

# **New York State Testing Program 2008: English Language Arts, Grades 3–8**

**Technical Report**

**Submitted  
October 2008**

**CTB/McGraw-Hill  
Monterey, California 93940**

---

## Copyright

---

Developed and published under contract with the New York State Education Department by CTB/McGraw-Hill LLC, a subsidiary of The McGraw-Hill Companies, Inc., 20 Ryan Ranch Road, Monterey, California 93940-5703. Copyright © 2008 by the New York State Education Department. Permission is hereby granted for New York State school administrators and educators to reproduce these materials, located online at <http://www.emsc.nysed.gov/ciai/testing/pubs.html>, in the quantities necessary for their school's use, but not for sale, provided copyright notices are retained as they appear in these publications. This permission does not apply to distribution of these materials, electronically or by other means, other than for school use.

# Table of Contents

---

<b>SECTION I: INTRODUCTION AND OVERVIEW .....</b>	<b>1</b>
INTRODUCTION .....	1
TEST PURPOSE .....	1
TARGET POPULATION .....	1
TEST USE AND DECISIONS BASED ON ASSESSMENT .....	1
<i>Scale Scores</i> .....	1
<i>Proficiency Level Cut Scores and Classification</i> .....	2
<i>Standard Performance Index Scores</i> .....	2
TESTING ACCOMMODATIONS .....	2
TEST TRANSCRIPTIONS .....	2
TEST TRANSLATIONS .....	3
<b>SECTION II: TEST DESIGN AND DEVELOPMENT .....</b>	<b>4</b>
TEST DESCRIPTION .....	4
TEST CONFIGURATION .....	4
TEST BLUEPRINT .....	5
2008 ITEM MAPPING BY NEW YORK STATE STANDARDS .....	18
NEW YORK STATE EDUCATOR’S INVOLVEMENT IN TEST DEVELOPMENT .....	18
CONTENT RATIONALE .....	19
ITEM DEVELOPMENT .....	19
ITEM REVIEW .....	20
MATERIALS DEVELOPMENT .....	21
ITEM SELECTION AND TEST CREATION (CRITERIA AND PROCESS) .....	21
PROFICIENCY AND PERFORMANCE STANDARDS .....	22
<b>SECTION III: VALIDITY .....</b>	<b>23</b>
CONTENT VALIDITY .....	23
CONSTRUCT (INTERNAL STRUCTURE) VALIDITY .....	24
<i>Internal Consistency</i> .....	24
<i>Unidimensionality</i> .....	24
<i>Minimization of Bias</i> .....	26
<b>SECTION IV: TEST ADMINISTRATION AND SCORING .....</b>	<b>28</b>
TEST ADMINISTRATION .....	28
SCORING PROCEDURES OF OPERATIONAL TESTS .....	28
SCORING MODELS .....	28
SCORING OF CONSTRUCTED-RESPONSE ITEMS .....	29
SCORER QUALIFICATIONS AND TRAINING .....	30
QUALITY CONTROL PROCESS .....	30
<b>SECTION V: OPERATIONAL TEST DATA COLLECTION AND CLASSICAL ANALYSIS .....</b>	<b>31</b>
DATA COLLECTION .....	31
DATA PROCESSING .....	31
CLASSICAL ANALYSIS AND CALIBRATION SAMPLE CHARACTERISTICS .....	33
CLASSICAL DATA ANALYSIS .....	37
<i>Item Difficulty and Response Distribution</i> .....	37
<i>Point-Biserial Correlation Coefficients</i> .....	44
<i>Distractor Analysis</i> .....	44
<i>Test Statistics and Reliability Coefficients</i> .....	44
<i>Speededness</i> .....	45
<i>Differential Item Functioning</i> .....	45

<b>SECTION VI: IRT SCALING AND EQUATING .....</b>	<b>48</b>
IRT MODELS AND RATIONALE FOR USE.....	48
CALIBRATION SAMPLE .....	49
CALIBRATION PROCESS .....	49
ITEM-MODEL FIT.....	50
LOCAL INDEPENDENCE.....	51
SCALING AND EQUATING .....	52
<i>Anchor Item Security</i> .....	54
<i>Anchor Item Evaluation</i> .....	54
ITEM PARAMETERS.....	61
TEST CHARACTERISTIC CURVES.....	67
SCORING PROCEDURE.....	70
<i>Weighting Constructed-Response Items in Grades 4 and 8</i> .....	71
RAW SCORE-TO-SCALE SCORE AND SEM CONVERSION TABLES .....	71
STANDARD PERFORMANCE INDEX.....	78
IRT DIF STATISTICS.....	80
<b>SECTION VII: RELIABILITY AND STANDARD ERROR OF MEASUREMENT .....</b>	<b>83</b>
TEST RELIABILITY .....	83
<i>Reliability for Total Test</i> .....	83
<i>Reliability of MC Items</i> .....	84
<i>Reliability of CR Items</i> .....	84
<i>Test Reliability for NCLB Reporting Categories</i> .....	84
STANDARD ERROR OF MEASUREMENT .....	89
PERFORMANCE LEVEL CLASSIFICATION CONSISTENCY AND ACCURACY .....	89
<i>Consistency</i> .....	90
<i>Accuracy</i> .....	91
<b>SECTION VIII: SUMMARY OF OPERATIONAL TEST RESULTS .....</b>	<b>93</b>
SCALE SCORE DISTRIBUTION SUMMARY .....	93
<i>Grade 3</i> .....	93
<i>Grade 4</i> .....	94
<i>Grade 5</i> .....	95
<i>Grade 6</i> .....	96
<i>Grade 7</i> .....	97
<i>Grade 8</i> .....	98
PERFORMANCE LEVEL DISTRIBUTION SUMMARY.....	99
<i>Grade 3</i> .....	100
<i>Grade 4</i> .....	101
<i>Grade 5</i> .....	102
<i>Grade 6</i> .....	103
<i>Grade 7</i> .....	104
<i>Grade 8</i> .....	105
<b>SECTION IX: LONGITUDINAL COMPARISON OF RESULTS .....</b>	<b>107</b>
<b>APPENDIX A—ELA PASSAGE SPECIFICATIONS .....</b>	<b>109</b>
<b>APPENDIX A—ELA PASSAGE SPECIFICATIONS .....</b>	<b>109</b>
<b>APPENDIX B—CRITERIA FOR ITEM ACCEPTABILITY.....</b>	<b>115</b>
<b>APPENDIX C—PSYCHOMETRIC GUIDELINES FOR OPERATIONAL ITEM SELECTION .....</b>	<b>117</b>

**APPENDIX D—FACTOR ANALYSIS RESULTS..... 119**  
**APPENDIX E—ITEMS FLAGGED FOR DIF ..... 121**  
**APPENDIX F—ITEM-MODEL FIT STATISTICS ..... 123**  
**APPENDIX G—DERIVATION OF THE GENERALIZED SPI PROCEDURE.. 129**  
**APPENDIX H—DERIVATION OF CLASSIFICATION CONSISTENCY AND  
ACCURACY ..... 135**  
    CLASSIFICATION CONSISTENCY..... 135  
    CLASSIFICATION ACCURACY..... 136  
**APPENDIX I—SCALE SCORE FREQUENCY DISTRIBUTIONS ..... 137**  
**REFERENCES..... 145**

