	Y
35	3PL	34.92	7	195515	7.46	521.37	Y
36	3PL	249.24	7	195515	64.74	521.37	Y
37	3PL	72.24	7	195515	17.44	521.37	Y
38	3PL	16.11	7	195515	2.44	521.37	Y
39	3PL	156.19	7	195515	39.87	521.37	Y
40	3PL	151.50	7	195515	38.62	521.37	Y
41	3PL	134.43	7	195515	34.06	521.37	Y
42	3PL	199.63	7	195515	51.48	521.37	Y
43	3PL	162.41	7	195515	41.53	521.37	Y
44	3PL	195.46	7	195515	50.37	521.37	Y
45	3PL	57.28	7	195515	13.44	521.37	Y
46	3PL	118.44	7	195515	29.78	521.37	Y
47	2PPC	138.76	17	195515	20.88	521.37	Y
48	2PPC	491.38	17	195515	81.35	521.37	Y
49	2PPC	1973.71	17	195515	335.57	521.37	Y
50	3PL	97.16	7	195515	24.10	521.37	Y
51	3PL	54.34	7	195515	12.65	521.37	Y
52	3PL	94.87	7	195515	23.49	521.37	Y
53	2PPC	1553.55	17	195515	263.52	521.37	Y
54	2PPC	420.19	17	195515	69.15	521.37	Y
55	2PPC	187.18	17	195515	29.19	521.37	Y
56	2PPC	293.21	17	195515	47.37	521.37	Y
57	2PPC	1479.47	35	195515	172.65	521.37	Y

Table 21. ELA Grade 7 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	*Z*-observed	*Z*-critical	Fit OK?
1	3PL	60.49	7	198014	14.30	528.04	Y
2	3PL	301.67	7	198014	78.75	528.04	Y
3	3PL	34.94	7	198014	7.47	528.04	Y
4	3PL	97.21	7	198014	24.11	528.04	Y
5	3PL	84.27	7	198014	20.65	528.04	Y
6	3PL	141.70	7	198014	36.00	528.04	Y
7	3PL	101.44	7	198014	25.24	528.04	Y
8	3PL	79.84	7	198014	19.47	528.04	Y
9	3PL	66.57	7	198014	15.92	528.04	Y
10	3PL	77.75	7	198014	18.91	528.04	Y
11	3PL	68.75	7	198014	16.50	528.04	Y
12	3PL	50.59	7	198014	11.65	528.04	Y
13	3PL	37.51	7	198014	8.15	528.04	Y
14	3PL	51.64	7	198014	11.93	528.04	Y
15	3PL	39.35	7	198014	8.65	528.04	Y
16	3PL	107.88	7	198014	26.96	528.04	Y
17	3PL	19.19	7	198014	3.26	528.04	Y
18	3PL	154.16	7	198014	39.33	528.04	Y
19	3PL	70.88	7	198014	17.07	528.04	Y
20	3PL	39.53	7	198014	8.69	528.04	Y
21	3PL	104.27	7	198014	26.00	528.04	Y
22	3PL	150.62	7	198014	38.38	528.04	Y
23	3PL	186.84	7	198014	48.06	528.04	Y
24	3PL	133.03	7	198014	33.68	528.04	Y
25	3PL	121.11	7	198014	30.50	528.04	Y
26	3PL	194.64	7	198014	50.15	528.04	Y
27	3PL	1354.66	7	198014	360.18	528.04	Y
28	3PL	1459.07	7	198014	388.08	528.04	Y
29	3PL	70.15	7	198014	16.88	528.04	Y
30	3PL	21.82	7	198014	3.96	528.04	Y
31	3PL	25.19	7	198014	4.86	528.04	Y
32	3PL	102.79	7	198014	25.60	528.04	Y
33	3PL	114.59	7	198014	28.75	528.04	Y
34	3PL	252.97	7	198014	65.74	528.04	Y
35	3PL	315.23	7	198014	82.38	528.04	Y
36	3PL	337.98	7	198014	88.46	528.04	Y
37	3PL	105.32	7	198014	26.28	528.04	Y
38	3PL	100.64	7	198014	25.03	528.04	Y
39	3PL	31.53	7	198014	6.56	528.04	Y
40	3PL	183.20	7	198014	47.09	528.04	Y
41	3PL	36.50	7	198014	7.88	528.04	Y

(Continued on next page)

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

Item	Model	Chi Square	DF	Total N	*Z*-observed	*Z*-critical	Fit OK?
42	3PL	122.34	7	198014	30.82	528.04	Y
43	3PL	38.71	7	198014	8.47	528.04	Y
44	3PL	67.07	7	198014	16.05	528.04	Y
45	3PL	101.57	7	198014	25.28	528.04	Y
46	3PL	367.33	7	198014	96.30	528.04	Y
47	2PPC	158.94	17	198014	24.34	528.04	Y
48	2PPC	623.53	17	198014	104.02	528.04	Y
49	2PPC	149.98	17	198014	22.81	528.04	Y
50	3PL	52.26	7	198014	12.10	528.04	Y
51	3PL	88.09	7	198014	21.67	528.04	Y
52	3PL	107.40	7	198014	26.83	528.04	Y
53	2PPC	489.60	17	198014	81.05	528.04	Y
54	2PPC	1116.08	17	198014	188.49	528.04	Y
55	2PPC	231.94	17	198014	36.86	528.04	Y
56	2PPC	518.36	17	198014	85.98	528.04	Y
57	2PPC	884.18	35	198014	101.50	528.04	Y

Table 22. ELA Grade 8 Item Fit Statistics

Item	Model	Chi Square	DF	Total N	*Z*-observed	*Z*-critical	Fit OK?
1	3PL	82.72	7	199015	20.24	530.71	Y
2	3PL	254.29	7	199015	66.09	530.71	Y
3	3PL	60.72	7	199015	14.36	530.71	Y
4	3PL	86.95	7	199015	21.37	530.71	Y
5	3PL	69.97	7	199015	16.83	530.71	Y
6	3PL	45.78	7	199015	10.36	530.71	Y
7	3PL	70.14	7	199015	16.88	530.71	Y
8	3PL	24.66	7	199015	4.72	530.71	Y
9	3PL	48.67	7	199015	11.14	530.71	Y
10	3PL	45.05	7	199015	10.17	530.71	Y
11	3PL	49.51	7	199015	11.36	530.71	Y
12	3PL	28.51	7	199015	5.75	530.71	Y
13	3PL	595.46	7	199015	157.27	530.71	Y
14	3PL	142.28	7	199015	36.16	530.71	Y
15	3PL	129.18	7	199015	32.65	530.71	Y
16	3PL	655.88	7	199015	173.42	530.71	Y
17	3PL	168.09	7	199015	43.05	530.71	Y
18	3PL	49.14	7	199015	11.26	530.71	Y
19	3PL	1062.06	7	199015	281.98	530.71	Y
20	3PL	119.80	7	199015	30.15	530.71	Y
21	3PL	28.78	7	199015	5.82	530.71	Y

(Continued on next page)

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

Item	Model	Chi Square	DF	Total N	Z-observed	Z-critical	Fit OK?
22	3PL	65.99	7	199015	15.77	530.71	Y
23	3PL	42.42	7	199015	9.47	530.71	Y
24	3PL	190.80	7	199015	49.12	530.71	Y
25	3PL	105.11	7	199015	26.22	530.71	Y
26	3PL	62.35	7	199015	14.79	530.71	Y
27	3PL	43.93	7	199015	9.87	530.71	Y
28	3PL	118.62	7	199015	29.83	530.71	Y
29	3PL	43.13	7	199015	9.66	530.71	Y
30	3PL	84.27	7	199015	20.65	530.71	Y
31	3PL	228.85	7	199015	59.29	530.71	Y
32	3PL	185.04	7	199015	47.58	530.71	Y
33	3PL	68.54	7	199015	16.45	530.71	Y
34	3PL	84.91	7	199015	20.82	530.71	Y
35	3PL	311.09	7	199015	81.27	530.71	Y
36	3PL	98.55	7	199015	24.47	530.71	Y
37	3PL	18.49	7	199015	3.07	530.71	Y
38	3PL	294.88	7	199015	76.94	530.71	Y
39	3PL	1060.47	7	199015	281.55	530.71	Y
40	3PL	178.43	7	199015	45.82	530.71	Y
41	3PL	100.68	7	199015	25.04	530.71	Y
42	3PL	100.35	7	199015	24.95	530.71	Y
43	3PL	29.71	7	199015	6.07	530.71	Y
44	3PL	169.19	7	199015	43.35	530.71	Y
45	3PL	149.61	7	199015	38.11	530.71	Y
46	3PL	39.85	7	199015	8.78	530.71	Y
47	2PPC	200.50	17	199015	31.47	530.71	Y
48	2PPC	1864.59	17	199015	316.86	530.71	Y
49	2PPC	365.77	17	199015	59.81	530.71	Y
50	3PL	68.09	7	199015	16.33	530.71	Y
51	3PL	129.02	7	199015	32.61	530.71	Y
52	3PL	87.76	7	199015	21.58	530.71	Y
53	2PPC	198.56	17	199015	31.14	530.71	Y
54	2PPC	465.38	17	199015	76.90	530.71	Y
55	2PPC	371.39	17	199015	60.78	530.71	Y
56	2PPC	1576.77	17	199015	267.50	530.71	Y
57	2PPC	1543.66	35	199015	180.32	530.71	Y

Local Independence

In using IRT models, one of the assumptions made is that the items are locally independent. That is, 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

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

The generalization to items with multiple response categories uses

$$d_{ja} \equiv x_{ja} - E_{ja},$$

where

$$E_{ja} \equiv E\left(x\middle|\hat{\theta}_a\right) = \sum_{k=1}^{m_j} kP_{jk2}\left(\hat{\theta}_a\right).$$

If a substantial number of items in the test demonstrate local dependence, these items may need to be calibrated separately. All pairs of items with Q_3 values greater than 0.20 were classified as locally dependent. The maximum value for this index is 1.00. 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 locally dependent in Grades 3 and 6. In Grade 4, two pairs of items are found to be locally dependent: items 14 and 19 (Q_3 = 0.284) and item 18 and 19 (Q_3 = 0.218). In Grade 5, one pair of items was found to be locally dependent: items 49 and 50 (Q_3 = 0.245). In Grade 7, two pairs of items were found to be locally dependent: items 48 and 49 (Q_3 = 0.215) and items 55 and 56 (Q_3 = 0.211). In Grade 8, two pairs of items were found to be locally dependent: items 48 and 49 (Q_3 = 0.296) and items 55 and 56 (Q_3 = 0.25). The magnitudes of these statistics were not sufficient to warrant any concern. Anchor items were excluded from Q_3 computation.

Scaling and Equating

The 2011 Grades 3–8 ELA Tests were calibrated and equated to the OP scales, using two separate equating procedures.

In the first equating procedure, the new 2011 OP forms were pre-equated to the corresponding 2010 assessments. Prior to pre-equating, FT items that were eligible for future OP administrations were then included in the NYS item pool. Items in the NYS item pool were items field-tested in 2009, 2008, 2007, 2006, and 2005. All items field-tested between 2005 and 2009 were equated to the NYS OP scales. For more details on equating of FT items

to the NYS OP scales, refer to *New York State Testing Program 2006: English Language Arts Grades 3–8*, page 56. The pool also included items owned by CTB/McGraw-Hill. These items consisted mostly of *TerraNova* items but also included items field-tested in New York State in 2010. *TerraNova* items were also equated to NYS OP scales.

At the pre-equating stage, the pool of FT items administered in years 2005 to 2009 and *TerraNova* items equated to the NYS OP scale were used to select the 2011 OP test forms using the following criteria:

- Content coverage of each form matched the test blueprint
- Psychometric properties of the items:
  - item fit
  - differential item functioning
  - item difficulty
  - item discrimination
  - omit rates
- Test characteristic curve (TCC) and standard error (SE) curve alignment of the 2011 forms with the target 2010 OP forms. (Note that the 2010 OP TCC and SE curves were based on OP parameters and the 2011 TCC and SE curves were based on FT parameters transformed to the OP scale.)

Although it was not possible to entirely avoid including flagged items in OP tests, the number of flagged items included in OP tests was small and content of all flagged items was carefully reviewed.

In the second equating procedure, the 2011 ELA OP data were re-calibrated after the 2011 OP administration. The equating data file included both the OP data and FT anchor forms data, the FT anchor records were matched to OP test data in two phases: exact match and fuzzy match. An exact match occurs when the school Bedscode (school unique ID) and student ID in both OP and FT data are the same. Fuzzy match includes all the following conditions:

a) at least ten characters of last name match (including blank spaces)
b) at least five characters of first name match (including blank spaces)
c) gender must be the same or one must be blank
d) school Bedscode must be the same or one must be blank
e) two of three parts of date of birth (MM or DD or YY) must be the same or one must be blank

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

The MC items contained in the FT anchor sets were representative of the content of the entire test for each grade. The equating was performed using a test characteristic curve (TCC) method (Stocking and Lord, 1983). TCC methods find the linear transformation (*M1* and *M2*) that transforms the original item parameter estimates (in theta metric) to the scale score metric and minimizes the difference in the relationship between raw scores and ability

estimates (i.e., TCC) defined by the FT anchor item parameter estimates from year 2010 and that relationship defined by the FT anchor item parameter estimates in new administration year 2011. This places the transformed parameters for the OP test items onto the New York State OP scale. In this procedure, new 2011 OP parameter estimates were obtained for all items. For the FT anchor items, the *a*-parameters and *b*-parameters were re-estimated within specified constraints (as described in "Calibration Process" subsection) while *c*-parameters of anchor items were fixed to their 2010 values.

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

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

where

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

The *A* and *B* values are derived from the input (2010 FT anchor parameter estimates) and estimate (2011 FT anchor parameter estimates) values of anchor items. Anchor input values are known item parameter estimates entered into equating. Anchor estimate values are parameter estimates for the same anchor items re-estimated during the equating procedure. The input and estimate anchor parameter estimates are expected to have similar values.

The *M1* and *M2* transformation parameters obtained in the Stocking and Lord equating process were used to transform item parameters obtained in a calibration process onto the final scale score metric. Table 23 presents the 2011 OP transformation constants for New York State Grades 3–8 ELA Tests.

Table 23. NYSTP ELA 2011 Final Transformation Constants

Grade	*M1*	*M2*
3	16.63	666.08
4	23.91	673.43
5	15.73	668.07
6	14.66	664.18
7	16.23	664.42
8	18.01	657.77

Anchor Item Security

In order for an equating to accurately place the items and forms onto the OP scale, it is important to keep the anchor items secure and to reduce anchor item exposure to students and teachers. Although the FT anchor forms were administered in four consecutive years: 2008,

2009, 2010, and 2011, they were administered only to small groups of NYS students each year. The FT anchor forms were developed, administered, collected, and scanned by CTB/McGraw-Hill. Given the "secure" status of these FT anchor forms, there is reason to believe that the item exposure effect was minimal.

Anchor Item Evaluation

Anchor items were evaluated using several procedures. Procedures 1 and 2 evaluate the overall anchor set, while procedure 3 evaluates individual anchor items.

1. Anchor set input and estimates of TCC alignment. The overall alignment of TCCs for the anchor set input and estimates was evaluated to determine the overall stability of anchor item parameters between 2010 and 2011 FT anchor form administrations.

2. Correlations of anchor input and estimates of *a*- and *b*-parameters. Correlations of anchor input and estimate of *a*- and *b*-parameters were evaluated for magnitude. Ideally, the correlations between anchor input and estimate for *a*-parameter should be at least 0.80 and the correlations for *b*-parameters should be at least 0.90. Items contributing to lower than expected correlations were flagged.

3. Iterative linking using Stocking and Lord's TCC method. This procedure, called the TCC method, minimizes the mean squared difference between the two TCCs: one based on 2010 FT anchor estimates and the other on transformed estimates from the 2011 equating of OP test forms. Differential item performance was evaluated by examining previous (input) and transformed (estimated) item parameters. Items with an absolute difference of parameters greater than two times the root mean square deviation were flagged.

In all cases, the overall TCC alignment for anchor set input and estimate was good. Correlations for *b*-parameter input and estimates ranged from 0.97 for Grade 7 to 1.00 for Grade 3. Correlations for *a*-parameter input and estimate ranged from 0.88 for Grades 7 and 8 to 0.97 for Grade 6. Correlations between *a*-parameter input and estimates and correlations between *b*-parameter input and estimates were above the NYS criterion for all grades.

Overall, TCC alignment for anchor set input and estimate was very good; correlations between parameter input and estimates were well above NYS criterion for Grades 3–8; in addition, no items was flagged using the Stocking and Lord's TCC method either. Therefore, no anchors were removed from any of the anchor sets.

Item Parameters

The OP test item parameters were estimated by the software PARDUX (Burket, 2002) and are presented in Tables 24–29. The parameter estimates are expressed in scale score 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 m_j in scale score metric for CR items.