## List of Tables

---

TABLE 1. NYSTP ELA 2008 TEST CONFIGURATION.....	4
TABLE 2. NYSTP ELA 2008 CLUSTER ITEMS.....	5
TABLE 3. NYSTP ELA 2008 TEST BLUEPRINT .....	6
TABLE 4A. NYSTP ELA 2008 OPERATIONAL TEST MAP, GRADE 3 .....	7
TABLE 4B. NYSTP ELA 2008 OPERATIONAL TEST MAP, GRADE 4 .....	8
TABLE 4C. NYSTP ELA 2008 OPERATIONAL TEST MAP, GRADE 5 .....	10
TABLE 4D. NYSTP ELA 2008 OPERATIONAL TEST MAP, GRADE 6 .....	12
TABLE 4E. NYSTP ELA 2008 OPERATIONAL TEST MAP, GRADE 7 .....	13
TABLE 4F. NYSTP ELA 2008 OPERATIONAL TEST MAP, GRADE 8.....	16
TABLE 5. NYSTP ELA 2008 STANDARD COVERAGE .....	18
TABLE 6. FACTOR ANALYSIS RESULTS FOR ELA TESTS (TOTAL POPULATION).....	25
TABLE 7A. NYSTP ELA GRADE 3 DATA CLEANING .....	31
TABLE 7B. NYSTP ELA GRADE 4 DATA CLEANING.....	32
TABLE 7C. NYSTP ELA GRADE 5 DATA CLEANING .....	32
TABLE 7D. NYSTP ELA GRADE 6 DATA CLEANING .....	32
TABLE 7E. NYSTP ELA GRADE 7 DATA CLEANING.....	33
TABLE 7F. NYSTP ELA GRADE 8 DATA CLEANING.....	33
TABLE 8A. GRADE 3 SAMPLE CHARACTERISTICS (N = 193433) .....	34
TABLE 8B. GRADE 4 SAMPLE CHARACTERISTICS (N = 195029) .....	34
TABLE 8C. GRADE 5 SAMPLE CHARACTERISTICS (N = 195928) .....	35
TABLE 8D. GRADE 6 SAMPLE CHARACTERISTICS (N = 198628) .....	35
TABLE 8E. GRADE 7 SAMPLE CHARACTERISTICS (N = 198390) .....	36
TABLE 8F. GRADE 8 SAMPLE CHARACTERISTICS (N = 206153).....	36
TABLE 9A. P-VALUES, SCORED RESPONSE DISTRIBUTIONS, AND POINT BISERIALS, GRADE 3.....	38
TABLE 9B. P-VALUES, SCORED RESPONSE DISTRIBUTIONS, AND POINT BISERIALS, GRADE 4.....	39
TABLE 9C. P-VALUES, SCORED RESPONSE DISTRIBUTIONS, AND POINT BISERIALS, GRADE 5.....	40
TABLE 9D. P-VALUES, SCORED RESPONSE DISTRIBUTIONS, AND POINT BISERIALS, GRADE 6.....	41

<b>TABLE 9E. P-VALUES, SCORED RESPONSE DISTRIBUTIONS, AND POINT BISERIALS, GRADE 7.....</b>	<b>42</b>
<b>TABLE 9F. P-VALUES, SCORED RESPONSE DISTRIBUTIONS, AND POINT BISERIALS, GRADE 8.....</b>	<b>43</b>
<b>TABLE 10. NYSTP ELA 2008 TEST FORM STATISTICS AND RELIABILITY</b>	<b>45</b>
<b>TABLE 11. NYSTP ELA 2008 CLASSICAL DIF SAMPLE N-COUNTS .....</b>	<b>46</b>
<b>TABLE 12. NUMBER OF ITEMS FLAGGED BY SMD AND MANTEL-HAENZSEL DIF METHODS .....</b>	<b>47</b>
<b>TABLE 13. NYSTP ELA 2008 CALIBRATION RESULTS.....</b>	<b>50</b>
<b>TABLE 14. NYSTP ELA 2008 FINAL TRANSFORMATION CONSTANTS.....</b>	<b>54</b>
<b>TABLE 15. ELA ANCHOR EVALUATION SUMMARY.....</b>	<b>56</b>
<b>TABLE 16A. 2007 OPERATIONAL ITEM PARAMETER ESTIMATES, GRADE 3 .....</b>	<b>61</b>
<b>TABLE 16B. 2007 OPERATIONAL ITEM PARAMETER ESTIMATES, GRADE 4.....</b>	<b>62</b>
<b>TABLE 16C. 2007 OPERATIONAL ITEM PARAMETER ESTIMATES, GRADE 5 .....</b>	<b>63</b>
<b>TABLE 16D. 2007 OPERATIONAL ITEM PARAMETER ESTIMATES, GRADE 6.....</b>	<b>64</b>
<b>TABLE 16E. 2007 OPERATIONAL ITEM PARAMETER ESTIMATES, GRADE 7 .....</b>	<b>65</b>
<b>TABLE 16F. 2007 OPERATIONAL ITEM PARAMETER ESTIMATES, GRADE 8.....</b>	<b>66</b>
<b>TABLE 17. GRADE 3 RAW SCORE TO SCALE SCORE (WITH STANDARD ERROR).....</b>	<b>72</b>
<b>TABLE 18. GRADE 4 RAW SCORE TO SCALE SCORE (WITH STANDARD ERROR).....</b>	<b>73</b>
<b>TABLE 19. GRADE 5 RAW SCORE TO SCALE SCORE (WITH STANDARD ERROR).....</b>	<b>74</b>
<b>TABLE 20. GRADE 6 RAW SCORE TO SCALE SCORE (WITH STANDARD ERROR).....</b>	<b>75</b>
<b>TABLE 21. GRADE 7 RAW SCORE TO SCALE SCORE (WITH STANDARD ERROR).....</b>	<b>76</b>
<b>TABLE 22. GRADE 8 RAW SCORE TO SCALE SCORE (WITH STANDARD ERROR).....</b>	<b>77</b>
<b>TABLE 23. SPI TARGET RANGES .....</b>	<b>79</b>

<b>TABLE 24. NUMBER OF ITEMS FLAGGED FOR DIF BY THE LINN-HARNISCH METHOD.....</b>	<b>82</b>
<b>TABLE 25. ELA 3–8 TESTS RELIABILITY AND STANDARD ERROR OF MEASUREMENT.....</b>	<b>83</b>
<b>TABLE 26 RELIABILITY AND STANDARD ERROR OF MEASUREMENT—MC ITEMS ONLY .....</b>	<b>84</b>
<b>TABLE 27 RELIABILITY AND STANDARD ERROR OF MEASUREMENT—CR ITEMS ONLY .....</b>	<b>84</b>
<b>TABLE 28A. GRADE 3 TEST RELIABILITY BY SUBGROUP.....</b>	<b>85</b>
<b>TABLE 28B. GRADE 4 TEST RELIABILITY BY SUBGROUP .....</b>	<b>86</b>
<b>TABLE 28C. GRADE 5 TEST RELIABILITY BY SUBGROUP.....</b>	<b>86</b>
<b>TABLE 28D. GRADE 6 TEST RELIABILITY BY SUBGROUP.....</b>	<b>87</b>
<b>TABLE 28E. GRADE 7 TEST RELIABILITY BY SUBGROUP .....</b>	<b>88</b>
<b>TABLE 28F. GRADE 8 TEST RELIABILITY BY SUBGROUP .....</b>	<b>88</b>
<b>TABLE 29. DECISION CONSISTENCY (ALL CUTS).....</b>	<b>91</b>
<b>TABLE 30. DECISION CONSISTENCY (LEVEL III CUT).....</b>	<b>91</b>
<b>TABLE 31. DECISION AGREEMENT (ACCURACY) .....</b>	<b>92</b>
<b>TABLE 32. ELA GRADES 3–8 SCALE SCORE DISTRIBUTION SUMMARY ...</b>	<b>93</b>
<b>TABLE 33. SCALE SCORE DISTRIBUTION SUMMARY, BY SUBGROUP, GRADE 3 .....</b>	<b>94</b>
<b>TABLE 34. SCALE SCORE DISTRIBUTION SUMMARY, BY SUBGROUP, GRADE 4.....</b>	<b>95</b>
<b>TABLE 35. SCALE SCORE DISTRIBUTION SUMMARY, BY SUBGROUP, GRADE 5 .....</b>	<b>96</b>
<b>TABLE 36. SCALE SCORE DISTRIBUTION SUMMARY, BY SUBGROUP, GRADE 6.....</b>	<b>97</b>
<b>TABLE 37. SCALE SCORE DISTRIBUTION SUMMARY, BY SUBGROUP, GRADE 7 .....</b>	<b>98</b>
<b>TABLE 38. SCALE SCORE DISTRIBUTION SUMMARY, BY SUBGROUP, GRADE 8 .....</b>	<b>99</b>
<b>TABLE 39. ELA GRADES 3–8 PERFORMANCE LEVEL CUT SCORES.....</b>	<b>100</b>
<b>TABLE 40. ELA GRADES 3–8 TEST PERFORMANCE LEVEL DISTRIBUTIONS.....</b>	<b>100</b>
<b>TABLE 41. PERFORMANCE LEVEL DISTRIBUTION SUMMARY, BY SUBGROUP, GRADE 3.....</b>	<b>101</b>

<b>TABLE 42. PERFORMANCE LEVEL DISTRIBUTION SUMMARY, BY SUBGROUP, GRADE 4.....</b>	<b>102</b>
<b>TABLE 43. PERFORMANCE LEVEL DISTRIBUTION SUMMARY, BY SUBGROUP, GRADE 5.....</b>	<b>103</b>
<b>TABLE 44. PERFORMANCE LEVEL DISTRIBUTION SUMMARY, BY SUBGROUP, GRADE 6.....</b>	<b>104</b>
<b>TABLE 45. PERFORMANCE LEVEL DISTRIBUTION SUMMARY, BY SUBGROUP, GRADE 7.....</b>	<b>105</b>
<b>TABLE 46. PERFORMANCE LEVEL DISTRIBUTION SUMMARY, BY SUBGROUP, GRADE 8.....</b>	<b>106</b>
<b>TABLE 47. ELA GRADES 3–8 TEST LONGITUDINAL RESULTS.....</b>	<b>107</b>
<b>TABLE A1. READABILITY SUMMARY INFORMATION FOR 2008 OPERATIONAL TEST PASSAGES.....</b>	<b>110</b>
<b>TABLE A2. NUMBER, TYPE, AND LENGTH OF PASSAGES.....</b>	<b>113</b>
<b>TABLE D1. FACTOR ANALYSIS RESULTS FOR ELA TESTS (SELECTED SUBPOPULATIONS).....</b>	<b>119</b>
<b>TABLE E1. NYSTP ELA 2008 CLASSICAL DIF ITEM FLAGS .....</b>	<b>121</b>
<b>TABLE F1. ELA ITEM FIT STATISTICS, GRADE 3.....</b>	<b>123</b>
<b>TABLE F2. ELA ITEM FIT STATISTICS, GRADE 4.....</b>	<b>124</b>
<b>TABLE F3. ELA ITEM FIT STATISTICS, GRADE 5 .....</b>	<b>125</b>
<b>TABLE F4. ELA ITEM FIT STATISTICS, GRADE 6.....</b>	<b>126</b>
<b>TABLE F5. ELA ITEM FIT STATISTICS, GRADE 7 .....</b>	<b>127</b>
<b>TABLE F6. ELA ITEM FIT STATISTICS, GRADE 8.....</b>	<b>128</b>
<b>TABLE I1. GRADE 3 ELA 2008 SS FREQUENCY DISTRIBUTION, STATE ...</b>	<b>137</b>
<b>TABLE I2. GRADE 4 ELA 2008 SS FREQUENCY DISTRIBUTION, STATE ...</b>	<b>138</b>
<b>TABLE I3. GRADE 5 ELA 2008 SS FREQUENCY DISTRIBUTION, STATE ...</b>	<b>139</b>
<b>TABLE I4. GRADE 6 ELA 2008 SS FREQUENCY DISTRIBUTION, STATE ...</b>	<b>140</b>
<b>TABLE I5. GRADE 7 ELA 2008 SS FREQUENCY DISTRIBUTION, STATE ...</b>	<b>141</b>
<b>TABLE I6. GRADE 8 ELA 2008 SS FREQUENCY DISTRIBUTION, STATE ...</b>	<b>143</b>

## **Section I: Introduction and Overview**

---

### ***Introduction***

An overview of the New York State Testing Program (NYSTP), Grades 3–8, English Language Arts (ELA) 2008 Operational (OP) Tests is provided in this report. The report contains information about operational 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 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 Grades 3–8 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 2008, nonpublic schools participated in all grade tests but were not well represented in the testing program. Subsequently, the New York State Education Department (NYSED) made a decision to exclude these schools from the data analyses. Public school students were required to take all State assessments administered at their grade level, except for a very small percentage of students with disabilities who took the New York State Alternate Assessment (NYSAA) for students with severe disabilities. For more detail on this exemption, please refer to the *School Administrator’s Manual for Public and Nonpublic Schools* (SAM), available online at <http://www.emsc.nysed.gov/osa/sam/gr3-8ela-08.pdf>

### ***Test Use and Decisions Based on Assessment***

The Grades 3–8 ELA Tests are used to measure the extent to which individual students achieve the New York State Learning Standards in ELA and to determine whether schools, districts, and the State meet the required progress targets specified in the New York State accountability system. There are several types of scores available from the Grades 3–8 ELA Tests and 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.” Uses of Grades 3–8 ELA Tests scores include: determining student progress within schools and districts, supporting registration of schools and districts, determining eligibility of students for additional instruction time, and providing 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 (Not Meeting Learning Standards), Level II (Partially Meeting Learning Standards), Level III (Meeting Learning Standards), and Level IV (Meeting Learning Standards with Distinction). The proficiency cut scores used to distinguish among Levels I, II, III, and IV were established during the process of Standard Setting. There is reason to believe and evidence to support the claim that New York State ELA proficiency cut scores reflect the abilities intended by the New York State Education Department. Performance of students on the Grades 3–8 ELA Tests in relation to proficiency level cut scores is reported in a form of performance level classification. The performances of schools, districts, and the State are reported as percentages of students in each performance level. Detailed information on a process of establishing 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 and 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 questions onto scannable answer sheets; and the

teachers transcribe the responses to the constructed-response 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. Braille Publishers, Inc. produced the braille editions for the previous Grades 4 and 8 Tests.

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 student 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 January 2008 over a two-day (Grades 3, 5, 7, and 8) or three-day (Grades 4 and 6) 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 3, Grades 4, 6, and 8) and a Listening section (Book 2). Students in Grades 3, 5, and 7 also completed an Editing Paragraph (in Book 2). The 2008 *Teacher’s Directions* available online (<http://www.emsc.nysed.gov/osa/elaei/gr3-5tdc08.pdf> and <http://www.emsc.nysed.gov/osa/elaei/gr6-8tdc08.pdf>) as well as the 2008 *School Administrator’s Manual* (<http://www.emsc.nysed.gov/osa/sam/gr3-8ela-08.pdf>) provide details on security, scheduling, classroom organization and preparation, test materials, and administration.