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

Table 24. 2011 Operational Item Parameter Estimates, Grade 3

Item	Max Pts	*a*-par/alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3
1	1	0.045	643.598	0.230	
2	1	0.076	634.613	0.291	
3	1	0.072	653.580	0.268	
4	1	0.071	637.544	0.210	
5	1	0.060	680.061	0.217	
6	1	0.027	658.882	0.200	
7	1	0.063	630.381	0.124	
8	1	0.045	631.424	0.230	
9	1	0.055	623.330	0.064	
10	1	0.061	630.034	0.098	
11	1	0.050	632.944	0.060	
12	1	0.057	657.690	0.217	
13	1	0.039	654.856	0.167	
14	1	0.040	645.770	0.100	
15	1	0.053	637.459	0.168	
16	1	0.029	649.541	0.050	
17	1	0.070	638.323	0.153	
18	1	0.056	640.505	0.251	
19	1	0.041	650.557	0.174	
20	1	0.045	643.695	0.181	
21	1	0.055	667.985	0.384	
22	1	0.054	656.813	0.222	
23	1	0.032	698.855	0.197	
24	1	0.066	661.824	0.352	
25	1	0.051	675.748	0.219	
26	1	0.053	683.302	0.266	
27	1	0.080	697.235	0.157	
28	1	0.053	673.820	0.179	
29	1	0.068	652.078	0.205	
30	1	0.035	649.879	0.230	
31	1	0.042	660.575	0.218	
32	1	0.048	660.992	0.268	
33	1	0.045	651.311	0.187	
34	1	0.062	679.050	0.303	
35	1	0.063	658.125	0.321	

(Continued on next page)

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

Item	Max Pts	*a*-par/alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3
36	1	0.042	607.075	0.200	
37	1	0.033	626.531	0.126	
38	1	0.040	615.142	0.200	
39	1	0.040	616.210	0.200	
40	1	0.067	651.185	0.361	
41	2	0.051	32.339	32.613	
42	2	0.062	38.699	39.542	
43	2	0.047	28.790	31.000	
44	1	0.032	635.061	0.094	
45	1	0.051	656.331	0.243	
46	1	0.059	658.332	0.268	
47	2	0.075	47.892	46.578	
48	2	0.071	45.264	44.709	
49	2	0.060	37.612	37.730	
50	2	0.051	32.720	33.628	
51	3	0.056	34.339	36.360	37.858

Table 25. 2011 Operational Item Parameter Estimates, Grade 4

Item	Max Pts	*a*-par/alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3	gamma4
1	1	0.025	636.219	0.200		
2	1	0.027	662.258	0.163		
3	1	0.065	640.839	0.213		
4	1	0.035	651.827	0.302		
5	1	0.041	651.227	0.156		
6	1	0.024	658.891	0.119		
7	1	0.037	660.286	0.242		
8	1	0.032	685.429	0.382		
9	1	0.025	692.461	0.206		
10	1	0.020	652.246	0.156		
11	1	0.036	686.792	0.303		
12	1	0.028	673.832	0.163		
13	1	0.034	668.277	0.168		
14	1	0.031	688.375	0.290		
15	1	0.032	625.521	0.079		
16	1	0.047	660.524	0.326		
17	1	0.034	618.680	0.087		
18	1	0.056	673.822	0.146		
19	1	0.029	683.699	0.202		
20	1	0.023	643.799	0.106		
21	1	0.048	629.684	0.195		

(Continued on next page)

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

Item	Max Pts	*a*-par/alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3	gamma4
22	1	0.040	645.202	0.127		
23	1	0.050	663.131	0.167		
24	1	0.046	636.123	0.321		
25	1	0.027	652.756	0.236		
26	1	0.020	641.024	0.096		
27	1	0.050	655.358	0.237		
28	1	0.027	625.119	0.088		
29	1	0.042	640.783	0.180		
30	1	0.054	651.495	0.253		
31	1	0.030	636.659	0.058		
32	1	0.014	666.135	0.137		
33	1	0.042	661.528	0.169		
34	1	0.034	649.226	0.128		
35	1	0.044	649.029	0.232		
36	1	0.045	675.866	0.196		
37	1	0.031	653.697	0.216		
38	1	0.036	667.378	0.184		
39	1	0.031	641.049	0.106		
40	1	0.048	660.975	0.378		
41	1	0.035	683.206	0.153		
42	1	0.054	711.315	0.203		
43	1	0.030	706.580	0.198		
44	1	0.011	628.662	0.200		
45	1	0.032	616.735	0.200		
46	1	0.016	636.276	0.200		
47	1	0.027	619.346	0.200		
48	1	0.015	662.351	0.124		
49	2	0.028	16.760	18.943		
50	2	0.042	26.305	28.114		
51	2	0.029	18.206	20.040		
52	1	0.023	688.695	0.417		
53	1	0.031	623.242	0.118		
54	1	0.023	693.037	0.384		
55	2	0.031	19.579	20.781		
56	2	0.025	16.773	17.343		
57	2	0.044	27.451	29.779		
58	2	0.046	30.082	30.785		
59	4	0.044	27.350	28.331	29.393	30.912

Table 26. 2011 Operational Item Parameter Estimates, Grade 5

Item	Max Pts	*a*-par/alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3	gamma4
1	1	0.045	625.105	0.117		
2	1	0.064	670.941	0.464		
3	1	0.056	659.818	0.144		
4	1	0.014	685.449	0.200		
5	1	0.052	659.479	0.182		
6	1	0.027	659.531	0.200		
7	1	0.054	655.686	0.155		
8	1	0.040	628.716	0.149		
9	1	0.064	639.757	0.232		
10	1	0.042	636.591	0.188		
11	1	0.034	639.882	0.121		
12	1	0.059	659.963	0.289		
13	1	0.078	657.886	0.184		
14	1	0.061	661.390	0.243		
15	1	0.088	655.595	0.332		
16	1	0.056	657.174	0.185		
17	1	0.026	707.815	0.220		
18	1	0.030	661.868	0.071		
19	1	0.059	657.516	0.166		
20	1	0.022	690.270	0.200		
21	1	0.037	645.206	0.200		
22	1	0.030	653.523	0.074		
23	1	0.036	672.715	0.258		
24	1	0.028	661.899	0.200		
25	1	0.055	660.760	0.221		
26	1	0.084	653.523	0.310		
27	1	0.082	647.749	0.209		
28	1	0.067	657.767	0.197		
29	1	0.093	651.639	0.306		
30	1	0.072	649.416	0.204		
31	1	0.080	666.648	0.230		
32	1	0.038	665.702	0.170		
33	1	0.041	678.580	0.199		
34	1	0.080	672.510	0.291		
35	1	0.071	670.441	0.189		
36	1	0.020	608.250	0.200		
37	1	0.029	636.260	0.200		
38	1	0.036	628.022	0.200		
39	1	0.030	638.721	0.200		
40	1	0.059	644.537	0.188		

(Continued on next page)

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

Item	Max Pts	*a*-par/alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3	gamma4
41	2	0.076	47.329	49.502		
42	2	0.065	40.729	43.153		
43	2	0.073	47.026	49.551		
44	1	0.041	652.800	0.228		
45	1	0.055	664.777	0.301		
46	1	0.029	629.763	0.190		
47	2	0.112	70.645	72.115		
48	2	0.067	41.917	43.826		
49	2	0.062	38.802	40.751		
50	2	0.053	33.232	35.093		
51	4	0.064	39.539	40.283	41.869	42.907

Table 27. 2011 Operational Item Parameter Estimates, Grade 6

Item	Max Pts	*a*-par/alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3	gamma4
1	1	0.015	652.749	0.200		
2	1	0.060	623.328	0.200		
3	1	0.068	624.343	0.200		
4	1	0.038	653.578	0.237		
5	1	0.089	648.259	0.215		
6	1	0.048	623.671	0.200		
7	1	0.044	668.827	0.191		
8	1	0.071	658.508	0.250		
9	1	0.046	656.860	0.095		
10	1	0.045	642.013	0.049		
11	1	0.035	666.658	0.191		
12	1	0.073	642.953	0.175		
13	1	0.070	641.379	0.152		
14	1	0.080	646.264	0.206		
15	1	0.041	665.334	0.189		
16	1	0.063	665.594	0.295		
17	1	0.060	638.680	0.152		
18	1	0.045	639.633	0.050		
19	1	0.060	663.033	0.246		
20	1	0.079	647.470	0.231		
21	1	0.093	650.855	0.175		
22	1	0.080	649.876	0.150		
23	1	0.065	642.940	0.180		
24	1	0.056	649.431	0.147		
25	1	0.047	662.680	0.164		

(Continued on next page)

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

Item	Max Pts	*a*-par/alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3	gamma4
26	1	0.067	639.422	0.276		
27	1	0.031	659.688	0.085		
28	1	0.037	666.140	0.083		
29	1	0.067	667.995	0.258		
30	1	0.050	641.780	0.200		
31	1	0.044	649.644	0.027		
32	1	0.098	663.771	0.362		
33	1	0.098	663.609	0.322		
34	1	0.026	690.989	0.422		
35	1	0.046	636.232	0.062		
36	1	0.069	653.585	0.183		
37	1	0.070	649.854	0.239		
38	1	0.044	638.713	0.053		
39	1	0.073	666.088	0.247		
40	1	0.047	646.181	0.095		
41	1	0.035	671.643	0.152		
42	1	0.042	627.686	0.200		
43	1	0.065	625.607	0.200		
44	1	0.056	653.723	0.287		
45	1	0.029	673.479	0.274		
46	1	0.049	651.959	0.138		
47	2	0.072	45.428	46.462		
48	2	0.071	44.790	45.359		
49	2	0.068	42.384	44.521		
50	1	0.065	646.675	0.210		
51	1	0.045	665.849	0.234		
52	1	0.057	637.685	0.195		
53	2	0.039	23.192	25.087		
54	2	0.052	32.305	33.888		
55	2	0.066	42.448	43.464		
56	2	0.064	41.477	42.755		
57	4	0.068	42.701	43.769	44.770	45.657

Table 28. 2011 Operational Item Parameter Estimates, Grade 7

Item	Max Pts	*a*-par/alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3	gamma4
1	1	0.073	651.035	0.278		
2	1	0.027	655.030	0.226		
3	1	0.039	652.399	0.148		
4	1	0.088	647.217	0.229		
5	1	0.059	663.967	0.167		
6	1	0.068	647.671	0.327		
7	1	0.051	642.901	0.204		
8	1	0.070	640.781	0.196		
9	1	0.079	647.391	0.202		
10	1	0.025	646.844	0.119		
11	1	0.051	670.468	0.220		
12	1	0.037	681.678	0.263		
13	1	0.052	660.910	0.155		
14	1	0.044	673.679	0.254		
15	1	0.067	633.978	0.217		
16	1	0.093	647.765	0.243		
17	1	0.061	641.864	0.196		
18	1	0.050	644.249	0.140		
19	1	0.073	652.599	0.163		
20	1	0.069	644.106	0.204		
21	1	0.091	646.929	0.199		
22	1	0.066	635.941	0.180		
23	1	0.075	664.919	0.225		
24	1	0.071	667.395	0.223		
25	1	0.039	667.664	0.158		
26	1	0.079	668.733	0.225		
27	1	0.060	710.414	0.361		
28	1	0.035	687.347	0.114		
29	1	0.035	654.792	0.148		
30	1	0.056	647.295	0.176		
31	1	0.043	653.629	0.256		
32	1	0.040	648.378	0.082		
33	1	0.104	640.681	0.290		
34	1	0.025	661.037	0.200		
35	1	0.064	654.674	0.163		
36	1	0.058	658.730	0.125		
37	1	0.075	637.871	0.253		
38	1	0.078	649.085	0.283		
39	1	0.051	661.959	0.256		
40	1	0.087	653.648	0.292		
41	1	0.043	661.877	0.184		

(Continued on next page)

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

Item	Max Pts	*a*-par/alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3	gamma4
42	1	0.025	635.028	0.200		
43	1	0.037	647.948	0.251		
44	1	0.034	672.687	0.190		
45	1	0.046	642.188	0.136		
46	1	0.021	643.765	0.200		
47	2	0.082	50.206	52.997		
48	2	0.063	39.285	40.458		
49	2	0.067	41.867	43.960		
50	1	0.043	631.597	0.200		
51	1	0.047	668.409	0.298		
52	1	0.045	659.132	0.369		
53	2	0.045	27.484	29.173		
54	2	0.051	33.579	33.319		
55	2	0.073	45.386	46.763		
56	2	0.055	34.404	36.338		
57	4	0.064	39.832	40.683	42.355	43.659

Table 29. 2011 Operational Item Parameter Estimates, Grade 8

Item	Max Pts	*a*-par/alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3	gamma4
1	1	0.026	644.803	0.200		
2	1	0.027	644.815	0.220		
3	1	0.053	647.885	0.293		
4	1	0.055	632.375	0.297		
5	1	0.053	648.248	0.133		
6	1	0.061	638.716	0.246		
7	1	0.019	648.685	0.092		
8	1	0.050	634.362	0.244		
9	1	0.056	637.487	0.209		
10	1	0.057	631.929	0.184		
11	1	0.046	633.821	0.149		
12	1	0.047	629.109	0.207		
13	1	0.038	636.431	0.220		
14	1	0.042	652.442	0.126		
15	1	0.059	658.229	0.313		
16	1	0.025	650.863	0.200		
17	1	0.058	620.574	0.082		
18	1	0.040	633.618	0.166		
19	1	0.034	625.550	0.200		
20	1	0.043	631.634	0.076		

(Continued on next page)

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

Item	Max Pts	*a*-par/alpha	*b*-par/ gamma1	*c*-par/ gamma2	gamma3	gamma4
21	1	0.042	619.496	0.100		
22	1	0.054	616.426	0.072		
23	1	0.051	627.696	0.194		
24	1	0.041	615.623	0.200		
25	1	0.025	637.817	0.085		
26	1	0.047	640.617	0.154		
27	1	0.056	633.048	0.160		
28	1	0.036	670.910	0.245		
29	1	0.061	629.183	0.234		
30	1	0.053	653.351	0.331		
31	1	0.048	646.519	0.105		
32	1	0.050	671.108	0.151		
33	1	0.034	698.323	0.171		
34	1	0.071	640.036	0.313		
35	1	0.033	649.473	0.066		
36	1	0.067	625.456	0.126		
37	1	0.039	646.720	0.188		
38	1	0.060	646.964	0.096		
39	1	0.042	629.846	0.220		
40	1	0.042	657.174	0.195		
41	1	0.059	645.791	0.160		
42	1	0.056	612.146	0.200		
43	1	0.039	653.062	0.126		
44	1	0.063	650.616	0.147		
45	1	0.033	593.541	0.200		
46	1	0.046	621.157	0.109		
47	2	0.067	40.017	43.109		
48	2	0.038	23.325	24.600		
49	2	0.045	27.890	29.366		
50	1	0.040	663.918	0.105		
51	1	0.023	680.670	0.133		
52	1	0.049	658.714	0.182		
53	2	0.057	34.518	36.951		
54	2	0.064	38.914	40.571		
55	2	0.051	31.194	32.573		
56	2	0.047	30.266	31.294		
57	4	0.051	30.903	31.343	32.550	33.429

Test Characteristic Curves

Test characteristic curves (TCCs) provide an overview of the tests in the IRT scale score metric. The 2010 and 2011 TCCs were generated using final OP item parameters for all test items administered in 2010 and 2011. TCCs are the summation of all the item characteristic

curves (ICCs) for items that contribute to the OP scale score. Standard error (SE) curves graphically show the amount of measurement error at different ability levels. The 2010 and 2011 TCCs and SE curves are presented in Figures 1–6. Following the adoption of the chain equating method by New York State, the TCCs for new OP test forms are compared to the previous year's TCCs rather than to the baseline 2006 test form TCCs. It should be noted that although the 2010 OP curves are considered to be target curves for the 2011 OP test TCCs, NYSED requested that the 2011 forms be more difficult than the 2010 forms, which was taken into consideration during new from selection. Note that in all figures the pink TCCs and SE curves represent the 2010 OP test and blue TCCs and SE curves represent 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 -axis is the proportion of the test that the students can answer correctly.

Figure 1. Grade 3 ELA 2010 and 2011 OP TCCs and SE curves

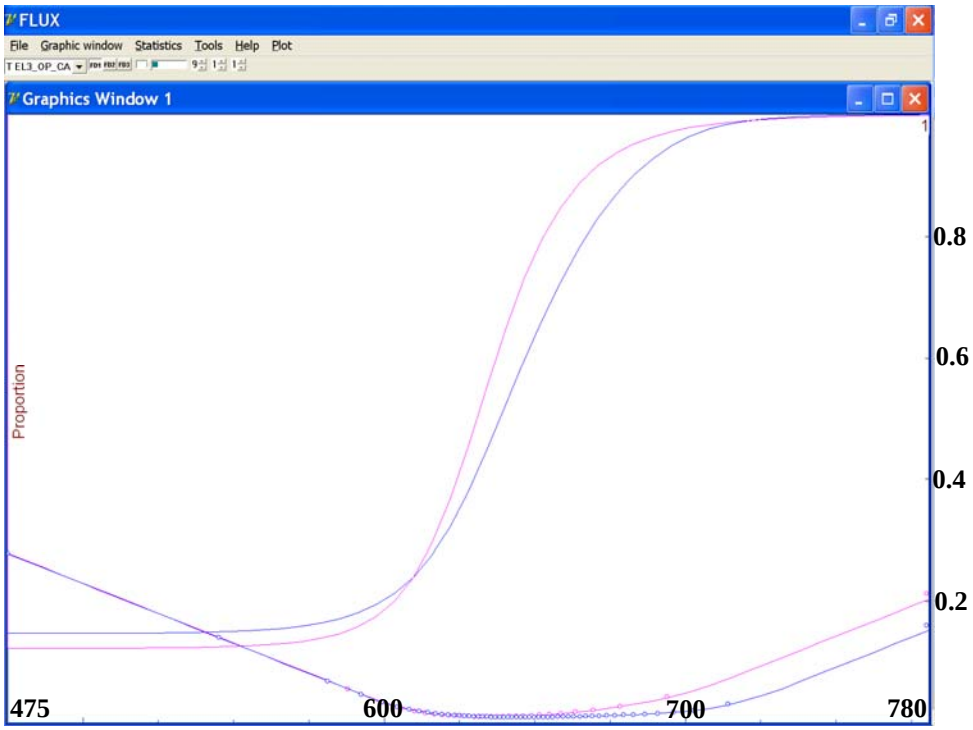


Figure 2. Grade 4 ELA 2010 and 2011 OP TCCs and SE curves

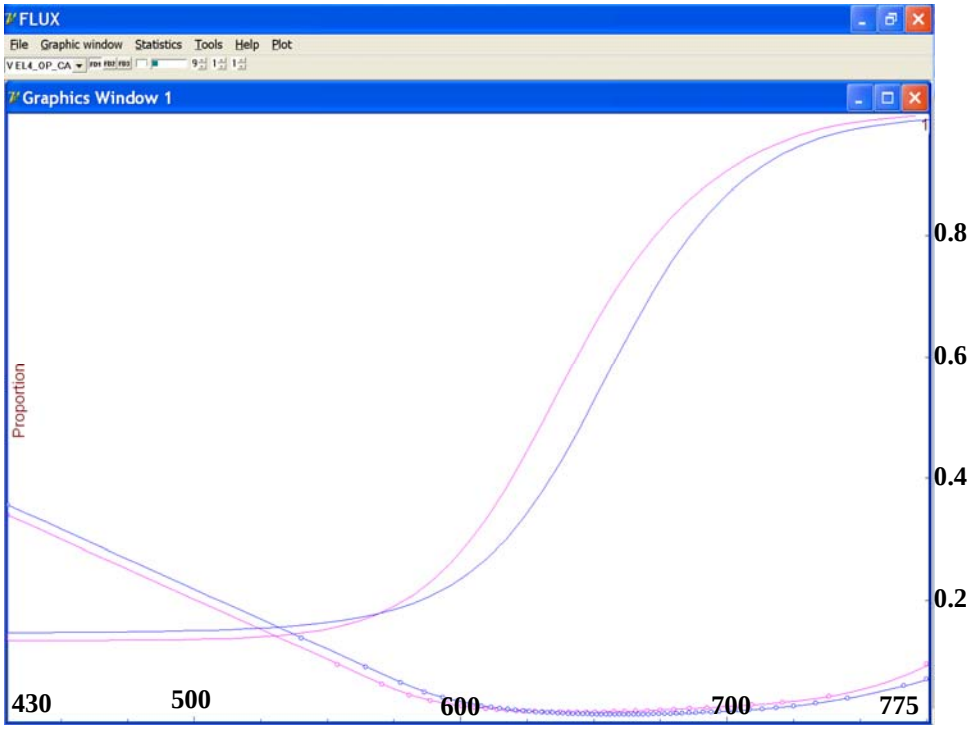


Figure 3. Grade 5 ELA 2010 and 2011 OP TCCs and SE curves

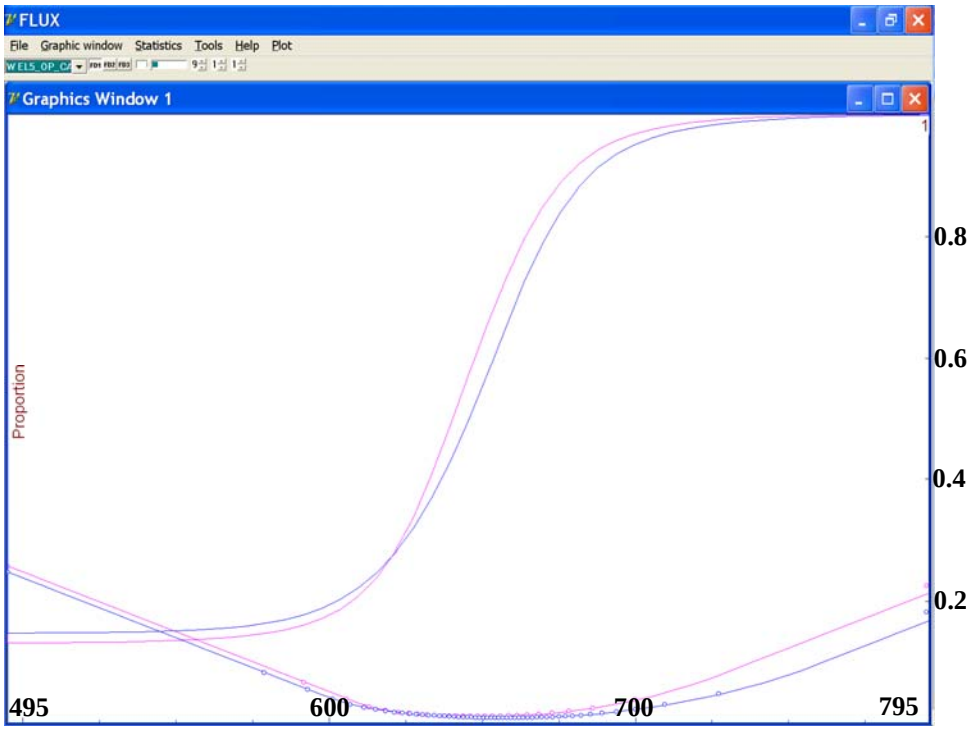


Figure 4. Grade 6 ELA 2010 and 2011 TCCs and SE curves

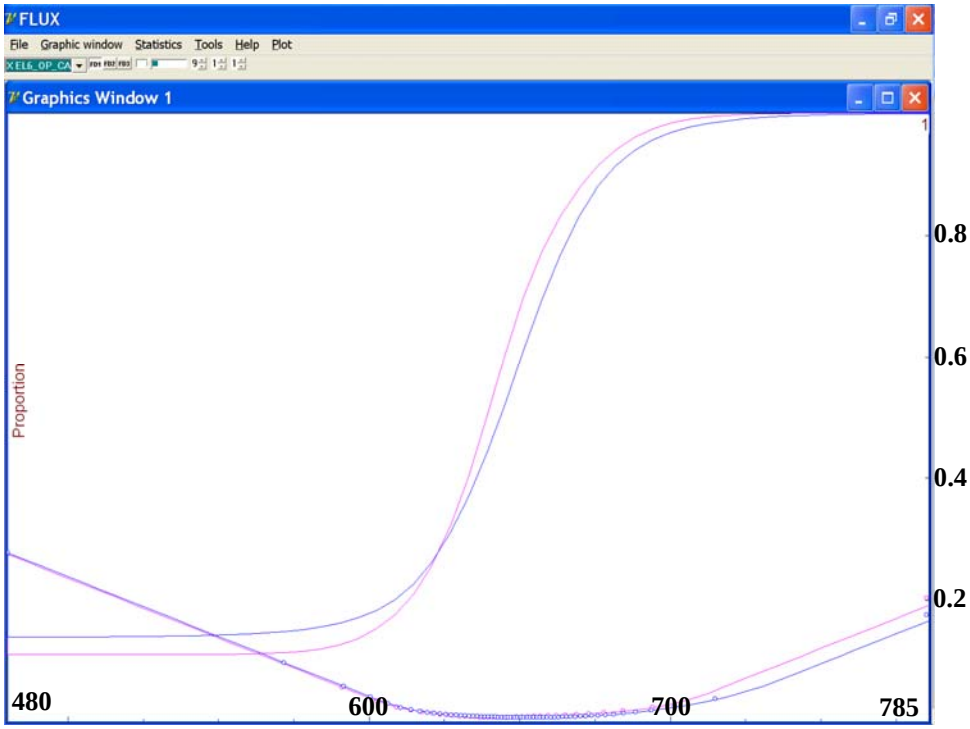


Figure 5. Grade 7 ELA 2010 and 2011 TCCs and SE curves

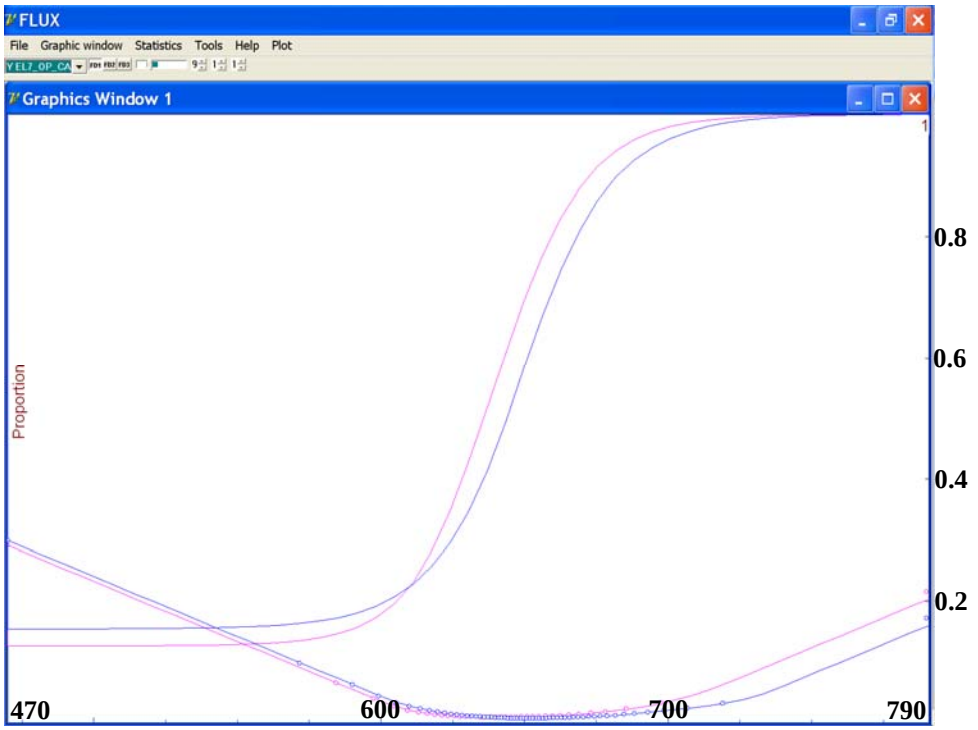
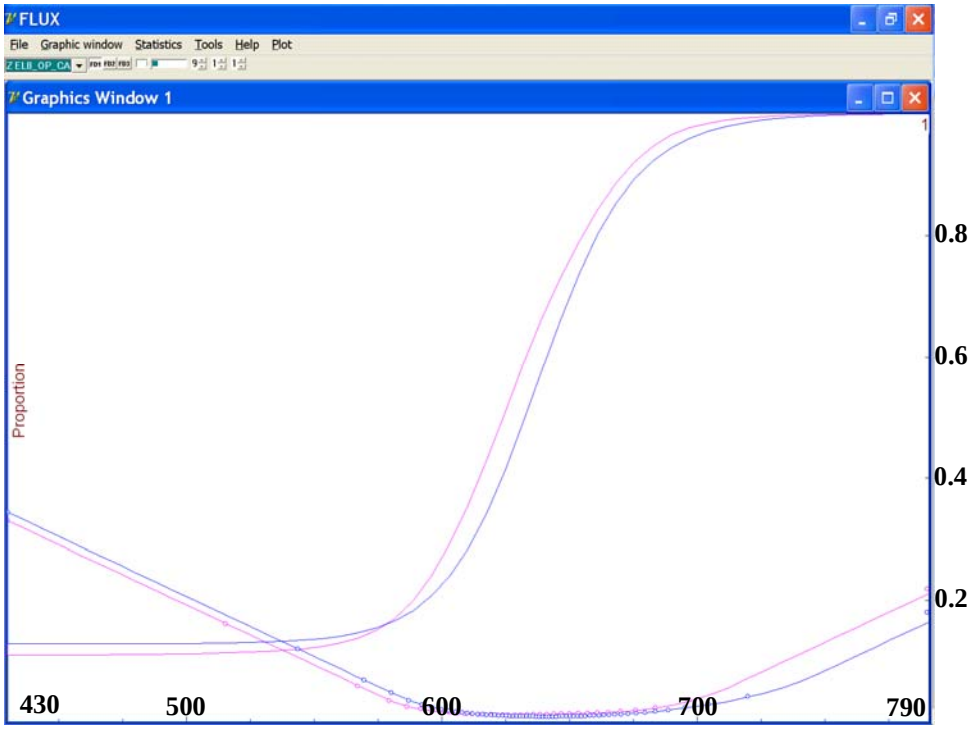


Figure 6. Grade 8 ELA 2010 and 2011 TCCs and SE curves



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

Scoring Procedure

New York State students were scored using the number correct (NC) scoring method. This method considers how many score points a student obtained on a test in determining his or her 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 CTB/McGraw-Hill's proprietary FLUX program. The inverse of the TCC procedure produces trait values based on unweighted raw scores. These estimates show negligible bias for tests with maximum possible raw scores of at least 30 points. All New York State 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 correct maximum likelihood estimates (MLE) showing negligible bias for tests of at least 30 items. For tests with a mixture of MC and CR items, the MLE and TCC estimates are even more similar (Yen, 1984).

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

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

where

x_i is a student's observed raw score on item *i*,

v_i is a non-optimal weight specified in a scoring process (v_i = 1 if no weights are specified), and

$\tilde{\theta}$ is a trait estimate.

Raw Score-to-Scale Score and SEM Conversion Tables

The scale score (SS) is the basic score for the New York State 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:

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

where

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

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

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

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

Weighted Raw Score	Scale Score	Standard Error
0	475	141
1	475	141
2	475	141
3	475	141
4	475	141
5	475	141
6	475	141
7	475	141
8	475	141
9	545	71
10	581	35
11	592	24
12	599	17
13	604	14
14	608	12
15	612	10
16	614	9
17	617	8
18	619	7
19	621	7
20	623	7
21	625	6
22	627	6
23	628	6
24	630	6
25	631	5
26	633	5
27	634	5
28	636	5
29	637	5
30	638	5
31	640	5
32	641	5
33	643	5
34	644	5
35	645	5
36	647	5
37	648	5
38	650	5
39	651	5
40	653	5
41	654	5

(Continued on next page)

.

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

Weighted Raw Score	Scale Score	Standard Error
42	656	5
43	658	5
44	659	5
45	661	5
46	663	6
47	665	6
48	667	6
49	669	6
50	671	6
51	674	6
52	677	7
53	680	7
54	683	7
55	687	8
56	691	8
57	696	9
58	703	11
59	714	16
60	780	82

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

Weighted Raw Score	Scale Score	Standard Error
0	430	180
1	430	180
2	430	180
3	430	180
4	430	180
5	430	180
6	430	180
7	430	180
8	430	180
9	430	180
10	430	180
11	540	70
12	564	46
13	578	33
14	586	25
15	593	20
16	599	17
17	604	15
18	608	14

(Continued on next page)

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

Weighted Raw Score	Scale Score	Standard Error
19	612	13
20	615	12
21	618	11
22	621	10
23	623	10
24	626	9
25	628	9
26	630	8
27	633	8
28	635	8
29	637	8
30	639	7
31	640	7
32	642	7
33	644	7
34	646	7
35	648	7
36	649	7
37	651	7
38	653	7
39	655	7
40	656	7
41	658	7
42	660	7
43	661	7
44	663	7
45	665	7
46	667	7
47	669	7
48	671	7
49	673	7
50	675	7
51	677	7
52	679	7
53	681	7
54	683	8
55	686	8
56	688	8
57	691	8
58	694	9
59	697	9
60	701	9

(Continued on next page)

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

Weighted Raw Score	Scale Score	Standard Error
61	704	10
62	709	10
63	713	11
64	719	12
65	725	13
66	733	16
67	745	20
68	766	31
69	775	36

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

Weighted Raw Score	Scale Score	Standard Error
0	495	126
1	495	126
2	495	126
3	495	126
4	495	126
5	495	126
6	495	126
7	495	126
8	495	126
9	495	126
10	579	42
11	593	28
12	601	20
13	607	16
14	611	13
15	615	11
16	618	10
17	621	9
18	624	8
19	626	8
20	629	7
21	631	7
22	632	7
23	634	6
24	636	6
25	638	6
26	639	6
27	641	5
28	642	5

(Continued on next page)

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

Weighted Raw Score	Scale Score	Standard Error
29	644	5
30	645	5
31	646	5
32	648	5
33	649	5
34	650	5
35	652	5
36	653	4
37	654	4
38	655	4
39	657	4
40	658	4
41	659	4
42	661	4
43	662	4
44	663	5
45	665	5
46	666	5
47	668	5
48	669	5
49	671	5
50	673	5
51	675	5
52	677	6
53	680	6
54	682	7
55	685	7
56	689	8
57	694	10
58	700	12
59	710	15
60	727	24
61	795	92

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

Weighted Raw Score	Scale Score	Standard Error
0	480	141
1	480	141
2	480	141
3	480	141
4	480	141
5	480	141
6	480	141
7	480	141
8	480	141
9	480	141
10	572	50
11	591	30
12	600	21
13	606	15
14	611	12
15	614	10
16	617	9
17	619	8
18	622	7
19	624	7
20	626	6
21	627	6
22	629	6
23	630	6
24	632	5
25	633	5
26	635	5
27	636	5
28	637	5
29	638	5
30	640	4
31	641	4
32	642	4
33	643	4
34	644	4
35	645	4
36	646	4
37	647	4
38	648	4
39	649	4
40	650	4

(Continued on next page)

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

Weighted Raw Score	Scale Score	Standard Error
41	651	4
42	652	4
43	653	4
44	655	4
45	656	4
46	657	4
47	658	4
48	659	4
49	660	4
50	662	4
51	663	4
52	664	4
53	666	4
54	667	5
55	668	5
56	670	5
57	672	5
58	674	5
59	676	6
60	678	6
61	681	7
62	684	7
63	689	8
64	694	10
65	701	13
66	715	19
67	785	89

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

Weighted Raw Score	Scale Score	Standard Error
0	470	152
1	470	152
2	470	152
3	470	152
4	470	152
5	470	152
6	470	152
7	470	152
8	470	152

(Continued on next page)

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

Weighted Raw Score	Scale Score	Standard Error
9	470	152
10	470	152
11	572	50
12	590	32
13	599	23
14	605	17
15	610	14
16	614	12
17	617	10
18	620	9
19	622	8
20	624	8
21	626	7
22	628	7
23	630	6
24	632	6
25	633	6
26	635	5
27	636	5
28	637	5
29	639	5
30	640	5
31	641	4
32	642	4
33	643	4
34	644	4
35	645	4
36	646	4
37	648	4
38	649	4
39	650	4
40	651	4
41	652	4
42	653	4
43	654	4
44	655	4
45	656	4
46	658	4
47	659	4
48	660	4
49	661	4
50	663	5

(Continued on next page)

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

51	664	5
52	666	5
53	667	5
54	669	5
55	670	5
56	672	5
57	674	6
58	676	6
59	679	6
60	682	7
61	685	7
62	688	8
63	693	9
64	699	11
65	707	13
66	719	17
67	790	87

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

Weighted Raw Score	Scale Score	Standard Error
0	430	175
1	430	175
2	430	175
3	430	175
4	430	175
5	430	175
6	430	175
7	430	175
8	430	175
9	544	61
10	569	35
11	580	24
12	587	18
13	592	14
14	596	12
15	600	11
16	603	10
17	606	9
18	608	8
19	610	8
20	612	7
21	614	7
22	616	6
23	618	6

(Continued on next page)

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

Weighted Raw Score	Scale Score	Standard Error
24	619	6
25	621	6
26	622	6
27	624	5
28	625	5
29	627	5
30	628	5
31	629	5
32	631	5
33	632	5
34	633	5
35	635	5
36	636	5
37	637	5
38	638	5
39	640	5
40	641	5
41	642	5
42	643	5
43	645	5
44	646	5
45	647	5
46	649	5
47	650	5
48	652	5
49	653	5
50	655	5
51	656	5
52	658	5
53	660	5
54	662	6
55	663	6
56	666	6
57	668	6
58	670	7
59	673	7
60	676	8
61	680	8
62	684	9
63	689	10
64	695	12
65	704	15
66	720	21
67	790	91

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, CTB/McGraw-Hill's scoring system looks not only at how many of those items the student answered correctly, but at additional information as well. In technical terms, the procedure CTB/McGraw-Hill uses to calculate the SPI is based on a combination of item response theory (IRT) and Bayesian methodology. In non-technical terms, the procedure takes into consideration the number of items related to the objective that the student answered correctly, the difficulty level of those items, as well as the student's performance on the rest of the test in which the objective is found. This use of additional information increases the accuracy of the SPI. Details on the SPI derivation procedure are provided in Appendix F.