**Table 1. NYSTP ELA 2008 Test Configuration**

Grade	Day	Book	Number of Items			Allotted Time ( minutes)	
			MC	CR*	Total**	Testing	Prep
3	1	1	20	1	21	40	10
	2	2	4	3	7	35	15
	Totals		24	4	28	75	25
4	1	1	28	0	28	45	10
	2	2	0	3	3	45	15
	3	3	0	4	4	60	10
	Totals		28	7	35	150	35
5	1	1	20	1	21	45	10
	2	2	4	2	6	30	15
	Totals		24	3	27	75	25
6	1	1	26	0	26	55	10
	2	2	0	4	4	45	15
	3	3	0	4	4	60	10
	Totals		26	8	34	160	35

(Continued on next page)

**Table 1. NYSTP ELA 2008 Test Configuration (cont.)**

Grade	Day	Book	Number of Items			Allotted Time ( minutes)	
			MC	CR*	Total**	Testing	Prep
7	1	1	26	2	28	55	10
	2	2	4	3	7	30	15
	Totals		30	5	35	85	25
8	1	1	26	0	26	55	10
	1	2	0	4	4	45	15
	2	3	0	4	4	60	10
	Totals		26	8	34	160	35

\*Does not reflect cluster-scoring. \*\* Reflects actual items in the test books.

In most cases, the test book item number is also the item number for the purposes of data analysis. The exception is that constructed-response items from Grades 4, 6, and 8 are cluster-scored. Table 2 lists the test book item numbers and the item numbers as scored. Because analyses are based on scored data, the latter item numbers will be referred to in this *Technical Report*.

**Table 2. NYSTP ELA 2008 Cluster Items**

Grade	Cluster Type	Contributing Book Items	Item Number for Data Analysis
4	Listening	29, 30, 31	29
4	Reading	32, 33, 34, 35	30
4	Writing Mechanics	31, 35	31
6	Listening	27, 28, 29, 30	27
6	Reading	31, 32, 33, 34	28
6	Writing Mechanics	30, 34	29
8	Listening	27, 28, 29, 30	27
8	Reading	31, 32, 33, 34	28
8	Writing Mechanics	30, 34	29

### ***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 standard at that grade and the emphasis placed on those performance indicators by the blueprint-specifications panel members. Table 3 shows the Grades 3–8 ELA Tests blueprint and actual number of score points in 2008 OP tests.

**Table 3. NYSTP ELA 2008 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	33	3	S1	10	10	33.0	33.0
			S2	14	15	47.0	50.0
			S3	6	5	20.0	17.0
4	39	3	S1	13	12	36.0	33.5
			S2	16	16	44.5	44.5
			S3	7	8	19.5	22.0
5	31	3	S1	12	13	43.0	46.0
			S2	10	10	36.0	36.0
			S3	6	5	21.0	18.0
6	39	3	S1	13	11	36.0	30.5
			S2	16	16	44.5	44.5
			S3	7	9	19.5	25.0
7	41	3	S1	15	17	39.0	45.0
			S2	15	13	39.0	34.0
			S3	8	8	22.0	21.0
8	39	3	S1	14	13	39.0	36.0
			S2	14	14	39.0	39.0
			S3	8	9	22.0	25.0

Tables 4a–4f present Grades 3–8 ELA Test item maps with the item type indicator, the maximum number of points obtainable from each item, the Learning Standard measured by each item, and the answer key.

**Table 4a. NYSTP ELA 2008 Operational Test Map, Grade 3**

Question	Type	Points	Standard	Performance Indicator	Answer Key
<b>Book 1</b>	<b>Reading</b>				
1	multiple choice	1	3	Evaluate the content by identifying important and unimportant details	B
2	multiple choice	1	2	Make predictions, draw conclusions, and make inferences about events and characters	A
3	multiple choice	1	2	Make predictions, draw conclusions, and make inferences about events and characters	B
4	multiple choice	1	2	Determine the meaning of unfamiliar words by using context clues, dictionaries, and other classroom resources	C
5	multiple choice	1	3	Evaluate the content by identifying whether events, actions, characters, and/or settings are realistic	D
6	multiple choice	1	1	Identify a conclusion that summarizes the main idea	B
7	multiple choice	1	1	Read unfamiliar texts to collect data, facts, and ideas	A
8	multiple choice	1	1	Read unfamiliar texts to collect data, facts, and ideas	B
9	multiple choice	1	1	Identify main ideas and supporting details in informational texts	A
10	multiple choice	1	1	Determine the meaning of unfamiliar words by using context clues, dictionaries, and other classroom resources	D
11	multiple choice	1	2	Summarize main ideas and supporting details from imaginative texts	C
12	multiple choice	1	2	Use graphic organizers to record significant details about characters and events in stories	A
13	multiple choice	1	2	Use knowledge of story structure, story elements, and key vocabulary to interpret stories	B
14	multiple choice	1	2	Use knowledge of story structure, story elements, and key vocabulary to interpret stories	A
15	multiple choice	1	3	Evaluate the content by identifying whether events, actions, characters, and/or settings are realistic	C
16	multiple choice	1	1	Read and understand written directions	B
17	multiple choice	1	1	Read and understand written directions	A
18	multiple choice	1	3	Evaluate the content by identifying important and unimportant details	D
19	multiple choice	1	3	Evaluate the content by identifying important and unimportant details	A

*(Continued on next page)*

**Table 4a. NYSTP ELA 2008 Operational Test Map, Grade 3 (cont.)**

Question	Type	Points	Standard	Performance Indicator	Answer Key
<b>Book 2</b>	<b>Listening and Writing</b>				
20	multiple choice	1	1	Determine the meaning of unfamiliar words by using context clues, dictionaries, and other classroom resources	B
21	short response	2	1	Use graphic organizers to record significant details from informational texts	n/a
22	multiple choice	1	2	Identify elements of character, plot, and setting to understand the author's message or intent	B
23	multiple choice	1	2	Identify elements of character, plot, and setting to understand the author's message or intent	D
24	multiple choice	1	2	Identify elements of character, plot, and setting to understand the author's message or intent	A
25	multiple choice	1	2	Identify elements of character, plot, and setting to understand the author's message or intent	C
26	short response	2	2	Identify elements of character, plot, and setting to understand the author's message or intent	n/a
27	short response	2	2	Identify elements of character, plot, and setting to understand the author's message or intent	n/a
28	editing paragraph	3	n/a	Use basic punctuation correctly; Capitalize words such as literary titles, holidays, and product names	n/a

**Table 4b. NYSTP ELA 2008 Operational Test Map, Grade 4**

Question	Type	Points	Standard	Performance Indicator	Answer Key
<b>Book 1</b>	<b>Reading</b>				
1	multiple choice	1	1	Locate information in a text that is needed to solve a problem	A
2	multiple choice	1	1	Use text features, such as captions, charts, tables, graphs, maps, notes, and other visuals, to understand and interpret informational texts	C
3	multiple choice	1	1	Use text features, such as captions, charts, tables, graphs, maps, notes, and other visuals, to understand and interpret informational texts	D
4	multiple choice	1	1	Use graphic organizers to record significant details from informational texts	D
5	multiple choice	1	1	Determine the meaning of unfamiliar words by using context clues, dictionaries, and other classroom resources	B
6	multiple choice	1	3	Evaluate the content by identifying the author's purpose	C