For the 2011 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 36. SPI Target Ranges

Grade	Objective	# Items	Total Points	Level III Cut SPI Target Range
3	1	15	17	80–90
	2	24	28	69–80
	3	9	12	64–75
4	1	20	21	65–77
	2	26	30	67–75
	3	10	15	59–71
5	1	23	26	74–83
	2	17	19	69–81
	3	8	13	71–81
6	1	19	23	67–77
	2	24	26	73–81
	3	11	15	70–80
7	1	24	27	79–88
	2	20	22	63–74
	3	10	15	72–82

(Continued on next page)

Table 36. SPI Target Ranges (cont.)

Grade	Objective	# Items	Total Points	Level III Cut SPI Target Range
8	1	25	28	75–84
	2	19	21	77–85
	3	10	15	69–79

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 NYS test design does not support longitudinal comparison of the SPI scores due to such factors as differences in numbers of items per learning objective from year to year, differences in item difficulties in a given learning objective from year to year, and the fact that the learning objective sub-scores are not equated. The SPI scores are diagnostic scores and are best used at the classroom level to give teachers some insight into their students' strengths and weaknesses.

IRT DIF Statistics

In addition to classical DIF analysis, an IRT-based Linn-Harnisch statistical procedure was used to detect DIF on the Grades 3–8 ELA Tests (Linn and Harnisch, 1981). In this procedure, item parameters (discrimination, location, and guessing) and the scale score (θ) for each examinee were estimated for the 3PL model or the 2PPC model in the case of CR items. The item parameters were based on data from the total population of examinees. Then the population was divided into NRC, gender, or ethnic groups, and the members in each group are sorted into ten equal score categories (deciles) based upon their location on the scale score (θ) scale. The expected proportion correct for each group, based on the model prediction, is compared to the observed (actual) proportion correct obtained by the group.

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

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

where

n_g is the number of examinees in decile g.

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

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

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

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

where

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

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

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

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

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

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

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

These indices are indicators of the degree to which members of a specific subgroup perform better or worse than expected on each item. Differences for decile groups provide an index for each of the ten regions on the scale score (θ) scale. The decile group difference (D_{ig}) can

be either positive or negative. When the difference (D_{ig}) is greater than or equal to 0.100, or less than or equal to -0.100, the item is flagged for potential DIF.

The following groups were analyzed using the IRT-based DIF analysis: Female, Male, Asian, Black, Hispanic, White, High Needs districts (by NRC code), Low Needs districts (by NRC code), and English language learners. Applying the Linn-Harnisch method revealed that one item was flagged in Grade 3; five items were flagged in Grades 4 and 6; three items were flagged in Grades 5 and 7; four items were flagged on the Grade 8 test, as is shown in Table 37. As indicated in the classical DIF analysis section, items flagged for DIF do not necessarily display bias.

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

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

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

Section VII: Reliability and Standard Error of Measurement

This section presents specific information on various test reliability statistics (RS) and standard error of measurement (SEM), as well as the results from a study of performance level classification accuracy and consistency. The data set for these studies includes all tested New York State public and charter school students who received valid scores. A study of inter-rater reliability was conducted by a vendor other than CTB/McGraw-Hill and is not included in this *Technical Report*.

Test Reliability

Test reliability is directly related to score stability and standard error and, as such, is an essential element of fairness and validity. Test reliability can be directly measured with an alpha statistic, or the alpha statistic can be used to derive the SEM. For the Grades 3–8 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 38 are the case counts (N-count), number of test items (# Items), Cronbach's alpha and associated SEM, and Feldt-Raju coefficient and associated SEM obtained for the total ELA Tests.

Table 38. 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	196476	51	60	0.91	2.97	0.91	2.90
4	197040	59	69	0.91	3.47	0.92	3.36
5	200195	51	61	0.90	3.05	0.91	2.97
6	198076	57	67	0.92	3.29	0.92	3.19
7	200140	57	67	0.92	3.26	0.92	3.18
8	201278	57	67	0.92	3.25	0.93	3.16

All the coefficients for total test reliability were in the range 0.90–0.93, which indicates high internal consistency. As expected, the lowest reliabilities were found for the shortest test (i.e., Grade 5), and the highest reliabilities were associated with the longer tests (Grades 4, 6, 7, and 8).

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 39 presents reliabilities for the MC subsets.

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

Grade	N-count	# Items	Cronbach's Alpha	SEM of Cronbach	Feldt-Raju	SEM of Feldt-Raju
3	196476	43	0.88	2.47	0.88	2.45
4	197040	51	0.89	2.83	0.90	2.82
5	200195	43	0.87	2.56	0.88	2.54
6	198076	49	0.90	2.71	0.90	2.69
7	200140	49	0.90	2.75	0.90	2.74
8	201278	49	0.91	2.67	0.91	2.65

Reliability of CR Items

Reliability coefficients were also computed for the subsets of CR items. The results are presented in Table 40.

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

Grade	N-count	# Items	# RS Points	Cronbach's Alpha	SEM of Cronbach	Feldt-Raju	SEM of Feldt-Raju
3	196476	8	17	0.78	1.52	0.79	1.50
4	197040	8	18	0.78	1.82	0.79	1.77
5	200195	8	18	0.80	1.52	0.81	1.47
6	198076	8	18	0.79	1.70	0.81	1.61
7	200140	8	18	0.79	1.58	0.81	1.53
8	201278	8	18	0.81	1.67	0.82	1.62

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 code (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 41A–41F, 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 0.86–0.94 range, indicating very good test internal consistency (reliability) for analyzed subgroups of examinees.

Table 41A. Grade 3 Test Reliability by Subgroup

Group	Subgroup	N-count	Cronbach's Alpha	SEM of Cronbach	Feldt-Raju	SEM of Feldt-Raju
State	All Students	196476	0.91	2.97	0.91	2.90
Gender	Female	96054	0.90	2.88	0.90	2.82
	Male	100422	0.91	3.04	0.92	2.96
Ethnicity	Asian	15960	0.91	2.75	0.91	2.68
	Black	36529	0.90	3.19	0.91	3.12
	Hispanic	45736	0.90	3.15	0.91	3.07
	American Indian	1090	0.91	3.09	0.91	3.02
	Multi-Racial	1563	0.91	2.91	0.91	2.84
	Unknown	272	0.91	2.84	0.91	2.77
	White	95326	0.90	2.80	0.90	2.74
NRC	New York City	70884	0.91	3.06	0.91	2.98
	Big 4 Cities	8117	0.91	3.32	0.92	3.23
	High Needs Urban/Suburban	15372	0.90	3.12	0.91	3.05
	High Needs Rural	11148	0.91	3.01	0.91	2.94
	Average Needs	57536	0.90	2.85	0.90	2.79
	Low Needs	27941	0.88	2.64	0.88	2.60
	Charter	4965	0.86	3.01	0.87	2.96
SWD	All Codes	28048	0.91	3.48	0.92	3.35
SUA	All Codes	48718	0.91	3.38	0.92	3.26
ELL	ELL=Y	18144	0.90	3.40	0.91	3.29
SWD/SUA	SUA=504 plan codes	24828	0.91	3.51	0.92	3.37
ELL/SUA	SUA=ELL codes	16502	0.90	3.39	0.90	3.28

Table 41B. Grade 4 Test Reliability by Subgroup

Group	Subgroup	N-count	Cronbach's Alpha	SEM of Cronbach	Feldt-Raju	SEM of Feldt-Raju
State	All Students	197040	0.91	3.47	0.92	3.36
Gender	Female	96348	0.91	3.41	0.91	3.32
	Male	100692	0.92	3.51	0.92	3.40
Ethnicity	Asian	15492	0.91	3.28	0.92	3.18
	Black	36838	0.91	3.64	0.91	3.54
	Hispanic	44920	0.91	3.61	0.91	3.51
	American Indian	955	0.91	3.59	0.91	3.48
	Multi-Racial	1377	0.92	3.40	0.92	3.30
	Unknown	248	0.92	3.40	0.92	3.30
	White	97210	0.91	3.33	0.91	3.24
NRC	New York City	70021	0.91	3.54	0.92	3.44
	Big 4 Cities	8254	0.91	3.72	0.92	3.62
	High Needs Urban/Suburban	15319	0.91	3.59	0.91	3.49
	High Needs Rural	11552	0.91	3.51	0.91	3.41
	Average Needs	58973	0.90	3.38	0.91	3.29
	Low Needs	28461	0.89	3.17	0.90	3.09
	Charter	3899	0.88	3.49	0.89	3.43
SWD	All Codes	29861	0.91	3.79	0.91	3.68
SUA	All Codes	49414	0.91	3.76	0.92	3.64
ELL	ELL=Y	16287	0.90	3.80	0.90	3.69
SWD/SUA	SUA=504 plan codes	27285	0.91	3.79	0.91	3.69
ELL/SUA	SUA=ELL codes	14785	0.90	3.79	0.90	3.69

Table 41C. Grade 5 Test Reliability by Subgroup

Group	Subgroup	N-count	Cronbach's Alpha	SEM of Cronbach	Feldt-Raju	SEM of Feldt-Raju
State	All Students	200195	0.90	3.05	0.91	2.97
Gender	Female	97835	0.90	2.99	0.90	2.91
	Male	102360	0.91	3.10	0.91	3.02
Ethnicity	Asian	16633	0.91	2.82	0.92	2.74
	Black	37296	0.90	3.27	0.90	3.19
	Hispanic	44110	0.90	3.23	0.90	3.15
	American Indian	969	0.90	3.22	0.91	3.14
	Multi-Racial	1315	0.91	2.99	0.91	2.91
	Unknown	240	0.90	3.01	0.91	2.93
	White	99632	0.89	2.91	0.90	2.83

(Continued on next page)

Table 41C. Grade 5 Test Reliability by Subgroup (cont.)

Group	Subgroup	N-count	Cronbach's Alpha	SEM of Cronbach	Feldt-Raju	SEM of Feldt-Raju
NRC	New York City	69263	0.91	3.13	0.91	3.04
	Big 4 Cities	7906	0.91	3.38	0.91	3.29
	High Needs Urban/Suburban	14996	0.90	3.21	0.91	3.13
	High Needs Rural	11529	0.90	3.14	0.91	3.06
	Average Needs	60624	0.89	2.97	0.90	2.90
	Low Needs	30062	0.87	2.72	0.87	2.67
	Charter	5207	0.88	3.19	0.89	3.13
SWD	All Codes	30698	0.90	3.50	0.91	3.39
SUA	All Codes	48998	0.90	3.45	0.91	3.34
ELL	ELL=Y	13890	0.89	3.53	0.90	3.42
SWD/SUA	SUA=504 plan codes	28298	0.90	3.51	0.90	3.41
ELL/SUA	SUA=ELL codes	12423	0.89	3.52	0.89	3.42

Table 41D. Grade 6 Test Reliability by Subgroup

Group	Subgroup	N-count	Cronbach's Alpha	SEM of Cronbach	Feldt-Raju	SEM of Feldt-Raju
State	All Students	198076	0.92	3.29	0.92	3.19
Gender	Female	96452	0.91	3.23	0.92	3.13
	Male	101624	0.92	3.34	0.92	3.24
Ethnicity	Asian	15218	0.93	3.10	0.93	3.00
	Black	37645	0.90	3.50	0.91	3.41
	Hispanic	43253	0.91	3.50	0.92	3.40
	American Indian	929	0.91	3.44	0.91	3.36
	Multi-Racial	1253	0.91	3.21	0.92	3.11
	Unknown	253	0.93	3.22	0.93	3.14
	White	99525	0.90	3.11	0.91	3.01
NRC	New York City	67620	0.92	3.44	0.92	3.33
	Big 4 Cities	7746	0.91	3.55	0.92	3.46
	High Needs Urban/Suburban	14604	0.91	3.42	0.91	3.33
	High Needs Rural	11576	0.91	3.33	0.91	3.23
	Average Needs	60990	0.90	3.16	0.91	3.07
	Low Needs	30045	0.89	2.91	0.89	2.84
	Charter	4853	0.88	3.38	0.89	3.32

(Continued on next page)

Table 41D. Grade 6 Test Reliability by Subgroup (cont.)

Group	Subgroup	N-count	Cronbach's Alpha	SEM of Cronbach	Feldt-Raju	SEM of Feldt-Raju
SWD	All Codes	30338	0.90	3.70	0.90	3.62
SUA	All Codes	44473	0.90	3.69	0.91	3.60
ELL	ELL=Y	11975	0.87	3.79	0.88	3.69
SWD/SUA	SUA=504 plan codes	27707	0.90	3.71	0.90	3.63
ELL/SUA	SUA=ELL codes	10139	0.87	3.79	0.88	3.69

Table 41E. Grade 7 Test Reliability by Subgroup

Group	Subgroup	N-count	Cronbach's Alpha	SEM of Cronbach	Feldt-Raju	SEM of Feldt-Raju
State	All Students	200140	0.92	3.26	0.92	3.18
Gender	Female	98063	0.91	3.16	0.92	3.09
	Male	102077	0.92	3.33	0.92	3.26
Ethnicity	Asian	15315	0.93	3.05	0.93	2.96
	Black	37951	0.90	3.49	0.91	3.43
	Hispanic	42976	0.91	3.48	0.91	3.41
	American Indian	982	0.91	3.40	0.92	3.33
	Multi-Racial	1111	0.92	3.17	0.92	3.09
	Unknown	264	0.93	3.31	0.93	3.22
	White	101541	0.90	3.07	0.91	3.00
NRC	New York City	68583	0.92	3.40	0.92	3.33
	Big 4 Cities	7527	0.92	3.57	0.92	3.50
	High Needs Urban/Suburban	14513	0.91	3.42	0.91	3.36
	High Needs Rural	11698	0.91	3.28	0.91	3.21
	Average Needs	61149	0.90	3.13	0.91	3.06
	Low Needs	32076	0.89	2.88	0.89	2.83
	Charter	3725	0.88	3.36	0.88	3.33
SWD	All Codes	30706	0.90	3.70	0.90	3.63
SUA	All Codes	43064	0.90	3.69	0.91	3.61
ELL	ELL = Y	10880	0.87	3.82	0.88	3.72
SWD/SUA	SUA=504 plan codes	27892	0.89	3.71	0.90	3.64
ELL/SUA	SUA=ELL codes	9080	0.87	3.82	0.88	3.72

Table 41F. Grade 8 Test Reliability by Subgroup

Group	Subgroup	N-count	Cronbach's Alpha	SEM of Cronbach	Feldt-Raju	SEM of Feldt-Raju
State	All Students	201278	0.92	3.25	0.93	3.16
Gender	Female	98306	0.92	3.15	0.92	3.06
	Male	102972	0.92	3.34	0.93	3.25
Ethnicity	Asian	15846	0.94	3.06	0.94	2.97
	Black	38001	0.91	3.46	0.91	3.38
	Hispanic	42651	0.92	3.45	0.92	3.37
	American Indian	979	0.92	3.44	0.92	3.35
	Multi-Racial	981	0.91	3.14	0.92	3.06
	Unknown	240	0.93	3.28	0.93	3.19
	White	102580	0.91	3.08	0.91	3.00
NRC	New York City	69837	0.92	3.39	0.92	3.31
	Big 4 Cities	7503	0.92	3.54	0.93	3.46
	High Needs Urban/Suburban	14233	0.92	3.38	0.92	3.31
NRC	High Needs Rural	11675	0.91	3.28	0.92	3.19
	Average Needs	61628	0.91	3.13	0.91	3.05
	Low Needs	32415	0.89	2.87	0.90	2.80
	Charter	2856	0.89	3.36	0.89	3.31
SWD	All Codes	30004	0.90	3.65	0.91	3.57
SUA	All Codes	42130	0.91	3.65	0.91	3.56
ELL	ELL = Y	10884	0.88	3.75	0.88	3.65
SWD/SUA	SUA=504 plan codes	27330	0.90	3.66	0.91	3.58
ELL/SUA	SUA=ELL codes	9171	0.88	3.75	0.88	3.65

Standard Error of Measurement

The SEM, as computed from Cronbach's alpha and the Feldt-Raju reliability statistics, are presented in Table 38. The SEMs ranged 2.97–3.47, 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 41A–41F. The SEMs associated with all reliability estimates for all subpopulations are in the range 2.60–3.82, 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 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 42 and 43 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 42 depicts the consistency study results based on the range of performance levels for all grades. Overall, between 75% and 81% of students were estimated to be classified

consistently to one of the four performance categories. The coefficient kappa, which indicates the consistency of the placement in the absence of chance, ranged 0.61–0.68.

Table 42. Decision Consistency (All Cuts)

Grade	N-count	Agreement	Inconsistency	Kappa
3	196476	0.75	0.25	0.61
4	197040	0.81	0.19	0.67
5	200195	0.76	0.24	0.62
6	198076	0.78	0.22	0.65
7	200140	0.78	0.22	0.66
8	201278	0.81	0.19	0.68

Table 43 depicts the consistency study results based on two performance levels (passing and not passing) as defined by the Level III cut. Overall, about 86%–88% 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 0.73–0.76.

Table 43. Decision Consistency (Level III Cut)

Grade	N-count	Agreement	Inconsistency	Kappa
3	196476	0.86	0.14	0.73
4	197040	0.88	0.12	0.75
5	200195	0.87	0.13	0.74
6	198076	0.88	0.12	0.76
7	200140	0.88	0.12	0.75
8	201278	0.88	0.12	0.76

Accuracy

The results of classification accuracy are presented in Table 44. 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 approximately 82%–86% of the time across all performance levels and approximately 90%–91% of the time in regards to the Level III cut score.

Table 44. Decision Agreement (Accuracy)

Grade	N-count	Accuracy					
		All Cuts	False Positive (All Cuts)	False Negative (All Cuts)	**Level III Cut**	False Positive (Level III Cut)	False Negative (Level III Cut)
3	196476	**0.82**	0.12	0.06	**0.90**	0.07	0.03
4	197040	**0.86**	0.08	0.06	**0.91**	0.04	0.04
5	200195	**0.83**	0.11	0.07	**0.91**	0.05	0.05
6	198076	**0.84**	0.11	0.05	**0.91**	0.06	0.03
7	200140	**0.84**	0.10	0.06	**0.91**	0.05	0.04
8	201278	**0.86**	0.10	0.04	**0.91**	0.06	0.03

Section VIII: Summary of Operational Test Results

This section summarizes the distribution of OP scale score results on the New York State 2011 Grades 3–8 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 code (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 includes examinees who are classified with disabilities and use one or more disability-related accommodations. Data include examinees with valid scores from all public and charter schools. Note that complete scale score frequency distribution tables are located in Appendix H.

Scale Score Distribution Summary

Scale score distribution summary tables are presented and discussed in Tables 45–51. In Table 45, scale score statistics for total populations of students from public and charter schools are presented. In Tables 46–51, 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 45. 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	196476	663.47	21.19	640	653	665	677	687
4	197040	671.84	28.98	639	658	675	688	704
5	200195	667.82	19.47	646	658	668	680	689
6	198076	662.62	18.11	642	652	663	674	681
7	200140	663.71	19.60	642	653	664	674	685
8	201278	655.28	22.15	629	642	656	668	680

Grade 3

Scale score statistics and N-counts of demographic groups for Grade 3 are presented in Table 46. The population scale score mean was 663.47 with a standard deviation of 21.19. By gender subgroup, Females outperformed Males, and the difference was more than 5 scale score points. Asian, Multi-Racial, 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 (669.11). Students from the Big 4 Cities achieved a lower scale score mean than their peers from

schools with other NRC designations and about a half of a standard deviation below the population mean. The SWD, SUA, and ELL subgroups scored, on average, approximately three-fourths of one standard deviation below the mean scale score for the population. The SWD/SUA subgroup, which had a scale score mean about 22 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 (671), Multi-Racial (667), White (669), Average Needs districts (667), and Low Needs districts (674).

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

Demographic Category (Subgroup)		N-count	SS Mean	SS Std Dev	10th %tile	25th %tile	50th %tile	75th %tile	90th %tile
State	All Students	196476	663.47	21.19	640	653	665	677	687
Gender	Female	96054	666.13	20.30	644	656	667	680	687
	Male	100422	660.92	21.71	637	650	663	674	683
Ethnicity	Asian	15960	669.11	21.05	647	659	671	680	691
	Black	36529	656.16	20.71	634	647	658	669	680
	Hispanic	45736	657.13	20.75	634	647	659	669	680
	American Indian	1090	659.50	20.71	636	650	661	671	683
	Multi-Racial	1563	665.01	20.84	641	654	667	677	687
	Unknown	272	667.20	20.84	641	656	667	680	691
	White	95326	668.38	19.94	647	658	669	680	691
NRC	New York City	70884	660.23	21.49	637	650	661	674	683
	Big 4 Cities	8117	650.82	23.89	627	640	653	665	677
	High Needs Urban/Suburban	15372	658.70	20.68	636	648	659	671	680
	High Needs Rural	11148	662.25	20.14	638	651	663	674	683
	Average Needs	57536	666.76	19.50	645	656	667	680	687
	Low Needs	27941	672.59	18.35	653	663	674	683	691
	Charter	4965	662.72	16.31	644	653	663	674	683
SWD	All Codes	28048	642.80	26.15	617	631	645	658	669
SUA	All Codes	48718	648.14	24.17	621	637	651	663	671
ELL	ELL=Y	18144	647.19	22.81	623	638	651	661	669
SWD/SUA	SUA=504 plan codes	24828	641.30	25.88	614	630	644	656	667
ELL/SUA	SUA=ELL codes	16502	647.89	22.09	625	640	651	661	669

Grade 4

Scale score statistics and N-counts of demographic groups for Grade 4 are presented in Table 47. The Grade 4 population (All Students) mean was 671.84, with a standard deviation of 28.98. By gender subgroup, Females outperformed Males, but the difference was less than 7 scale score points. Asian, Multi-Racial, 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 (681.30). Students from the Big 4 Cities achieved a lower scale score mean than their peers from schools with other NRC designations and about a half of a standard deviation below the population mean. The SWD/SUA subgroup had a scale score mean nearly 30 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 (677), Asian (683), White (679), Average Needs districts (677), and Low Needs districts (686).

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

Demographic Category (Subgroup)		N-count	SS Mean	SS Std Dev	10th %tile	25th %tile	50th %tile	75th %tile	90th %tile
State	All Students	197040	671.84	28.98	639	658	675	688	704
Gender	Female	96348	675.25	27.57	644	660	677	691	704
	Male	100692	668.58	29.91	635	655	671	686	701
Ethnicity	Asian	15492	681.30	28.86	649	667	683	697	713
	Black	36838	662.45	28.58	630	649	665	679	694
	Hispanic	44920	663.62	28.71	633	651	667	681	694
	American Indian	955	665.04	27.52	635	651	667	681	694
	Multi-Racial	1377	675.00	30.43	642	660	675	691	709
	Unknown	248	673.02	30.96	644	659	675	691	704
	White	97210	677.71	27.24	648	663	679	694	709
NRC	New York City	70021	668.77	29.68	637	655	671	686	701
	Big 4 Cities	8254	655.24	32.46	621	640	658	675	688
	High Needs Urban/Suburban	15319	663.72	28.48	633	651	667	681	694
	High Needs Rural	11552	667.36	26.87	637	655	669	683	697
	Average Needs	58973	675.23	26.37	646	661	677	691	704
	Low Needs	28461	684.84	25.32	656	671	686	701	713
	Charter	3899	669.22	21.87	642	656	671	683	694
SWD	All Codes	29861	643.44	35.61	608	628	648	665	679
SUA	All Codes	49414	650.19	33.67	615	637	655	671	683
ELL	ELL=Y	16287	647.32	32.40	612	635	653	667	677
SWD/SUA	SUA=504 plan codes	27285	642.08	35.29	608	628	648	663	677
ELL/SUA	SUA=ELL codes	14785	648.30	31.45	615	637	653	667	679

Grade 5

Scale score summary statistics for Grade 5 students are in Table 48. Overall, the scale score mean was 667.82, with a standard deviation of 19.47. The difference between mean scale scores by gender groups was about 5 scale score units. Female, Asian, Multi-Racial, 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 668: Female (669), Asian (675), Multi-Racial (669), White (671), Average Needs (669) and Low Needs districts (675).

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

Demographic Category (Subgroup)		N-count	SS Mean	SS Std Dev	10th %tile	25th %tile	50th %tile	75th %tile	90th %tile
State	All Students	200195	667.82	19.47	646	658	668	680	689
Gender	Female	97835	669.91	18.70	649	659	669	680	689
	Male	102360	665.82	19.98	644	657	666	677	685
Ethnicity	Asian	16633	675.02	21.85	653	665	675	685	700
	Black	37296	661.31	19.06	641	652	662	673	682
	Hispanic	44110	662.58	18.34	642	653	663	673	682
	American Indian	969	662.16	19.86	642	653	663	673	682
	Multi-Racial	1315	669.55	19.06	648	659	669	680	689
	Unknown	240	670.65	19.26	648.5	659	669	682	694
	White	99632	671.40	18.39	652	662	671	682	689
NRC	New York City	69263	666.40	20.00	645	655	666	677	689
	Big 4 Cities	7906	656.17	21.49	634	646	658	669	680
	High Needs Urban/Suburban	14996	662.32	18.98	642	653	663	673	682
	High Needs Rural	11529	664.24	18.99	644	655	665	675	685
	Average Needs	60624	669.64	17.70	650	661	669	680	689
	Low Needs	30062	676.12	17.02	658	668	675	685	694
	Charter	5207	663.47	15.64	645	654	663	673	682
SWD	All Codes	30698	649.02	22.70	626	639	652	662	671
SUA	All Codes	48998	653.02	21.69	631	644	655	666	675
ELL	ELL=Y	13890	649.53	20.95	626	641	653	662	669
SWD/SUA	SUA=504 plan codes	28298	648.36	22.52	626	639	652	662	669
ELL/SUA	SUA=ELL codes	12423	650.02	20.62	629	642	653	662	669

Grade 6

Scale score summary statistics for Grade 6 students are in Table 49. The scale score mean was 662.62, with a standard deviation of 18.11. Female, Asian, Multi-Racial, 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 four-fifths of a standard deviation below the mean scale score for the population. The ELL subgroup, which had a scale score mean more than 26 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 (664), Asian (668), Multi-Racial (666), White (667), Average Needs districts (666), and Low Needs districts (672).

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

Demographic Category (Subgroup)		N-count	SS Mean	SS Std Dev	10th %tile	25th %tile	50th %tile	75th %tile	90th %tile
State	All Students	198076	662.62	18.11	642	652	663	674	681
Gender	Female	96452	664.48	17.71	644	655	664	674	684
	Male	101624	660.85	18.30	640	651	662	672	681
Ethnicity	Asian	15218	667.22	19.98	645	658	668	678	689
	Black	37645	655.89	16.15	637	647	657	666	674
	Hispanic	43253	655.71	16.90	636	646	657	666	674
	American Indian	929	657.90	16.48	638	649	659	667	676
	Multi-Racial	1253	665.40	17.53	646	656	666	674	684
	Unknown	253	662.42	27.52	641	655	664	674	689
	White	99525	667.47	17.19	648	658	667	676	684
NRC	New York City	67620	658.19	18.36	637	648	659	668	678
	Big 4 Cities	7746	653.78	16.95	633	644	655	664	672
	High Needs Urban/Suburban	14604	658.27	16.47	638	649	659	668	676
	High Needs Rural	11576	661.69	16.35	643	652	663	672	681
	Average Needs	60990	665.94	16.37	647	657	666	676	684
	Low Needs	30045	671.72	17.19	653	663	672	681	689
	Charter	4853	659.54	13.50	643	651	659	668	676
SWD	All Codes	30338	645.30	18.10	627	636	647	656	664
SUA	All Codes	44473	646.99	17.98	627	638	648	658	666
ELL	ELL=Y	11975	640.82	18.05	624	633	643	651	658
SWD/SUA	SUA=504 plan codes	27707	644.77	17.84	626	636	646	656	663
ELL/SUA	SUA=ELL codes	10139	640.93	18.12	624	633	643	651	658

Grade 7

Scale score statistics and N-counts of demographic groups for Grade 7 are presented in Table 50. The population scale score mean was 663.71 and the population standard deviation was 19.60. By gender subgroup, Females outperformed Males, the difference was about one-fourth of a standard deviation. Female, Asian, Multi-Racial, 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 (669.28). 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 approximately four-fifths of a standard deviation below the mean scale score for the population. The ELL subgroup, which had a scale score mean more than 25 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 664: Female (666), Asian (670), Multi-Racial (667), White (669), Average Needs districts (667), and Low Needs districts (672).

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

Demographic Category (Subgroup)		N-count	SS Mean	SS Std Dev	10th %tile	25th %tile	50th %tile	75th %tile	90th %tile
State	All Students	200140	663.71	19.60	642	653	664	674	685
Gender	Female	98063	666.38	18.54	645	655	666	676	688
	Male	102077	661.14	20.23	640	651	661	672	685
Ethnicity	Asian	15315	669.28	22.28	645	659	670	682	693
	Black	37951	656.04	17.77	637	646	656	666	676
	Hispanic	42976	656.33	18.66	637	648	658	667	676
	American Indian	982	658.96	19.69	640	649	659	670	682
	Multi-Racial	1111	666.33	18.59	644	655	667	676	688
	Unknown	264	662.58	24.17	639	651	663	674	688
	White	101541	668.88	18.19	649	659	669	679	688
NRC	New York City	68583	659.26	19.72	639	649	659	670	682
	Big 4 Cities	7527	652.03	22.22	630	642	654	664	674
	High Needs Urban/Suburban	14513	657.70	18.60	637	649	659	669	679
	High Needs Rural	11698	662.77	17.62	643	653	663	672	682
	Average Needs	61149	667.27	17.64	648	658	667	679	688
	Low Needs	32076	673.47	17.29	654	664	672	685	693
	Charter	3725	659.65	13.97	643	651	660	669	676
SWD	All Codes	30706	645.15	21.31	626	637	648	656	664
SUA	All Codes	43064	646.62	21.37	626	639	649	658	667
ELL	ELL=Y	10880	638.79	23.39	620	632	642	651	658
SWD/SUA	SUA=504 plan codes	27892	644.65	21.19	626	636	646	656	664
ELL/SUA	SUA=ELL codes	9080	639.05	22.89	620	633	642	651	658

Grade 8

Scale score statistics and N-counts of demographic groups for Grade 8 are presented in Table 51. The population scale score mean was 655.28 with a standard deviation of 22.15. By gender subgroup, Females outperformed Males, but the difference was less than 7 scale score points. Female, Asian, Multi-Racial, 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 656: Female (658), Asian (663), Multi-Racial (660), White (662), Average Needs districts (660), and Low Needs districts (666).

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

Demographic Category (Subgroup)		N-count	SS Mean	SS Std Dev	10th %tile	25th %tile	50th %tile	75th %tile	90th %tile
State	All Students	201278	655.28	22.15	629	642	656	668	680
Gender	Female	98306	658.66	21.54	633	646	658	670	684
	Male	102972	652.06	22.24	627	640	653	666	676
Ethnicity	Asian	15846	661.33	25.10	632	649	663	676	689
	Black	38001	646.94	19.93	624	636	647	658	670
	Hispanic	42651	646.85	20.71	624	636	647	660	670
	American Indian	979	647.37	20.05	624	636	647	660	670
	Multi-Racial	981	659.77	21.85	635	647	660	670	684
	Unknown	240	653.94	26.58	628	641	655	668	682
	White	102580	660.98	20.87	637	649	662	673	684
NRC	New York City	69837	649.97	21.50	625	638	650	663	676
	Big 4 Cities	7503	641.11	24.00	616	629	642	655	666
	High Needs Urban/Suburban	14233	649.48	20.51	625	638	650	662	673
	High Needs Rural	11675	653.73	19.97	631	642	655	666	676
	Average Needs	61628	659.25	20.33	636	647	660	670	684
	Low Needs	32415	667.08	20.21	645	656	666	680	689
	Charter	2856	651.13	16.64	632	641.5	652	662	670
SWD	All Codes	30004	634.00	21.84	610	624	636	647	656
SUA	All Codes	42130	635.33	22.04	612	625	637	649	658
ELL	ELL = Y	10884	626.65	22.10	606	618	629	640	647
SWD/SUA	SUA=504 plan codes	27330	633.53	21.41	610	624	636	646	656
ELL/SUA	SUA=ELL codes	9171	626.86	22.42	606	618	629	640	647

Performance Level Distribution Summary

Students are classified as Level I (Below Standards), Level II (Meets Basic Standards), Level III (Meets Proficiency Standards), and Level IV (Exceeds Proficiency Standards). The

original proficiency cut scores used to distinguish among Levels I, II, III, and IV established during the process of Standard Setting in 2006 were adjusted after the 2010 OP test administration to reflect a change in the test administration window between the 2008–2009 and 2009–2010 school years and the State's policy decision to align the proficiency standards with Grade 8 student performance on the NYS Regents ELA examination.

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

Table 52 shows the ELA cut scores used for classification of students to the four performance level categories in 2011.

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

Content	Grade	2010 NYS cut scores 'Theoretical Cuts'			**2011 NYS cut scores** (Minimum scale score in each proficiency level in 2010 - 'Operational Cuts')		
		Level			**Level**		
		II	**III**	**IV**	**II**	**III**	**IV**
	Column	1	2	3	4	5	6
ELA	3	643	662	694	**644**	**663**	**694**
	4	637	668	720	**637**	**671**	**722**
	5	647	666	700	**648**	**668**	**700**
	6	644	662	694	**644**	**662**	**694**
	7	642	664	698	**642**	**665**	**698**
	8	627	658	699	**628**	**658**	**699**

Tables 53–59 show the performance level distribution for all examinees from public and charter schools with valid scores. Table 53 presents performance level data for total populations of students in Grades 3–8. Tables 54–59 contain performance level data for selected subgroups of students. In general, these distributions reflect the same achievement trends in the scale score summary discussion. More Female students were classified in Level III and above categories as compared to Male students. Similarly more Asian and White students were classified in Level III and above categories as compared to 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. Please note that the case counts for the Unknown subgroup are very low and are heavily influenced by very high and/or very low achieving individual students.

Table 53. 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	196476	11.52	32.50	51.39	4.59	55.98
4	197040	8.20	31.96	57.38	2.46	59.84
5	200195	10.41	31.78	53.39	4.41	57.80
6	198076	11.55	32.55	51.93	3.96	55.89
7	200140	9.23	39.37	47.82	3.58	51.40
8	201278	7.47	45.50	45.23	1.80	47.03

Grade 3

Performance level distributions and N-counts of demographic groups for Grade 3 are presented in Table 54. Statewide, 55.98% of third-graders were Level III and Level IV. 14.07% of Male students were Level I, as compared to only 8.85% of Female students. The percentage of students in Levels III and IV varied widely by ethnicity and NRC subgroups. About 76% of Low Needs district students and about 69% 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 52%–71% 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, Multi-Racial, White, Average Needs districts, and Low Needs districts.

Table 54. 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	196476	11.52	32.50	51.39	4.59	55.98
Gender	Female	96054	8.85	30.03	55.25	5.87	61.11
	Male	100422	14.07	34.87	47.70	3.37	51.07
Ethnicity	Asian	15960	6.74	24.00	61.79	7.47	69.26
	Black	36529	18.34	42.97	37.20	1.49	38.70
	Hispanic	45736	16.81	41.33	40.17	1.69	41.86
	American Indian	1090	15.32	38.35	44.22	2.11	46.33
	Multi-Racial	1563	10.11	30.45	53.29	6.14	59.44
	Unknown	272	10.66	28.31	54.41	6.62	61.03
	White	95326	7.14	25.66	60.51	6.68	67.19
NRC	New York City	70884	14.29	37.47	44.88	3.36	48.24
	Big 4 Cities	8117	27.56	42.74	28.45	1.26	29.70
	High Needs Urban/Suburban	15372	15.72	39.25	42.79	2.24	45.03
	High Needs Rural	11148	12.62	34.15	49.62	3.61	53.23
	Average Needs	57536	7.94	28.33	58.27	5.45	63.73
	Low Needs	27941	3.97	20.01	66.96	9.07	76.02
	Charter	4965	8.64	39.80	49.33	2.24	51.56
SWD	All Codes	28048	41.97	39.29	18.25	0.49	18.74
SUA	All Codes	48718	31.46	42.53	25.38	0.63	26.01
ELL	ELL=Y	18144	30.86	47.60	21.35	0.19	21.54
SWD/SUA	SUA=504 plan codes	24828	44.53	39.43	15.73	0.32	16.05
ELL/SUA	SUA=ELL codes	16502	29.42	48.66	21.72	0.20	21.92

Grade 4

Performance level distributions and N-counts of demographic groups for Grade 4 are presented in Table 55. 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 (79.62%), and the SWD/SUA subgroup had the lowest (18.3%). Students in the Black, Hispanic, and American Indian subgroups had percentages classified in Levels III and IV below 50%, which was more than 12% below the other ethnic subgroups. More than twice as many Big 4 Cities students were in Level I than the State population. About a fourth of the students with ELL, SWD, or SUA status were in Level I (over three times the Statewide rate of 8.2%) 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, Multi-Racial, White, Average Needs districts, and Low Needs districts.