*(Continued on next page)*

**Table 4b. NYSTP ELA 2008 Operational Test Map, Grade 4 (cont.)**

Question	Type	Points	Standard	Performance Indicator	Answer Key
<b>Book 1</b>	<b>Reading</b>				
7	multiple choice	1	1	Collect and interpret data, facts, and ideas from unfamiliar texts	B
8	multiple choice	1	2	Use specific evidence from stories to identify themes; describe characters, their actions and motivations; relate a sequence of events	B
9	multiple choice	1	2	Make predictions, draw conclusions, and make inferences about events and characters	D
10	multiple choice	1	3	Evaluate the content by identifying whether events, actions, characters, and/or settings are realistic	C
11	multiple choice	1	2	Make predictions, draw conclusions, and make inferences about events and characters	C
12	multiple choice	1	2	Use knowledge of story structure, story elements, and key vocabulary to interpret stories	B
13	multiple choice	1	2	Use knowledge of story structure, story elements, and key vocabulary to interpret stories	C
14	multiple choice	1	2	Use specific evidence from stories to identify themes; describe characters, their actions and motivations; relate a sequence of events	D
15	multiple choice	1	2	Use graphic organizers to record significant details about characters and events in stories	A
16	multiple choice	1	2	Use graphic organizers to record significant details about characters and events in stories	D
17	multiple choice	1	1	Identify a main idea and supporting details in informational texts	C
18	multiple choice	1	1	Locate information in a text that is needed to solve a problem	A
19	multiple choice	1	1	Identify a main idea and supporting details in informational texts	D
20	multiple choice	1	1	Determine the meaning of unfamiliar words by using context clues, dictionaries, and other classroom resources	C
21	multiple choice	1	1	Collect and interpret data, facts, and ideas from unfamiliar texts	B
22	multiple choice	1	1	Identify a conclusion that summarizes the main idea	C
23	multiple choice	1	3	Evaluate the content by identifying the author's purpose	A
24	multiple choice	1	2	Use specific evidence from stories to identify themes; describe characters, their actions and motivations; relate a sequence of events	B
25	multiple choice	1	2	Use specific evidence from stories to identify themes; describe characters, their actions and motivations; relate a sequence of events	B

*(Continued on next page)*

**Table 4b. NYSTP ELA 2008 Operational Test Map, Grade 4 (cont.)**

Question	Type	Points	Standard	Performance Indicator	Answer Key
<b>Book 1</b>	<b>Reading</b>				
26	multiple choice	1	2	Use graphic organizers to record significant details about characters and events in stories	D
27	multiple choice	1	3	Evaluate the content by identifying important and unimportant details	A
28	multiple choice	1	2	Explain the difference between fact and fiction	C
<b>Book 2</b>	<b>Listening and Writing</b>				
29–31	short and extended response	4	2	Listening/Writing cluster	n/a
<b>Book 3</b>	<b>Reading and Writing</b>				
32–35	short and extended response	4	3	Reading/Writing cluster	n/a
<b>Book 2 &amp; Book 3</b>	<b>Writing Mechanics</b>				
31 & 35	extended response	3	n/a	Writing Mechanics cluster	n/a

**Table 4c. NYSTP ELA 2008 Operational Test Map, Grade 5**

Question	Type	Points	Standard	Performance Indicator	Answer Key
<b>Book 1</b>	<b>Reading</b>				
1	multiple choice	1	3	Evaluate information, ideas, opinions, and themes in texts by identifying a central idea and supporting details	B
2	multiple choice	1	1	Read to collect and interpret data, facts, and ideas from multiple sources	C
3	multiple choice	1	1	Determine the meaning of unfamiliar words by using context clues, a dictionary, or a glossary	C
4	multiple choice	1	1	Read to collect and interpret data, facts, and ideas from multiple sources	B
5	multiple choice	1	2	Identify literary elements, such as setting, plot, and character, of different genres	A
6	multiple choice	1	2	Identify literary elements, such as setting, plot, and character, of different genres	D

*(Continued on next page)*

**Table 4c. NYSTP ELA 2008 Operational Test Map, Grade 5 (cont.)**

Question	Type	Points	Standard	Performance Indicator	Answer Key
<b>Book 1</b>	<b>Reading</b>				
7	multiple choice	1	3	Evaluate information, ideas, opinions, and themes in texts by identifying a central idea and supporting details	B
8	multiple choice	1	2	Determine the meaning of unfamiliar words by using context clues, a dictionary, or a glossary	C
9	multiple choice	1	2	Identify literary elements, such as setting, plot, and character, of different genres	B
10	multiple choice	1	2	Define the characteristics of different genres	C
11	multiple choice	1	1	Read to collect and interpret data, facts, and ideas from multiple sources	A
12	multiple choice	1	1	Read to collect and interpret data, facts, and ideas from multiple sources	A
13	multiple choice	1	1	Distinguish between fact and opinion	D
14	multiple choice	1	3	Evaluate information, ideas, opinions, and themes in texts by identifying a central idea and supporting details	B
15	multiple choice	1	1	Determine the meaning of unfamiliar words by using context clues, a dictionary, or a glossary	C
16	multiple choice	1	2	Identify literary elements, such as setting, plot, and character, of different genres	C
17	multiple choice	1	2	Identify literary elements, such as setting, plot, and character, of different genres	A
18	multiple choice	1	2	Identify literary elements, such as setting, plot, and character, of different genres	B
19	multiple choice	1	2	Identify literary elements, such as setting, plot, and character, of different genres	B
20	multiple choice	1	2	Determine the meaning of unfamiliar words by using context clues, a dictionary, or a glossary	D
21	short response	2	3	Evaluate information, ideas, opinions, and themes in texts by identifying a central idea and supporting details	n/a
22	multiple choice	1	1	Identify essential details for note taking	A
23	multiple choice	1	1	Identify information that is implicit rather than stated	A
24	multiple choice	1	1	Identify essential details for note taking	C
25	multiple choice	1	1	Identify essential details for note taking	C
26	short response	2	1	Identify essential details for note taking	n/a
27	editing paragraph	3	n/a	Observe the rules of punctuation, capitalization, and spelling; use correct grammatical construction	n/a

**Table 4d. NYSTP ELA 2008 Operational Test Map, Grade 6**

Question	Type	Points	Standard	Performance Indicator	Answer Key
<b>Book 1</b>	<b>Reading</b>				
1	multiple choice	1	3	Evaluate information, ideas, opinions, and themes by identifying a central idea and supporting details	C
2	multiple choice	1	2	Identify literary elements (e.g., setting, plot, character, rhythm, and rhyme) of different genres	A
3	multiple choice	1	2	Read, view, and interpret literary text	B
4	multiple choice	1	2	Read, view, and interpret literary text	B
5	multiple choice	1	2	Recognize how the author uses literary devices, such as simile, metaphor, and personification, to create meaning	A
6	multiple choice	1	2	Identify the ways in which characters change and develop throughout a story	D
7	multiple choice	1	2	Define the characteristics of different genres	C
8	multiple choice	1	1	Read to collect and interpret data, facts, and ideas from multiple sources	C
9	multiple choice	1	1	Recognize organizational formats to assist in comprehension of informational texts	A
10	multiple choice	1	1	Identify information that is implied rather than stated	B
11	multiple choice	1	1	Compare and contrast information about one topic from multiple sources	A
12	multiple choice	1	3	Identify different perspectives (such as social, cultural, ethnic, historical) on an issue presented in one or more than one text	C
13	multiple choice	1	1	Recognize organizational formats to assist in comprehension of informational texts	D
14	multiple choice	1	1	Read to collect and interpret data, facts, and ideas from multiple sources	A
15	multiple choice	1	1	Compare and contrast information about one topic from multiple sources	A
16	multiple choice	1	3	Evaluate information, ideas, opinions, and themes by identifying a central idea and supporting details	C
17	multiple choice	1	2	Identify literary elements (e.g., setting, plot, character, rhythm, and rhyme) of different genres	B
18	multiple choice	1	2	Read, view, and interpret literary text	D
19	multiple choice	1	2	Identify literary elements (e.g., setting, plot, character, rhythm, and rhyme) of different genres	A
20	multiple choice	1	2	Identify the ways in which characters change and develop throughout a story	B