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

Demographic Category (Subgroup)		N-count	Level I %	Level II %	Level III %	Level IV %	Levels III & IV %
State	All Students	197040	8.20	31.96	57.38	2.46	59.84
Gender	Female	96348	6.14	29.60	61.09	3.18	64.27
	Male	100692	10.18	34.23	53.83	1.76	55.59
Ethnicity	Asian	15492	5.35	20.67	68.46	5.52	73.98
	Black	36838	12.74	42.89	43.60	0.78	44.37
	Hispanic	44920	11.90	41.05	46.26	0.79	47.05
	American Indian	955	10.16	39.90	49.32	0.63	49.95
	Multi-Racial	1377	6.54	30.86	57.81	4.79	62.60
	Unknown	248	6.05	32.66	58.87	2.42	61.29
	White	97210	5.24	25.36	66.04	3.36	69.40
NRC	New York City	70021	9.84	35.96	51.86	2.34	54.20
	Big 4 Cities	8254	19.69	46.09	33.66	0.57	34.23
	High Needs Urban/Suburban	15319	11.87	40.89	46.46	0.78	47.24
	High Needs Rural	11552	9.40	37.48	52.04	1.07	53.12
	Average Needs	58973	5.75	28.29	63.56	2.39	65.96
	Low Needs	28461	2.67	17.70	74.46	5.16	79.62
	Charter	3899	6.69	38.47	54.04	0.80	54.83
SWD	All Codes	29861	32.59	47.21	20.00	0.20	20.20
SUA	All Codes	49414	24.64	47.03	28.08	0.26	28.34
ELL	ELL=Y	16287	25.67	51.78	22.46	0.09	22.55
SWD/SUA	SUA=504 plan codes	27285	33.97	47.73	18.20	0.10	18.30
ELL/SUA	SUA=ELL codes	14785	24.20	52.69	23.02	0.09	23.11

Grade 5

Performance level distributions and N-counts of demographic groups for Grade 5 are presented in Table 56. About 57.8% 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 (78.44%). Students in the American Indian, Black, and Hispanic subgroups had rates less than 45% of students classified in Levels III and IV, approximately 17% less than other ethnic subgroups. Over two times as many Big 4 Cities students were in

Level I than the State population. About 30%–48% of the students with ELL, SWD, or SUA status were in Level I (approximately three times as many as the Statewide rate of 10.41%), yet only about 16%–25% 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, Multi-Racial, White, Average Needs districts, and Low Needs districts.

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

Demographic Category (Subgroup)		N-count	Level I %	Level II %	Level III %	Level IV %	Levels III & IV %
State	All Students	200195	10.41	31.78	53.39	4.41	57.80
Gender	Female	97835	8.08	29.84	56.73	5.35	62.08
	Male	102360	12.64	33.64	50.21	3.50	53.71
Ethnicity	Asian	16633	6.41	19.62	63.89	10.08	73.96
	Black	37296	17.06	41.83	39.26	1.84	41.10
	Hispanic	44110	14.87	40.47	42.65	2.01	44.66
	American Indian	969	15.07	41.69	40.66	2.58	43.24
	Multi-Racial	1315	9.73	28.59	55.51	6.16	61.67
	Unknown	240	8.33	32.50	51.25	7.92	59.17
	White	99632	6.59	26.15	61.79	5.47	67.26
NRC	New York City	69263	11.95	35.11	48.42	4.52	52.94
	Big 4 Cities	7906	26.78	41.54	30.51	1.18	31.68
	High Needs Urban/Suburban	14996	15.68	39.76	42.74	1.82	44.56
	High Needs Rural	11529	12.93	37.39	47.14	2.53	49.67
	Average Needs	60624	7.55	29.51	58.72	4.22	62.94
	Low Needs	30062	3.30	18.26	70.48	7.97	78.44
	Charter	5207	12.23	42.69	43.48	1.59	45.07
SWD	All Codes	30698	38.76	43.26	17.63	0.36	17.98
SUA	All Codes	48998	30.68	44.23	24.48	0.61	25.09
ELL	ELL=Y	13890	36.25	47.29	16.30	0.16	16.46
SWD/SUA	SUA=504 plan codes	28298	39.94	43.55	16.26	0.26	16.52
ELL/SUA	SUA=ELL codes	12423	34.89	48.17	16.78	0.15	16.94

Grade 6

Performance level distributions and N-counts of demographic groups for Grade 6 are presented in Table 57. Statewide, 55.89% of Grade 6 students were classified in Levels III and IV. As was seen in other grades, the Low Need subgroup had the most students classified in these two proficiency levels (79.01%), and the ELL, SWD, and SUA subgroups had the fewest. Students in the American Indian, Black, and Hispanic subgroups had about 37%– 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, Multi-Racial, White, Average Needs districts, and Low Needs districts.

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

Demographic Category (Subgroup)		N-count	Level I %	Level II %	Level III %	Level IV %	Levels III & IV %
State	All Students	198076	11.55	32.55	51.93	3.96	55.89
Gender	Female	96452	9.27	31.01	54.84	4.88	59.72
	Male	101624	13.71	34.03	49.18	3.08	52.26
Ethnicity	Asian	15218	8.67	23.20	60.85	7.27	68.12
	Black	37645	18.02	44.53	36.52	0.93	37.45
	Hispanic	43253	19.46	42.70	36.65	1.19	37.84
	American Indian	929	15.61	39.18	43.92	1.29	45.21
	Multi-Racial	1253	8.46	28.49	57.62	5.43	63.05
	Unknown	253	11.86	27.67	53.75	6.72	60.47
	White	99525	6.11	25.05	63.04	5.80	68.84
NRC	New York City	67620	17.03	39.20	41.17	2.60	43.77
	Big 4 Cites	7746	23.29	43.44	32.34	0.93	33.27
	High Needs Urban/Suburban	14604	15.68	39.47	43.32	1.53	44.85
	High Needs Rural	11576	10.59	35.65	51.02	2.74	53.76
	Average Needs	60990	6.88	27.95	60.54	4.63	65.18
	Low Needs	30045	3.20	17.79	70.39	8.62	79.01
	Charter	4853	10.14	44.01	44.76	1.09	45.85
SWD	All Codes	30338	41.08	44.23	14.51	0.18	14.69
SUA	All Codes	44473	36.71	45.31	17.77	0.21	17.98
ELL	ELL=Y	11975	51.76	41.90	6.30	0.03	6.34
SWD/SUA	SUA=504 plan codes	27707	42.27	44.31	13.31	0.10	13.42
ELL/SUA	SUA=ELL codes	10139	51.11	42.48	6.37	0.04	6.41

Grade 7

Performance level distributions and N-counts of demographic groups for Grade 7 are presented in Table 58. In Grade 7, 51.4% 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 75% of Big 4 Cities students were in Levels I and II. About 75% of Low Needs students were in Levels III and IV. About 5% 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 85–96% 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, Multi-Racial, White, Average Needs districts, and Low Needs districts.

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

Demographic Category (Subgroup)		N-count	Level I %	Level II %	Level III %	Level IV %	Levels III & IV %
State	All Students	200140	9.23	39.37	47.82	3.58	51.40
Gender	Female	98063	6.72	36.30	52.43	4.56	56.99
	Male	102077	11.65	42.31	43.39	2.65	46.04
Ethnicity	Asian	15315	7.59	26.09	59.23	7.08	66.31
	Black	37951	14.77	53.57	30.91	0.75	31.66
	Hispanic	42976	15.17	51.76	32.17	0.89	33.06
	American Indian	982	11.61	49.29	37.37	1.73	39.10
	Multi-Racial	1111	7.74	34.65	52.39	5.22	57.61
	Unknown	264	12.88	38.64	43.18	5.30	48.48
	White	101541	4.88	30.77	59.10	5.25	64.35
NRC	New York City	68583	12.87	47.14	37.73	2.26	39.99
	Big 4 Cities	7527	22.93	50.90	25.39	0.78	26.17
	High Needs Urban/Suburban	14513	13.66	49.84	35.31	1.19	36.50
	High Needs Rural	11698	8.47	42.59	46.57	2.37	48.94
	Average Needs	61149	5.52	34.23	56.00	4.25	60.25
	Low Needs	32076	2.56	22.25	67.43	7.76	75.19
	Charter	3725	7.57	54.15	37.58	0.70	38.28
SWD	All Codes	30706	34.67	53.36	11.82	0.15	11.97
SUA	All Codes	43064	31.75	53.76	14.24	0.25	14.49
ELL	ELL=Y	10880	47.99	47.96	4.03	0.03	4.05
SWD/SUA	SUA=504 plan codes	27892	35.46	53.62	10.81	0.10	10.92
ELL/SUA	SUA=ELL codes	9080	47.48	48.66	3.83	0.03	3.87

Grade 8

Performance level distributions and N-counts of demographic groups for Grade 8 are presented in Table 59. In Grade 8, 47.03% of the students were in Levels III and IV. About 12% more Female than Male students were in Levels III or IV. Over 69% of American Indian, Black, and Hispanic students were in Levels I and II. Over 72% 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 88% 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, Multi-Racial, White, Average Needs districts, and Low Needs districts.

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

Demographic Category (Subgroup)		N-count	Level I %	Level II %	Level III %	Level IV %	Levels III & IV %
State	All Students	201278	7.47	45.50	45.23	1.80	47.03
Gender	Female	98306	5.16	41.78	50.58	2.49	53.07
	Male	102972	9.68	49.06	40.13	1.13	41.26
Ethnicity	Asian	15846	7.03	32.02	57.18	3.77	60.95
	Black	38001	11.41	60.42	27.73	0.44	28.17
	Hispanic	42651	12.19	58.63	28.74	0.43	29.18
	American Indian	979	11.24	57.92	30.13	0.72	30.85
	Multi-Racial	981	4.99	39.76	52.40	2.85	55.25
	Unknown	240	8.33	49.58	39.17	2.92	42.08
	White	102580	4.10	36.53	56.82	2.56	59.37
NRC	New York City	69837	10.23	54.63	34.15	0.99	35.14
	Big 4 Cities	7503	19.82	58.50	21.36	0.32	21.68
	High Needs Urban/Suburban	14233	10.87	54.74	33.60	0.79	34.39
	High Needs Rural	11675	7.14	50.00	41.67	1.19	42.86
	Average Needs	61628	4.35	40.28	53.31	2.06	55.37
	Low Needs	32415	1.90	25.87	68.01	4.21	72.23
	Charter	2856	6.37	58.89	34.35	0.39	34.73
SWD	All Codes	30004	28.92	61.90	9.13	0.05	9.18
SUA	All Codes	42130	27.16	61.52	11.22	0.10	11.32
ELL	ELL=Y	10884	41.41	56.20	2.39	0.00	2.39
SWD/SUA	SUA=504 plan codes	27330	29.59	62.06	8.30	0.04	8.34
ELL/SUA	SUA=ELL codes	9171	40.74	56.73	2.53	0.00	2.53

Section IX: Longitudinal Comparison of Results

This section provides longitudinal comparison of OP scale score results on the New York State 2006–2011 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 60.

Table 60. 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	2011	196476	663.47	21.19	11.52	32.50	51.39	4.59	55.98
	2010	196425	667.90	33.09	13.77	31.47	38.11	16.66	54.77
	2009	198123	669.97	35.81	4.75	19.37	65.17	10.72	75.89
	2008	195231	669.00	39.41	5.84	23.92	57.84	12.40	70.24
	2007	198320	666.99	42.23	8.92	23.89	57.29	9.90	67.20
	2006	185533	668.79	40.91	8.53	22.47	61.92	7.07	69.00
4	2011	197040	671.84	28.98	8.20	31.96	57.38	2.46	59.84
	2010	199254	672.82	29.50	8.34	34.82	50.87	5.97	56.84
	2009	195634	669.93	34.72	4.28	18.76	69.69	7.27	76.96
	2008	196367	666.40	39.90	7.34	21.37	62.85	8.44	71.29
	2007	197306	664.70	39.52	7.79	24.17	59.82	8.22	68.04
	2006	190847	665.73	40.74	8.92	22.40	59.94	8.74	68.68
5	2011	200195	667.82	19.47	10.41	31.78	53.39	4.41	57.80
	2010	197200	672.41	32.09	11.54	35.90	39.71	12.85	52.56
	2009	197522	675.47	34.58	0.62	17.09	68.72	13.57	82.29
	2008	197318	667.35	30.89	1.78	20.45	71.83	5.94	77.77
	2007	201841	665.39	37.98	4.89	26.88	61.37	6.86	68.24
	2006	201138	662.69	41.17	6.38	26.45	54.86	12.31	67.17
6	2011	198076	662.62	18.11	11.55	32.55	51.93	3.96	55.89
	2010	197845	664.48	24.67	11.30	34.44	47.40	6.85	54.26
	2009	197674	667.31	27.64	0.13	18.87	71.98	9.02	81.00
	2008	199689	661.45	30.03	1.63	31.20	62.49	4.68	67.17
	2007	204237	661.47	33.98	2.46	34.22	53.93	9.40	63.32
	2006	204104	656.52	40.85	7.28	32.24	48.88	11.60	60.48
7	2011	200140	663.71	19.60	9.23	39.37	47.82	3.58	51.40
	2010	199943	667.91	31.29	10.35	39.53	38.94	11.18	50.12
	2009	202400	667.19	27.06	0.42	19.15	73.51	6.91	80.42
	2008	205946	662.30	29.29	1.75	27.90	67.79	2.56	70.35
	2007	211545	654.84	38.23	5.90	36.22	51.91	5.98	57.89
	2006	210518	652.29	40.95	8.03	35.55	48.66	7.76	56.42

(Continued on next page)

Table 60. 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
8	2011	201278	655.28	22.15	7.47	45.50	45.23	1.80	47.03
	2010	204080	659.07	31.11	8.95	39.98	43.37	7.70	51.06
	2009	207083	661.09	30.82	1.72	29.66	63.75	4.87	68.62
	2008	207646	657.26	37.66	4.95	38.53	50.80	5.73	56.53
	2007	213676	655.39	39.32	6.12	36.75	51.45	5.68	57.13
	2006	212138	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 2011 administrations and the scale scores from the 2006–2011 administrations can be directly compared, the performance level results between 2006–2009 OP tests and 2010–2011 OP tests are not directly comparable because of re-setting the proficiency level cut score values after the 2010 OP test administration.

As seen in Table 60, an increase in scale score means was observed for all ELA grades except Grades 3 and 5 between the 2006 and 2010 test administrations. Grade 3 mean scale score dropped 1 scale score point in 2010, and Grade 5 mean scale score dropped 3 scale score points in 2010. In 2011, the mean scale score for all grades dropped from about 1 scale score point for Grade 4 to above 4 scale score points for Grades 3 and 5. Moderate gain was observed for Grades 4, 5, 6, and 8 for which total gains were 5 or 6 scale score points between the 2006 and 2011 test administrations. The largest gain in scale score points between the 2006 and 2011 test administrations was noted for Grade 7 (11 scale score points). Grade 3 dropped more than 5 scale score points between the 2006 and 2011 test administrations. Relatively steady yearly gain was noticed for Grade 7 with the 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 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, and a moderate decline (approximately 4 points) for Grade 3 and a slight decline (1 point) for Grade 4 between 2010 and 2011. 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, then a slight decline (2–3 scale score points) between 2009 and 2010, and then a moderate decline (approximately 4 points) between 2010 and 2011. 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 points increase was observed between 2008 and 2009. A moderate mean scale score decline (3 scale score points) was observed between 2009 and 2010 and then a slight decline (approximately 2 points) between 2010 and 2011.

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 18 scale score points in 2011. 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 2011. The standard deviation for Grade 5 decreased from approximately 40 scale score points in 2006 to about 31 scale score points in 2008 and then increased to approximately 35 scale score points in 2009, and then it decreased to 32 scale score points in 2010 and to 19 scale score points in 2011. 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 and dropped to around 30 scale score points in 2009, and then it increased to approximately 31 scale score points in 2010 and then decreased to 19 and 22 scale score points in 2011.

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, CTB/McGraw-Hill 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-stereotypic representation of economic classes, races, cultures, ages, genders, and persons with disabilities in context and art

For Constructed-Response Items:

Check that the content of each item is

- designed to assess the targeted performance indicator
- appropriate for the grade level being tested
- presented at a reading level suitable for the grade level being tested
- appropriate in context
- written so that a student possessing knowledge or skill being tested can construct a response that 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 clue to the correct response
- checked for accuracy and documented against reliable, up-to-date sources (including rubrics)
- free of cultural, racial, ethnic, age, gender, disability, or other apparent bias

Check that the format of each item is

- appropriate for the question being asked and 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 2011 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 2011 item selection will include 2005, 2006, 2007, 2008, 2009, and 2010 FT items and items owned by CTB/McGraw-Hill. These items consisted mostly of *TerraNova* items but also included items field-tested in New York State in 2010. 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 they 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 yet not bias if the content is a necessary part of the construct that is measured. That is, some items may be flagged for DIF not out of chance and still not represent bias.
- Verify that the items will be administered in the same relative positions in both the FT and OP forms (e.g., the first item in a FT form should also be the first item in an OP form). When that is impossible, please ensure that they are in the same one-third section of the forms.
- Evaluate the alignment of TCC and SE curves of the proposed 2011 OP forms and the 2010 OP forms. Select the 2011 forms to be more difficult than the 2010 forms.
- Try to get the best scale coverage—make sure that MC items cover a wide range of the scale.
- Provide the research department with the following item selection information:
    - Percentage of score points per learning standard (target, 2011 full selection, 2011 MC items only)
    - Item number in 2011 OP book
    - Item unique identification number, item type, FT year, FT form, and FT item number
    - Item classical statistics (p-values, point biserials, etc.)
    - ITEMWIN output (including TCCs)
    - Summary file with IRT item parameters for selected items

Appendix 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 accommodation (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)

Grade	Subgroup	Initial Eigenvalues			
		Component	Total	% of Variance	Cumulative %
3	ELL	**1**	**9.65**	**18.92**	**18.92**
		2	1.64	3.22	22.15
		3	1.24	2.42	24.57
		4	1.08	2.11	26.68
		5	1.06	2.07	28.75
		6	1.05	2.05	30.80
		7	1.03	2.03	32.82
		8	1.01	1.99	34.81
		9	1.01	1.97	36.79
	SWD	**1**	**10.62**	**20.81**	**20.81**
		2	1.72	3.38	24.19
		3	1.48	2.91	27.10
		4	1.10	2.17	29.27
		5	1.09	2.14	31.41
		6	1.03	2.01	33.42
		7	1.00	1.96	35.38
	SUA	**1**	**10.57**	**20.72**	**20.72**
		2	1.70	3.33	24.05
		3	1.39	2.73	26.78
		4	1.10	2.16	28.94
		5	1.05	2.07	31.01
		6	1.02	2.00	33.00
		7	1.00	1.96	34.96
	SWD /SUA	**1**	**10.24**	**20.07**	**20.07**
		2	1.70	3.34	23.41
		3	1.48	2.90	26.31
		4	1.12	2.19	28.50
		5	1.10	2.16	30.65
		6	1.03	2.03	32.68
		7	1.01	1.99	34.67
		8	1.00	1.96	36.63

(Continued on next page)

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

Grade	Subgroup	Initial Eigenvalues			
		Component	Total	% of Variance	Cumulative %
3	ELL /SUA	**1**	**9.39**	**18.42**	**18.42**
		2	1.64	3.22	21.64
		3	1.23	2.41	24.05
		4	1.08	2.12	26.17
		5	1.06	2.07	28.24
		6	1.05	2.06	30.30
		7	1.04	2.04	32.34
		8	1.02	1.99	34.33
		9	1.01	1.97	36.30
4	ELL	**1**	**9.33**	**15.81**	**15.81**
		2	1.47	2.49	18.29
		3	1.26	2.13	20.42
		4	1.20	2.03	22.45
		5	1.14	1.93	24.38
		6	1.08	1.83	26.21
		7	1.06	1.79	28.01
		8	1.05	1.78	29.78
		9	1.02	1.73	31.51
		10	1.01	1.70	33.22
	SWD	**1**	**10.17**	**17.24**	**17.24**
		2	1.61	2.74	19.97
		3	1.38	2.33	22.30
		4	1.24	2.10	24.40
		5	1.16	1.96	26.36
		6	1.10	1.86	28.22
		7	1.04	1.76	29.97
		8	1.03	1.74	31.71
		9	1.01	1.71	33.43
	SUA	**1**	**10.46**	**17.73**	**17.73**
		2	1.57	2.66	20.39
		3	1.34	2.27	22.66
		4	1.21	2.06	24.72
		5	1.15	1.95	26.67
		6	1.07	1.82	28.49
		7	1.03	1.75	30.23
		8	1.01	1.71	31.94

(Continued on next page)

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

Grade	Subgroup	Initial Eigenvalues			
		Component	Total	% of Variance	Cumulative %
4	SWD /SUA	**1**	**9.82**	**16.65**	**16.65**
		2	1.63	2.76	19.40
		3	1.36	2.31	21.71
		4	1.25	2.12	23.83
		5	1.16	1.96	25.79
		6	1.1	1.87	27.66
		7	1.04	1.77	29.42
		8	1.03	1.75	31.17
		9	1.01	1.72	32.89
	ELL /SUA	**1**	**9.16**	**15.52**	**15.52**
		2	1.46	2.47	17.99
		3	1.26	2.13	20.12
		4	1.2	2.03	22.15
		5	1.15	1.95	24.10
		6	1.08	1.83	25.93
		7	1.06	1.79	27.72
		8	1.05	1.79	29.50
		9	1.02	1.73	31.23
		10	1.01	1.71	32.95
		11	1.00	1.70	34.65
5	ELL	**1**	**8.27**	**16.22**	**16.22**
		2	1.54	3.02	19.24
		3	1.24	2.43	21.67
		4	1.10	2.15	23.82
		5	1.09	2.13	25.95
		6	1.06	2.07	28.02
		7	1.05	2.05	30.07
		8	1.02	2.01	32.08
		9	1.02	2.00	34.08
	SWD	**1**	**9.08**	**17.80**	**17.80**
		2	1.65	3.24	21.04
		3	1.27	2.48	23.52
		4	1.15	2.25	25.77
		5	1.07	2.10	27.86
		6	1.05	2.06	29.92
		7	1.01	1.98	31.90
		8	1.00	1.96	33.86

(Continued on next page)

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

Grade	Subgroup	Initial Eigenvalues			
		Component	Total	% of Variance	Cumulative %
5	SUA	**1**	**9.36**	**18.36**	**18.36**
		2	1.59	3.12	21.48
		3	1.22	2.39	23.86
		4	1.13	2.22	26.08
		5	1.05	2.05	28.13
		6	1.04	2.03	30.16
		7	1.01	1.98	32.14
	SWD /SUA	**1**	**8.86**	**17.36**	**17.36**
		2	1.65	3.24	20.60
		3	1.27	2.49	23.09
		4	1.14	2.24	25.33
		5	1.08	2.11	27.44
		6	1.06	2.07	29.51
		7	1.02	2.00	31.50
		8	1.00	1.97	33.47
	ELL /SUA	**1**	**8.20**	**16.07**	**16.07**
		2	1.55	3.03	19.10
		3	1.24	2.43	21.53
		4	1.10	2.15	23.68
		5	1.09	2.14	25.82
		6	1.06	2.08	27.90
		7	1.05	2.05	29.95
		8	1.03	2.02	31.97
		9	1.01	1.98	33.96
		10	1.01	1.98	35.94
6	ELL	**1**	**7.91**	**13.88**	**13.88**
		2	1.67	2.93	16.81
		3	1.47	2.58	19.38
		4	1.27	2.23	21.61
		5	1.16	2.03	23.64
		6	1.13	1.98	25.61
		7	1.11	1.94	27.55

(Continued on next page)

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

Grade	Subgroup	Initial Eigenvalues			
		Component	Total	% of Variance	Cumulative %
6	ELL	8	1.06	1.86	29.41
		9	1.05	1.84	31.26
		10	1.03	1.81	33.07
		11	1.02	1.79	34.85
		12	1.01	1.78	36.63
		13	1.01	1.77	38.40
	SWD	**1**	**9.41**	**16.52**	**16.52**
		2	1.77	3.10	19.62
		3	1.44	2.53	22.14
		4	1.32	2.31	24.45
		5	1.11	1.95	26.41
		6	1.09	1.91	28.31
		7	1.06	1.86	30.17
		8	1.04	1.82	31.99
		9	1.02	1.79	33.78
		10	1.00	1.76	35.54
	SUA	**1**	**9.68**	**16.99**	**16.99**
		2	1.75	3.07	20.05
		3	1.47	2.57	22.63
		4	1.30	2.27	24.90
		5	1.09	1.92	26.82
		6	1.08	1.89	28.71
		7	1.05	1.84	30.55
		8	1.03	1.80	32.35
		9	1.01	1.77	34.12
	SWD /SUA	**1**	**9.18**	**16.11**	**16.11**
		2	1.77	3.10	19.21
		3	1.44	2.52	21.73
		4	1.32	2.32	24.05
		5	1.12	1.97	26.02
		6	1.09	1.91	27.93
		7	1.06	1.87	29.80
		8	1.04	1.83	31.62

(Continued on next page)

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

Grade	Subgroup	Initial Eigenvalues			
		Component	Total	% of Variance	Cumulative %
6	SWD /SUA	9	1.02	1.80	33.42
		10	1.01	1.77	35.19
	ELL /SUA	**1**	**7.93**	**13.91**	**13.91**
		2	1.65	2.90	16.81
		3	1.48	2.60	19.41
		4	1.27	2.22	21.63
		5	1.16	2.03	23.66
		6	1.13	1.98	25.63
		7	1.11	1.95	27.59
		8	1.07	1.88	29.47
		9	1.06	1.86	31.32
		10	1.04	1.82	33.14
		11	1.03	1.81	34.95
		12	1.02	1.79	36.74
		13	1.00	1.76	38.49
7	ELL	**1**	**7.68**	**13.48**	**13.48**
		2	1.75	3.08	16.55
		3	1.25	2.20	18.75
		4	1.20	2.10	20.85
		5	1.13	1.97	22.82
		6	1.10	1.94	24.76
		7	1.08	1.90	26.66
		8	1.07	1.88	28.54
		9	1.06	1.86	30.40
		10	1.05	1.84	32.24
		11	1.04	1.82	34.07
		12	1.03	1.81	35.88
		13	1.01	1.77	37.65
		14	1.01	1.77	39.42
	SWD	**1**	**9.31**	**16.33**	**16.33**
		2	1.86	3.27	19.60
		3	1.35	2.37	21.97
		4	1.19	2.08	24.06
		5	1.10	1.93	25.99
		6	1.07	1.88	27.87
		7	1.05	1.83	29.70
		8	1.02	1.79	31.49
		9	1.02	1.78	33.27

(Continued on next page)

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

Grade	Subgroup	Initial Eigenvalues			
		Component	Total	% of Variance	Cumulative %
7	SWD	10	1.01	1.77	35.04
	SUA	**1**	**9.67**	**16.97**	**16.97**
		2	1.85	3.25	20.23
		3	1.33	2.33	22.56
		4	1.21	2.12	24.67
		5	1.09	1.90	26.58
		6	1.05	1.85	28.42
		7	1.04	1.82	30.24
		8	1.02	1.79	32.03
		9	1.02	1.78	33.81
	SWD /SUA	**1**	**9.10**	**15.96**	**15.96**
		2	1.86	3.27	19.23
		3	1.35	2.37	21.6
		4	1.19	2.08	23.68
		5	1.10	1.94	25.61
		6	1.07	1.87	27.49
		7	1.05	1.84	29.33
		8	1.02	1.80	31.13
		9	1.02	1.79	32.92
		10	1.02	1.79	34.71
		11	1.00	1.76	36.47
	ELL /SUA	**1**	**7.65**	**13.42**	**13.42**
		2	1.76	3.08	16.50
		3	1.25	2.19	18.69
		4	1.19	2.09	20.78
		5	1.13	1.99	22.76
		6	1.11	1.95	24.71
		7	1.19	1.93	26.64
		8	1.07	1.88	28.52
		9	1.06	1.86	30.38
		10	1.06	1.86	32.24
		11	1.04	1.83	34.07
		12	1.03	1.81	35.87

(Continued on next page)

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

Grade	Subgroup	Initial Eigenvalues			
		Component	Total	% of Variance	Cumulative %
7	ELL /SUA	13	1.02	1.79	37.66
		14	1.01	1.77	39.43
8	ELL	**1**	**8.02**	**14.06**	**14.06**
		2	1.87	3.29	17.35
		3	1.46	2.56	19.91
		4	1.18	2.08	21.98
		5	1.14	2.00	23.99
		6	1.11	1.94	25.93
		7	1.09	1.91	27.84
		8	1.08	1.89	29.73
		9	1.06	1.85	31.58
		10	1.04	1.82	33.40
		11	1.02	1.78	35.19
		12	1.01	1.78	36.96
	SWD	**1**	**9.72**	**17.05**	**17.05**
		2	1.85	3.25	20.30
		3	1.53	2.68	22.98
		4	1.19	2.08	25.06
		5	1.09	1.90	26.97
		6	1.07	1.87	28.84
		7	1.04	1.82	30.66
		8	1.02	1.78	32.44
	SUA	**1**	**10.11**	**17.73**	**17.73**
		2	1.88	3.29	21.02
		3	1.53	2.69	23.71
		4	1.17	2.05	25.76
		5	1.07	1.87	27.63
		6	1.06	1.85	29.48
		7	1.03	1.81	31.29
		8	1.01	1.78	33.07
	SWD /SUA	**1**	**9.49**	**16.66**	**16.66**
		2	1.84	3.23	19.89
		3	1.52	2.67	22.56
		4	1.20	2.10	24.66
		5	1.09	1.91	26.57
		6	1.08	1.89	28.45
		7	1.04	1.83	30.28
		8	1.02	1.79	32.07
		9	1.00	1.76	33.83

(Continued on next page)

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

Grade	Subgroup	Initial Eigenvalues			
		Component	Total	% of Variance	Cumulative %
8	ELL /SUA	**1**	**8.09**	**14.20**	**14.20**
		2	1.89	3.31	17.51
		3	1.45	2.55	20.05
		4	1.18	2.07	22.12
		5	1.14	2.00	24.12
		6	1.10	1.94	26.06
		7	1.10	1.93	27.99
		8	1.08	1.89	29.88
		9	1.05	1.85	31.73
		10	1.03	1.81	33.54
		11	1.02	1.79	35.33
		12	1.02	1.78	37.11
		13	1.00	1.76	38.87

Appendix E—Items Flagged for DIF

These tables support the DIF information in Section V, "Operational Test Data Collection and Classical Analysis," and Section VI, "IRT Scaling and Equating." They include item numbers, focal group, and directions of DIF and DIF statistics. Table E1 shows items flagged by the SMD and Mantel-Haenszel methods, and Table E2 presents items flagged by the Linn-Harnisch method. 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 2011 Classical DIF Item Flags

Grade	Item #	Subgroup	DIF	SMD	Mantel-Haenszel	Delta
3	18	Asian	Against	No Flag	557.817	-1.599
3	20	Asian	Against	No Flag	697.724	-1.566
3	20	Hispanic	Against	-0.101	1919.360	-1.639
3	41	Black	Against	-0.162	No Flag	No Flag
3	46	Asian	Against	-0.102	1095.430	-1.679
3	46	Black	Against	-0.110	No Flag	No Flag
3	50	Asian	In Favor	0.143	No Flag	No Flag
3	50	Black	In Favor	0.170	No Flag	No Flag
3	50	Hispanic	In Favor	0.144	No Flag	No Flag
3	50	ELL	In Favor	0.101	No Flag	No Flag
3	50	High Needs	In Favor	0.114	No Flag	No Flag
3	51	Asian	In Favor	0.144	No Flag	No Flag
3	51	Hispanic	In Favor	0.101	No Flag	No Flag
3	51	Female	In Favor	0.113	No Flag	No Flag
4	2	Asian	Against	-0.130	1407.490	-1.713
4	13	Asian	Against	-0.136	1468.050	-1.767
4	35	Asian	Against	No Flag	575.809	-1.555
4	35	Hispanic	Against	No Flag	1929.980	-1.690
4	35	ELL	Against	-0.134	1492.040	-1.656
4	49	Black	In Favor	0.106	No Flag	No Flag
4	51	Asian	In Favor	0.111	No Flag	No Flag
4	51	Black	In Favor	0.115	No Flag	No Flag
4	51	Hispanic	In Favor	0.100	No Flag	No Flag
4	54	Asian	Against	-0.104	No Flag	No Flag
4	55	Asian	In Favor	0.105	No Flag	No Flag
4	55	Black	In Favor	0.120	No Flag	No Flag
4	55	Hispanic	In Favor	0.115	No Flag	No Flag
4	55	ELL	In Favor	0.145	No Flag	No Flag
4	56	Asian	In Favor	0.165	No Flag	No Flag
4	56	Black	In Favor	0.102	No Flag	No Flag
4	56	Hispanic	In Favor	0.122	No Flag	No Flag
4	56	ELL	In Favor	0.142	No Flag	No Flag

(Continued on next page)

Table E1. NYSTP ELA 2011 Classical DIF Item Flags (cont.)

Grade	Item #	Subgroup	DIF	SMD	Mantel-Haenszel	Delta
4	56	High Needs	In Favor	0.104	No Flag	No Flag
4	58	Asian	In Favor	0.122	No Flag	No Flag
4	59	Asian	In Favor	0.138	No Flag	No Flag
4	59	Female	In Favor	0.104	No Flag	No Flag
5	5	ELL	Against	-0.108	No Flag	No Flag
5	9	Asian	Against	No Flag	727.959	-2.271
5	47	ELL	Against	-0.105	No Flag	No Flag
5	48	Hispanic	In Favor	0.106	No Flag	No Flag
5	48	ELL	In Favor	0.100	No Flag	No Flag
5	50	Hispanic	In Favor	0.102	No Flag	No Flag
5	50	Female	In Favor	0.101	No Flag	No Flag
5	51	Hispanic	In Favor	0.120	No Flag	No Flag
5	51	Female	In Favor	0.119	No Flag	No Flag
6	30	Female	Against	No Flag	3259.182	-1.859
6	34	Asian	Against	-0.113	No Flag	No Flag
6	47	ELL	Against	-0.146	No Flag	No Flag
6	48	ELL	Against	-0.112	No Flag	No Flag
6	51	Asian	Against	-0.148	1496.094	-1.723
6	53	ELL	In Favor	0.124	No Flag	No Flag
6	54	Asian	In Favor	0.112	No Flag	No Flag
6	54	ELL	In Favor	0.118	No Flag	No Flag
6	55	Asian	In Favor	0.176	No Flag	No Flag
6	55	ELL	In Favor	0.152	No Flag	No Flag
6	56	Asian	In Favor	0.152	No Flag	No Flag
6	57	Asian	In Favor	0.189	No Flag	No Flag
6	57	Female	In Favor	0.131	No Flag	No Flag
6	57	ELL	In Favor	0.203	No Flag	No Flag
7	4	ELL	Against	-0.102	No Flag	No Flag
7	6	Asian	Against	No Flag	673.517	-1.696
7	6	ELL	Against	-0.111	No Flag	No Flag
7	19	Asian	Against	No Flag	1104.738	-2.001
7	19	Hispanic	Against	-0.152	3774.372	-2.236
7	19	ELL	Against	-0.161	1382.357	-2.074
7	19	High Needs	Against	No Flag	2348.111	-1.538
7	25	Asian	Against	-0.114	No Flag	No Flag
7	25	Black	Against	-0.102	No Flag	No Flag
7	25	Hispanic	Against	-0.125	No Flag	No Flag
7	33	Asian	Against	No Flag	233.770	-1.601
7	33	ELL	Against	-0.105	No Flag	No Flag
7	43	Asian	Against	No Flag	1077.735	-1.693
7	50	Asian	Against	No Flag	478.594	-1.538
7	53	Black	In Favor	0.102	No Flag	No Flag

(Continued on next page)

Table E1. NYSTP ELA 2011 Classical DIF Item Flags (cont.)