*(Continued on next page)*

**Table 4d. NYSTP ELA 2008 Operational Test Map, Grade 6 (cont.)**

Question	Type	Points	Standard	Performance Indicator	Answer Key
<b>Book 1</b>	<b>Reading</b>				
21	multiple choice	1	2	Determine the meaning of unfamiliar words by using context clues, a dictionary, or a glossary	D
22	multiple choice	1	1	Compare and contrast information about one topic from multiple sources	B
23	multiple choice	1	1	Read to collect and interpret data, facts, and ideas from multiple sources	A
24	multiple choice	1	1	Read to collect and interpret data, facts, and ideas from multiple sources	C
25	multiple choice	1	1	Distinguish between fact and opinion	A
26	multiple choice	1	3	Evaluate information, ideas, opinions, and themes by identifying a central idea and supporting details	C
<b>Book 2</b>	<b>Listening and Writing</b>				
27–30	short and extended response	5	2	Listening/Writing cluster	n/a
<b>Book 3</b>	<b>Reading and Writing</b>				
31–34	short and extended response	5	3	Reading/Writing cluster	n/a
<b>Book 2 &amp; Book 3</b>	<b>Writing Mechanics</b>				
30 & 34	extended response	3	n/a	Writing Mechanics cluster	n/a

**Table 4e. NYSTP ELA 2008 Operational Test Map, Grade 7**

Question	Type	Points	Standard	Performance Indicator	Answer Key
<b>Book 1</b>	<b>Reading</b>				
1	multiple choice	1	1	Interpret data, facts, and ideas from informational texts by applying thinking skills, such as define, classify, and infer	A
2	multiple choice	1	3	Evaluate the validity and accuracy of information, ideas, themes, opinions, and experiences in text to evaluate examples, details, or reasons used to support ideas	C
3	multiple choice	1	1	Make, confirm, or revise predictions	D

*(Continued on next page)*

**Table 4e. NYSTP ELA 2008 Operational Test Map, Grade 7 (cont.)**

Question	Type	Points	Standard	Performance Indicator	Answer Key
<b>Book 1</b>	<b>Reading</b>				
4	multiple choice	1	1	Use knowledge of structure, content, and vocabulary to understand informational text	D
5	multiple choice	1	2	Identify the author's point of view, such as first-person narrator and omniscient narrator	D
6	multiple choice	1	2	Interpret characters, plot, setting, and theme, using evidence from the text	C
7	multiple choice	1	2	Interpret characters, plot, setting, and theme, using evidence from the text	C
8	multiple choice	1	2	Interpret characters, plot, setting, and theme, using evidence from the text	A
9	multiple choice	1	3	Evaluate the validity and accuracy of information, ideas, themes, opinions, and experiences in text to evaluate examples, details, or reasons used to support ideas	B
10	multiple choice	1	2	Determine the meaning of unfamiliar words by using context clues, a dictionary, a glossary, and structural analysis (i.e., looking at roots, prefixes, and suffixes of words)	D
11	multiple choice	1	1	Determine the meaning of unfamiliar words by using context clues, a dictionary, a glossary, and structural analysis (i.e., looking at roots, prefixes, and suffixes of words)	B
12	multiple choice	1	1	Interpret data, facts, and ideas from informational texts by applying thinking skills, such as define, classify, and infer	C
13	multiple choice	1	3	Evaluate the validity and accuracy of information, ideas, themes, opinions, and experiences in text to evaluate examples, details, or reasons used to support ideas	A
14	multiple choice	1	1	Identify a purpose for reading	C
15	multiple choice	1	1	Determine the meaning of unfamiliar words by using context clues, a dictionary, a glossary, and structural analysis (i.e., looking at roots, prefixes, and suffixes of words)	D
16	multiple choice	1	1	Draw conclusions and make inferences on the basis of explicit and implied information	D
17	multiple choice	1	1	Condense, combine, or categorize new information from one or more sources	B
18	multiple choice	1	1	Draw conclusions and make inferences on the basis of explicit and implied information	B
19	multiple choice	1	1	Use knowledge of structure, content, and vocabulary to understand informational text	C

*(Continued on next page)*

**Table 4e. NYSTP ELA 2008 Operational Test Map, Grade 7 (cont.)**

Question	Type	Points	Standard	Performance Indicator	Answer Key
<b>Book 1</b>	<b>Reading</b>				
20	multiple choice	1	2	Interpret characters, plot, setting, and theme, using evidence from the text	C
21	multiple choice	1	2	Identify the author's point of view, such as first-person narrator and omniscient narrator	C
22	multiple choice	1	2	Determine how the use and meaning of literary devices (symbolism, metaphor and simile, alliteration, personification, flashback, and foreshadowing) convey the author's message or intent	D
23	multiple choice	1	3	Evaluate the validity and accuracy of information, ideas, themes, opinions, and experiences in text to evaluate examples, details, or reasons used to support ideas	D
24	multiple choice	1	2	Recognize how the author's use of language creates images or feelings	A
25	multiple choice	1	2	Recognize how the author's use of language creates images or feelings	B
26	multiple choice	1	2	Interpret characters, plot, setting, and theme, using evidence from the text	A
27	short response	2	2	Interpret characters, plot, setting, and theme, using evidence from the text	n/a
28	short response	2	3	Evaluate the validity and accuracy of information, ideas, themes, opinions, and experiences in text to evaluate examples, details, or reasons used to support ideas	n/a
<b>Book 2</b>	<b>Listening and Writing</b>				
29	multiple choice	1	1	Recall significant ideas and details, and describe relationships between and among them	B
30	multiple choice	1	1	Make, confirm, or revise predictions by distinguishing between relevant and irrelevant oral information	A
31	multiple choice	1	1	Recall significant ideas and details, and describe relationships between and among them	C
32	multiple choice	1	1	Recall significant ideas and details, and describe relationships between and among them	C
33	short response	2	1	Draw conclusions and make inferences on the basis of explicit information	n/a
34	short response	2	3	Form an opinion or judgment about the validity and accuracy of information, ideas, opinions, themes, and experiences	n/a
35	editing paragraph	3	n/a	Observe rules of punctuation, capitalization, and spelling; use correct grammatical construction	n/a

**Table 4f. NYSTP ELA 2008 Operational Test Map, Grade 8**

Question	Type	Points	Standard	Performance Indicator	Answer Key
<b>Book 1</b>	<b>Reading</b>				
1	multiple choice	1	2	Interpret characters, plot, setting, theme, and dialogue, using evidence from the text	B
2	multiple choice	1	2	Determine how the use and meaning of literary devices, such as symbolism, metaphor and simile, alliteration, personification, flashback, and foreshadowing, convey the author’s message or intent	A
3	multiple choice	1	2	Interpret characters, plot, setting, theme, and dialogue, using evidence from the text	C
4	multiple choice	1	2	Interpret characters, plot, setting, theme, and dialogue, using evidence from the text	A
5	multiple choice	1	2	Interpret characters, plot, setting, theme, and dialogue, using evidence from the text	D
6	multiple choice	1	1	Condense, combine, or categorize new information from one or more sources	C
7	multiple choice	1	1	Draw conclusions and make inferences on the basis of explicit and implied information	A
8	multiple choice	1	3	Evaluate the validity and accuracy of information, ideas, themes, opinions, and experiences in texts to evaluate examples, details, or reasons used to support ideas	C
9	multiple choice	1	1	Apply thinking skills, such as define, classify, and infer, to interpret data, facts, and ideas from informational texts	B
10	multiple choice	1	1	Determine the meaning of unfamiliar words by using context clues, a dictionary, a glossary, and structural analysis (i.e., looking at roots, prefixes, and suffixes of words)	A
11	multiple choice	1	2	Identify the author’s point of view, such as first-person narrator and omniscient narrator	A
12	multiple choice	1	2	Recognize how the author’s use of language creates images or feelings	B
13	multiple choice	1	2	Recognize how the author’s use of language creates images or feelings	B
14	multiple choice	1	2	Determine how the use and meaning of literary devices, such as symbolism, metaphor and simile, alliteration, personification, flashback, and foreshadowing, convey the author’s message or intent	A
15	multiple choice	1	2	Identify poetic elements, such as repetition, rhythm, and rhyming patterns, in order to interpret poetry	C
16	multiple choice	1	1	Apply thinking skills, such as define, classify, and infer, to interpret data, facts, and ideas from informational texts	D

*(Continued on next page)*

**Table 4f. NYSTP ELA 2008 Operational Test Map, Grade 8 (cont.)**

Question	Type	Points	Standard	Performance Indicator	Answer Key
<b>Book 1</b>	<b>Reading</b>				
17	multiple choice	1	1	Use knowledge of structure, content, and vocabulary to understand informational text	C
18	multiple choice	1	3	Evaluate the validity and accuracy of information, ideas, themes, opinions, and experiences in texts to evaluate examples, details, or reasons used to support ideas	B
19	multiple choice	1	1	Use knowledge of structure, content, and vocabulary to understand informational text	A
20	multiple choice	1	1	Determine the meaning of unfamiliar words by using context clues, a dictionary, a glossary, and structural analysis (i.e., looking at roots, prefixes, and suffixes of words)	C
21	multiple choice	1	2	Identify the author's point of view, such as first-person narrator and omniscient narrator	C
22	multiple choice	1	2	Interpret characters, plot, setting, theme, and dialogue, using evidence from the text	C
23	multiple choice	1	2	Interpret characters, plot, setting, theme, and dialogue, using evidence from the text	D
24	multiple choice	1	2	Interpret characters, plot, setting, theme, and dialogue, using evidence from the text	A
25	multiple choice	1	3	Evaluate the validity and accuracy of information, ideas, themes, opinions, and experiences in texts to identify cultural and ethnic values and their impact on content	A
26	multiple choice	1	3	Evaluate the validity and accuracy of information, ideas, themes, opinions, and experiences in texts to evaluate examples, details, or reasons used to support ideas	D
<b>Book 2</b>	<b>Listening and Writing</b>				
27–30	short and extended response	5	1	Listening/Writing cluster	n/a
<b>Book 3</b>	<b>Reading and Writing</b>				
31–34	short and extended response	5	3	Reading/Writing cluster	n/a
<b>Book 2 &amp; Book 3</b>	<b>Writing Mechanics</b>				
30 & 34	extended response	3	n/a	Writing Mechanics cluster	n/a

## 2008 Item Mapping by New York State Standards

**Table 5. NYSTP ELA 2008 Standard Coverage**

Grade	Standard	MC Item #s	CR Item #s	Total Items	Total Points
3	S1	6, 7, 8, 9, 10, 16, 17, 20	21	9	10
3	S2	2, 3, 4, 11, 12, 13, 14, 22, 23, 24, 25	26, 27	13	15
3	S3	1, 5, 15, 18, 19	n/a	5	5
4	S1	1, 2, 3, 4, 5, 7, 17, 18, 19, 20, 21, 22	n/a	12	12
4	S2	8, 9, 11, 12, 13, 14, 15, 16, 24, 25, 26, 28	29	13	16
4	S3	6, 10, 23, 27	31	5	8
5	S1	2, 3, 4, 11, 12, 13, 15, 22, 23, 24, 25	26	12	13
5	S2	5, 6, 8, 9, 10, 16, 17, 18, 19, 20	n/a	10	10
5	S3	1, 7, 14	21	4	5
6	S1	8, 9, 10, 11, 13, 14, 15, 22, 23, 24, 25	n/a	11	11
6	S2	2, 3, 4, 5, 6, 7, 17, 18, 19, 20, 21	27	12	16
6	S3	1, 12, 16, 26	29	5	9
7	S1	1, 3, 4, 11, 12, 14, 15, 16, 17, 18, 19, 29, 30, 31, 32	33	16	17
7	S2	5, 6, 7, 8, 10, 20, 21, 22, 24, 25, 26	27	12	13
7	S3	2, 9, 13, 23	28, 34	6	8
8	S1	6, 7, 9, 10, 16, 17, 19, 20	27	9	13
8	S2	1, 2, 3, 4, 5, 11, 12, 13, 14, 15, 21, 22, 23, 24	n/a	14	14
8	S3	8, 18, 25, 26	29	5	9

### *New York State Educator's Involvement in Test Development*

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

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

The selected participants must be certified and have both teaching and testing experience. The majority of 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 constructed-response items than others.)
- how much emphasis to place on each assessable performance indicator (For example, some performance indicators encompass a wider range of skills than others, necessitating a broader range of items in order to fully assess the performance indicator.)
- how the limitations, if any, were to be applied to the assessable performance indicators (For example, some portions of a performance indicator may be more appropriately assessed in the classroom than on a paper-and-pencil test.)
- what general examples of items could be used
- what the test blueprint was to be for each grade

The committees were composed of teachers from around the state 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 2008 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, in March 2007. 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 lead-content 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 New York State Testing Program and in the 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 10 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 (constructed-response items)
- the appropriateness of the correct response and distractors (multiple-choice 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 forms and submitted the field test forms to NYSED for their review and approval. The Grades 3–8 ELA Field Tests were administered to students across New York State during the week of January 22–26, 2007, using the State Sampling Matrix to ensure appropriate sampling of students. In addition, CTB/McGraw-Hill, in conjunction with NYSED test specialists, developed a field test *Teacher’s Directions and School Administrator’s Manual* to help ensure that the field tests were administered in a uniform manner to all participating students. Field test forms were assigned to participants at the school (grade) level while balancing the demographic statistics across forms, in order to proactively sample the students.

After administration of the field tests, rangefinding meetings were conducted in March 2007 in New York State to examine a sampling of the short- and extended-student responses. Committees of New York State educators with content and grade-level expertise were again assembled. Each committee was composed of approximately eight to ten participants per grade level. CTB/McGraw-Hill staff facilitated the meetings, and NYSED staff reviewed the decisions made by the committees and verified that the decisions made were consistent across grades. The committees’ charge was to select student responses that exemplified each score point of each constructed-response item. These responses, in conjunction with the rubrics, were then used by CTB/McGraw-Hill scoring staff to score the constructed response field test items.

## ***Item Selection and Test Creation (Criteria and Process)***

The third year of operational NYSTP Grades 3–8 ELA Tests were administered in January 2008. The test items were selected from the pool of items primarily field-tested in 2006 and 2007, using the data from those field tests. CTB/McGraw-Hill made preliminary selections for each grade. The selections were reviewed for alignment with the test design, blueprint, and the research guidelines for item selection (Appendix C). Item selection for the NYSTP 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 field-test item pool.

Item selection for the operational 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).

NYSED staff (including their content and research representative experts) traveled to CTB/McGraw-Hill in Monterey in July 2007 to finalize item selection and test creation. There, they discussed the content and data of the proposed selections, explored alternate selections for consideration, determined the final item selections, and ordered those items (assigned positions) in the operational test books. The final test forms were approved by the final eyes committee that consisted of approximately 20 participants across all grade levels. After the approval by NYSED, the tests were produced and administered in January 2008.