Grade	Item #	Subgroup	DIF	SMD	Mantel-Haenszel	Delta
7	54	Black	In Favor	0.127	No Flag	No Flag
7	54	Hispanic	In Favor	0.107	No Flag	No Flag
7	54	Female	In Favor	0.123	No Flag	No Flag
7	54	ELL	In Favor	0.153	No Flag	No Flag
7	56	Asian	In Favor	0.141	No Flag	No Flag
7	56	Black	In Favor	0.119	No Flag	No Flag
7	56	Hispanic	In Favor	0.130	No Flag	No Flag
7	56	Female	In Favor	0.106	No Flag	No Flag
7	56	ELL	In Favor	0.162	No Flag	No Flag
7	57	Asian	In Favor	0.121	No Flag	No Flag
7	57	Hispanic	In Favor	0.114	No Flag	No Flag
7	57	Female	In Favor	0.179	No Flag	No Flag
8	11	ELL	Against	-0.112	No Flag	No Flag
8	17	Asian	Against	No Flag	373.990	-1.663
8	17	ELL	Against	-0.113	746.131	-1.610
8	25	Asian	In Favor	0.129	1310.650	1.982
8	25	Black	In Favor	0.135	2102.918	1.594
8	25	Hispanic	In Favor	0.130	2083.011	1.509
8	25	High Needs	In Favor	0.118	No Flag	No Flag
8	34	Asian	Against	No Flag	1306.601	-2.282
8	34	Hispanic	Against	No Flag	1393.184	-1.513
8	34	Female	Against	No Flag	2369.414	-1.614
8	34	ELL	Against	-0.151	1041.503	-1.703
8	49	Asian	In Favor	0.105	No Flag	No Flag
8	55	Asian	In Favor	0.135	No Flag	No Flag
8	55	Hispanic	In Favor	0.108	No Flag	No Flag
8	55	ELL	In Favor	0.205	No Flag	No Flag
8	56	ELL	In Favor	0.200	No Flag	No Flag
8	57	Asian	In Favor	0.120	No Flag	No Flag
8	57	Female	In Favor	0.161	No Flag	No Flag
8	57	ELL	In Favor	0.147	No Flag	No Flag

In Table E2, note that positive values of D_{ig} indicate DIF in favor of a focal group and negative values of D_{ig} indicate DIF against a focal group.

Table E2. Items Flagged for DIF by the Linn-Harnisch Method

Grade	Item	Focal Group	Direction	Magnitude (D_{ig})
3	41	Black	Against	-0.106
4	13	Asian	Against	-0.114
4	35	ELL	Against	-0.108
4	55	ELL	In Favor	0.124
4	56	Asian	In Favor	0.119
4	56	ELL	In Favor	0.111
4	59	Asian	In Favor	0.105
5	48	ELL	In Favor	0.105
5	50	ELL	In Favor	0.104
5	51	ELL	In Favor	0.107
6	51	Asian	Against	-0.112
6	54	ELL	In Favor	0.101
6	55	ELL	In Favor	0.120
6	55	Asian	In Favor	0.146
6	56	Asian	In Favor	0.117
6	57	ELL	In Favor	0.189
6	57	Asian	In Favor	0.179
7	19	ELL	Against	-0.137
7	54	ELL	In Favor	0.147
7	56	ELL	In Favor	0.138
8	34	ELL	Against	-0.123
8	55	ELL	In Favor	0.171
8	55	Asian	In Favor	0.105
8	56	ELL	In Favor	0.184
8	56	Asian	In Favor	0.111
8	57	ELL	In Favor	0.163
8	57	Asian	In Favor	0.138

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

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

where

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

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

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

where

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

and

$$\gamma_{i0} = 0$$

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

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

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

where

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

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

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

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

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

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

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

and

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

Solving these equations for r_j and s_j produces

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

and

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

where

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

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

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

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

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

where

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

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

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

and

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

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 (correct score) (OPM), x_j / n_j, as

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

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

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

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

Check on Consistency and Adjustment of Weight Given to Prior Estimate

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

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

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

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

Possible Violations of the Assumptions

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

If the items in the standard do not permit guessing, it is reasonable to assume $\hat{T}_j$, the expected proportion correct of 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 H*H contingency table consist of the joint probabilities of the row and column observed category classifications.

That two administrations are independent implies that if X_1 and X_2 represent the raw score random variables on the two administrations, then, conditioned on θ, X_1 and X_2 are independent and identically distributed. Consequently, the conditional bivariate distribution of X_1 and X_2 is

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

The marginal bivariate distribution of X_1 and X_2 can be expressed as follows:

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

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

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

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

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

The agreement index (classification consistency) can be computed as

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

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

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

Then, the coefficient kappa (Cohen, 1960) is

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

Classification Accuracy

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

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

where

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

Appendix H—Scale Score Frequency Distributions

Tables H1–H6 depict the scale score (SS) distributions, by frequency (N-count), percent, cumulative frequency, and cumulative percent. The data in the tables include all public and charter school students with valid scale scores.

Table H1. Grade 3 ELA 2011 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
475	298	0.15	298	0.15
545	185	0.09	483	0.25
581	257	0.13	740	0.38
592	324	0.16	1064	0.54
599	387	0.20	1451	0.74
604	453	0.23	1904	0.97
608	532	0.27	2436	1.24
612	574	0.29	3010	1.53
614	648	0.33	3658	1.86
617	623	0.32	4281	2.18
619	713	0.36	4994	2.54
621	740	0.38	5734	2.92
623	793	0.40	6527	3.32
625	877	0.45	7404	3.77
627	838	0.43	8242	4.19
628	931	0.47	9173	4.67
630	993	0.51	10166	5.17
631	1096	0.56	11262	5.73
633	1160	0.59	12422	6.32
634	1286	0.65	13708	6.98
636	1463	0.74	15171	7.72
637	1644	0.84	16815	8.56
638	1753	0.89	18568	9.45
640	1942	0.99	20510	10.44
641	2119	1.08	22629	11.52
643	2395	1.22	25024	12.74
644	2691	1.37	27715	14.11
645	3068	1.56	30783	15.67
647	3448	1.75	34231	17.42

(Continued on next page)

Table H1. Grade 3 ELA 2011 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
648	3708	1.89	37939	19.31
650	4260	2.17	42199	21.48
651	4661	2.37	46860	23.85
653	5247	2.67	52107	26.52
654	5722	2.91	57829	29.43
656	6210	3.16	64039	32.59
658	6784	3.45	70823	36.05
659	7633	3.88	78456	39.93
661	8035	4.09	86491	44.02
663	8579	4.37	95070	48.39
665	9161	4.66	104231	53.05
667	9512	4.84	113743	57.89
669	9949	5.06	123692	62.96
671	10199	5.19	133891	68.15
674	10389	5.29	144280	73.43
677	10205	5.19	154485	78.63
680	9866	5.02	164351	83.65
683	8905	4.53	173256	88.18
687	7959	4.05	181215	92.23
691	6245	3.18	187460	95.41
696	4553	2.32	192013	97.73
703	2772	1.41	194785	99.14
714	1351	0.69	196136	99.83
780	340	0.17	196476	100.00

Table H2. Grade 4 ELA 2011 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
430	407	0.21	407	0.21
540	247	0.13	654	0.33
564	358	0.18	1012	0.51
578	444	0.23	1456	0.74
586	471	0.24	1927	0.98
593	562	0.29	2489	1.26
599	596	0.30	3085	1.57

(Continued on next page)

Table H2. Grade 4 ELA 2011 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
604	686	0.35	3771	1.91
608	790	0.40	4561	2.31
612	819	0.42	5380	2.73
615	861	0.44	6241	3.17
618	976	0.50	7217	3.66
621	995	0.50	8212	4.17
623	1155	0.59	9367	4.75
626	1226	0.62	10593	5.38
628	1244	0.63	11837	6.01
630	1308	0.66	13145	6.67
633	1465	0.74	14610	7.41
635	1553	0.79	16163	8.2
637	1742	0.88	17905	9.09
639	1877	0.95	19782	10.04
640	2007	1.02	21789	11.06
642	2173	1.10	23962	12.16
644	2419	1.23	26381	13.39
646	2532	1.29	28913	14.67
648	2763	1.40	31676	16.08
649	2954	1.50	34630	17.58
651	3068	1.56	37698	19.13
653	3370	1.71	41068	20.84
655	3700	1.88	44768	22.72
656	3978	2.02	48746	24.74
658	4250	2.16	52996	26.9
660	4608	2.34	57604	29.23
661	4992	2.53	62596	31.77
663	5148	2.61	67744	34.38
665	5586	2.83	73330	37.22
667	5811	2.95	79141	40.16
669	6067	3.08	85208	43.24
671	6337	3.22	91545	46.46
673	6655	3.38	98200	49.84
675	7139	3.62	105339	53.46
677	7073	3.59	112412	57.05
679	7432	3.77	119844	60.82

(Continued on next page)

Table H2. Grade 4 ELA 2011 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
681	7485	3.80	127329	64.62
683	7563	3.84	134892	68.46
686	7576	3.84	142468	72.30
688	7288	3.70	149756	76.00
691	7096	3.60	156852	79.60
694	6857	3.48	163709	83.08
697	6369	3.23	170078	86.32
701	5861	2.97	175939	89.29
704	5179	2.63	181118	91.92
709	4466	2.27	185584	94.19
713	3729	1.89	189313	96.08
719	2887	1.47	192200	97.54
725	2109	1.07	194309	98.61
733	1395	0.71	195704	99.32
745	858	0.44	196562	99.76
766	360	0.18	196922	99.94
775	118	0.06	197040	100.00

Table H3. Grade 5 ELA 2011 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
495	305	0.15	305	0.15
579	165	0.08	470	0.23
593	252	0.13	722	0.36
601	252	0.13	974	0.49
607	317	0.16	1291	0.64
611	380	0.19	1671	0.83
615	427	0.21	2098	1.05
618	515	0.26	2613	1.31
621	534	0.27	3147	1.57
624	624	0.31	3771	1.88
626	675	0.34	4446	2.22
629	756	0.38	5202	2.60
631	831	0.42	6033	3.01
632	888	0.44	6921	3.46
634	985	0.49	7906	3.95

(Continued on next page)

Table H3. Grade 5 ELA 2011 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
636	1102	0.55	9008	4.50
638	1259	0.63	10267	5.13
639	1425	0.71	11692	5.84
641	1531	0.76	13223	6.61
642	1672	0.84	14895	7.44
644	1791	0.89	16686	8.33
645	1990	0.99	18676	9.33
646	2172	1.08	20848	10.41
648	2399	1.20	23247	11.61
649	2643	1.32	25890	12.93
650	2946	1.47	28836	14.40
652	3138	1.57	31974	15.97
653	3518	1.76	35492	17.73
654	3839	1.92	39331	19.65
655	4151	2.07	43482	21.72
657	4537	2.27	48019	23.99
658	4944	2.47	52963	26.46
659	5407	2.70	58370	29.16
661	5806	2.90	64176	32.06
662	6332	3.16	70508	35.22
663	6669	3.33	77177	38.55
665	7299	3.65	84476	42.20
666	7696	3.84	92172	46.04
668	8230	4.11	100402	50.15
669	8759	4.38	109161	54.53
671	9209	4.60	118370	59.13
673	9654	4.82	128024	63.95
675	9912	4.95	137936	68.90
677	9998	4.99	147934	73.89
680	9982	4.99	157916	78.88
682	9712	4.85	167628	83.73
685	9130	4.56	176758	88.29
689	8055	4.02	184813	92.32
694	6557	3.28	191370	95.59
700	4636	2.32	196006	97.91
710	2785	1.39	198791	99.30
727	1188	0.59	199979	99.89
795	216	0.11	200195	100.00

Table H4. Grade 6 ELA 2011 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
480	148	0.07	148	0.07
572	88	0.04	236	0.12
591	132	0.07	368	0.19
600	149	0.08	517	0.26
606	240	0.12	757	0.38
611	297	0.15	1054	0.53
614	366	0.18	1420	0.72
617	414	0.21	1834	0.93
619	509	0.26	2343	1.18
622	564	0.28	2907	1.47
624	648	0.33	3555	1.79
626	713	0.36	4268	2.15
627	840	0.42	5108	2.58
629	910	0.46	6018	3.04
630	1018	0.51	7036	3.55
632	1091	0.55	8127	4.10
633	1183	0.60	9310	4.70
635	1333	0.67	10643	5.37
636	1376	0.69	12019	6.07
637	1458	0.74	13477	6.80
638	1668	0.84	15145	7.65
640	1633	0.82	16778	8.47
641	1936	0.98	18714	9.45
642	2030	1.02	20744	10.47
643	2136	1.08	22880	11.55
644	2372	1.20	25252	12.75
645	2471	1.25	27723	14.00
646	2694	1.36	30417	15.36
647	2833	1.43	33250	16.79
648	3213	1.62	36463	18.41
649	3175	1.60	39638	20.01
650	3556	1.80	43194	21.81
651	3822	1.93	47016	23.74
652	4032	2.04	51048	25.77
653	4198	2.12	55246	27.89
655	4555	2.30	59801	30.19

(Continued on next page)

Table H4. Grade 6 ELA 2011 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
656	4818	2.43	64619	32.62
657	5135	2.59	69754	35.22
658	5533	2.79	75287	38.01
659	5806	2.93	81093	40.94
660	6270	3.17	87363	44.11
662	6513	3.29	93876	47.39
663	6766	3.42	100642	50.81
664	7078	3.57	107720	54.38
666	7673	3.87	115393	58.26
667	7858	3.97	123251	62.22
668	8142	4.11	131393	66.33
670	8268	4.17	139661	70.51
672	8400	4.24	148061	74.75
674	8405	4.24	156466	78.99
676	8067	4.07	164533	83.07
678	7428	3.75	171961	86.82
681	7037	3.55	178998	90.37
684	6112	3.09	185110	93.45
689	5123	2.59	190233	96.04
694	3794	1.92	194027	97.96
701	2438	1.23	196465	99.19
715	1223	0.62	197688	99.80
785	388	0.20	198076	100.00

Table H5. Grade 7 ELA 2011 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
470	283	0.14	283	0.14
572	144	0.07	427	0.21
590	191	0.10	618	0.31
599	247	0.12	865	0.43
605	301	0.15	1166	0.58
610	388	0.19	1554	0.78
614	450	0.22	2004	1.00
617	527	0.26	2531	1.26

(Continued on next page)

Table H5. Grade 7 ELA 2011 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
620	547	0.27	3078	1.54
622	666	0.33	3744	1.87
624	696	0.35	4440	2.22
626	780	0.39	5220	2.61
628	793	0.40	6013	3.00
630	963	0.48	6976	3.49
632	1053	0.53	8029	4.01
633	1151	0.58	9180	4.59
635	1257	0.63	10437	5.21
636	1373	0.69	11810	5.90
637	1445	0.72	13255	6.62
639	1590	0.79	14845	7.42
640	1797	0.90	16642	8.32
641	1836	0.92	18478	9.23
642	2059	1.03	20537	10.26
643	2249	1.12	22786	11.39
644	2429	1.21	25215	12.60
645	2640	1.32	27855	13.92
646	2779	1.39	30634	15.31
648	3062	1.53	33696	16.84
649	3336	1.67	37032	18.50
650	3390	1.69	40422	20.20
651	3839	1.92	44261	22.12
652	4003	2.00	48264	24.12
653	4395	2.20	52659	26.31
654	4550	2.27	57209	28.58
655	5013	2.5	62222	31.09
656	5051	2.52	67273	33.61
658	5338	2.67	72611	36.28
659	5741	2.87	78352	39.15
660	6022	3.01	84374	42.16
661	6354	3.17	90728	45.33
663	6537	3.27	97265	48.60
664	6911	3.45	104176	52.05
666	7182	3.59	111358	55.64
667	7360	3.68	118718	59.32

(Continued on next page)

Table H5. Grade 7 ELA 2011 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
669	7644	3.82	126362	63.14
670	7932	3.96	134294	67.10
672	8080	4.04	142374	71.14
674	8006	4.00	150380	75.14
676	8316	4.16	158696	79.29
679	7891	3.94	166587	83.24
682	7681	3.84	174268	87.07
685	7165	3.58	181433	90.65
688	6336	3.17	187769	93.82
693	5198	2.60	192967	96.42
699	3806	1.90	196773	98.32
707	2269	1.13	199042	99.45
719	919	0.46	199961	99.91
790	179	0.09	200140	100.00

Table H6. Grade 8 ELA 2011 SS Frequency Distribution, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
430	147	0.07	147	0.07
544	86	0.04	233	0.12
569	131	0.07	364	0.18
580	224	0.11	588	0.29
587	277	0.14	865	0.43
592	369	0.18	1234	0.61
596	452	0.22	1686	0.84
600	492	0.24	2178	1.08
603	537	0.27	2715	1.35
606	653	0.32	3368	1.67
608	681	0.34	4049	2.01
610	810	0.40	4859	2.41
612	791	0.39	5650	2.81
614	873	0.43	6523	3.24
616	992	0.49	7515	3.73
618	1029	0.51	8544	4.24
619	1133	0.56	9677	4.81

(Continued on next page)

Table H6. Grade 8 ELA 2011 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
621	1145	0.57	10822	5.38
622	1249	0.62	12071	6.00
624	1438	0.71	13509	6.71
625	1527	0.76	15036	7.47
627	1671	0.83	16707	8.30
628	1826	0.91	18533	9.21
629	1977	0.98	20510	10.19
631	2092	1.04	22602	11.23
632	2315	1.15	24917	12.38
633	2498	1.24	27415	13.62
635	2791	1.39	30206	15.01
636	2819	1.40	33025	16.41
637	3092	1.54	36117	17.94
638	3343	1.66	39460	19.60
640	3505	1.74	42965	21.35
641	3843	1.91	46808	23.26
642	4010	1.99	50818	25.25
643	4265	2.12	55083	27.37
645	4521	2.25	59604	29.61
646	4762	2.37	64366	31.98
647	5269	2.62	69635	34.60
649	5384	2.67	75019	37.27
650	5686	2.82	80705	40.10
652	6012	2.99	86717	43.08
653	6330	3.14	93047	46.23
655	6712	3.33	99759	49.56
656	6863	3.41	106622	52.97
658	7277	3.62	113899	56.59
660	7470	3.71	121369	60.30
662	7702	3.83	129071	64.13
663	8053	4.00	137124	68.13
666	8076	4.01	145200	72.14
668	8013	3.98	153213	76.12
670	7997	3.97	161210	80.09
673	7620	3.79	168830	83.88
676	7454	3.70	176284	87.58

(Continued on next page)

Table H6. Grade 8 ELA 2011 SS Frequency Distribution, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
680	6807	3.38	183091	90.96
684	5931	2.95	189022	93.91
689	4987	2.48	194009	96.39
695	3656	1.82	197665	98.20
704	2249	1.12	199914	99.32
720	1072	0.53	200986	99.85
790	292	0.15	201278	100.00

References

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

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

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

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

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

Cattell, R.B. (1966). The Scree Test for the Number of Factors. *Multivariate Behavioral Research* 1:245–276.

CTB/McGraw-Hill (1996). *TerraNova™ Assessment Series (1st Ed.)*. Monterey, CA: CTB/McGraw-Hill.

CTB/McGraw-Hill (2000). *TerraNova™ Assessment Series (2nd Ed.)* Monterey, CA: CTB/McGraw-Hill.

CTB/McGraw-Hill (2006). *TerraNova™ Assessment Series (3rd Ed.)* Monterey, CA: CTB/McGraw-Hill.

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

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

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

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

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

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

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

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

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

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

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

Kolen, M.J. and R.L. Brennan (1995). *Test Equating: Methods and Practices.* New York: 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.

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