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 sites:

- <http://www.emsc.nysed.gov/osa/sam/gr3-8ela-08.pdf>
- <http://www.nysedregents.org/testing/elaei/08exams/home.htm>

### ***Proficiency and Performance Standards***

Proficiency cut score recommendations and the drafting of performance standards occurred at the NYSTP ELA standard setting review held in Albany in June 2006. The results were reviewed by a measurement review committee and were approved in August 2006. For each grade level, there are four proficiency levels. Three cut points demarcate the performance standards needed to demonstrate each ascending level of proficiency. For details on standard setting method, participants, achievement levels, and results (impact), refer to the *Bookmark Standard Setting Technical Report 2006 for Grades 3, 4, 5, 6, 7, and 8 English Language Arts* and *New York State ELA Measurement Review Technical Report 2006 for English Language Arts*.

## **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 the NYSTP, the content is defined by detailed, written specifications and blueprints that describe New York State content standards and define the skills that must be measured to assess these content standards (see Tables 3–5 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 field tests for their alignment with test blueprint. Educators also participated in a process of establishing scoring rubrics (during rangefinding meetings) for constructed-response items. Section II, "Test Design and Development," contains more information specific to the item review process. An independent study of alignment between the New York State curriculum and

the New York State Grades 3–8 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 New York State Grades 3–8 ELA Tests is supported by several types of evidence that can be obtained from the ELA test data.

#### **Internal Consistency**

Empirical studies of the internal structure of the test provide one type of evidence of construct validity. For example, high internal consistency constitutes evidence of validity. This is because high coefficients imply that the test questions are measuring the same domain of skill and are reliable and consistent. Reliability coefficients of the tests for total populations and subgroups of students are presented in Section VII, “Reliability and Standard Error of Measurement.” For the total population, the reliability coefficients (Cronbach’s alpha) ranged from 0.84–0.90, and for most subgroups the reliability coefficient was equal or greater than 0.80 (the exception was for Grade 5 students from districts classified as low need). 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 Q1 statistics (Yen, 1981) and the results are described in detail in Section VI, “IRT Scaling and Equating.” It was found that except for item 9 in Grade 3 test, item 21 in Grade 5 test, and item 4 in Grade 7 test, all other items on the 2008 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 of the grades. In addition, 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 principle components accounting for more than 10 percent 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 6.

**Table 6. Factor Analysis Results for ELA Tests (Total Population)**

Grade	Initial Eigenvalues			
	Component	Total	% of Variance	Cumulative %
3	<b>1</b>	<b>6.29</b>	<b>22.47</b>	<b>22.47</b>
	2	1.28	4.58	27.04
	3	1.07	3.83	30.87
4	<b>1</b>	<b>7.79</b>	<b>25.11</b>	<b>25.11</b>
	2	1.30	4.21	29.32
5	<b>1</b>	<b>5.54</b>	<b>20.50</b>	<b>20.50</b>
	2	1.12	4.16	24.66
	3	1.03	3.81	28.47
	4	1.00	3.71	32.18
6	<b>1</b>	<b>7.20</b>	<b>24.83</b>	<b>24.83</b>
	2	1.17	4.03	28.87
	3	1.05	3.64	32.50

(Continued on next page)

**Table 6. Factor Analysis Results for ELA Tests (Total Population) (cont.)**

Grade	Initial Eigenvalues			
	Component	Total	% of Variance	Cumulative %
7	<b>1</b>	<b>6.99</b>	<b>19.98</b>	<b>19.98</b>
	2	1.17	3.36	23.33
	3	1.16	3.32	26.65
	4	1.05	3.01	29.66
	5	1.02	2.93	32.59
8	<b>1</b>	<b>6.42</b>	<b>22.14</b>	<b>22.14</b>
	2	1.17	4.03	26.18

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 an 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: limited English proficiency (LEP) students, 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 LEP, SWD and SUA classifications are provided in Table D1 of Appendix D.

### **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 field test materials was reviewed by at least these same people.

In the third procedure, New York State educators who reviewed all field test 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 field test stage were closely examined for content bias and avoided during the operational test construction, DIF analyses were conducted again on operational 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 operational test item selection. Only those items deemed free of bias were included in the operational 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* (2008) located at <http://www.emsc.nysed.gov/3-8/archived.htm#scoring>.

### ***Test Administration***

NYSTP Grades 3–8 ELA Tests were administered at the classroom level during January 2008. The testing window for Grades 3, 4, and 5 was January 7–11. The testing window for Grades 6, 7, and 8 was January 14–18. The makeup test administration window for Grades 3, 4, and 5 was January 14–18, and for Grades 6, 7, and 8, it was January 22–25. 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 operational 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 operational scoring. A district ELA leader trained 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 2007–08 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 constructed-response items was based primarily on the scoring guides, which were created by CTB/McGraw-Hill handscoring and content development specialists with guidance from NYSED and New York State teachers during rangefinding sessions conducted after each field test. The CTB ELA handscoring team was composed of six supervisors, each representing one grade. Supervisors are selected on the basis of their handscoring experiences along with their educational and professional backgrounds.

In March 2007, CTB/McGraw-Hill staff met with groups of teachers from across the state in rangefinding sessions. Sets of actual student responses from the field tests were reviewed and discussed openly, and consensus scores were agreed upon by the teachers based on the teaching methods and criteria across the state, as well as on NYSED policies. In addition, a DVD was created to further explain each section of the scoring guides. Trainers used these materials to train scoring committee members on the criteria for scoring constructed-response items. Scoring Leader Handbooks were also distributed to outline the responsibilities of the scoring roles. CTB/McGraw-Hill handscoring staff also conducted training sessions in New York City to better equip these teachers and administrators with enhanced knowledge of scoring principles and criteria.

Scoring was conducted with pen and pencil scoring as opposed to electronic scoring, and each scoring committee member evaluated actual student papers instead of electronically scanned papers. All scoring committee members were trained by previously trained and approved trainers along with guidance from scoring guides, the ELA Frequently Asked Questions (FAQs) document, and a DVD 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 operational test was conducted by qualified administrators and teachers. Trainers used the scoring guides and DVDs to train scoring committee members on the criteria for scoring constructed-response 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 test 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, ELA FAQs, and DVD, they called the New York State ELA Helpline. This call center was established to help teachers and administrators during operational scoring. The helpline staff consisted of trained CTB/McGraw-Hill handscoring personnel who answered questions by phone, fax, or email. When a member of the staff was unable to resolve an issue, they deferred to NYSED for a scoring decision. A quality check was also performed on each completed box of scored tests to certify that all questions were scored and that the scoring committee members darkened each score on the answer document appropriately. To affirm that all schools across the state adhered to scoring guidelines and policies, approximately five percent of the schools’ results are audited each year by an outside vendor.

## Section V: Operational Test Data Collection and Classical Analysis

---

### *Data Collection*

Operational 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 late March and early April 2008. These data were used for all data analysis. Phase 2 involved submitting “straggler files” to CTB/McGraw-Hill in mid-April 2008. 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 7a–7f.

**Table 7a. NYSTP ELA Grade 3 Data Cleaning**

Exclusion Rule	# Deleted	# Cases Remain
Initial # of cases		195850
Out of grade	152	195698
No grade	6	195692
Duplicate record	0	195692
Non-public and out-of-district schools	2258	193434
Missing values for ALL items on OP form	1	193433
Out-of-range CR scores	0	193433

**Table 7b. NYSTP ELA Grade 4 Data Cleaning**

Exclusion Rule	# Deleted	# Cases Remain
Initial # of cases		208071
Out of grade	221	207850
No grade	12	207838
Duplicate record	0	207838
Non-public and out-of-district schools	12802	195036
Missing values for ALL items on OP form	7	195029
Out-of-range CR scores	0	195029

**Table 7c. NYSTP ELA Grade 5 Data Cleaning**

Exclusion Rule	# Deleted	# Cases Remain
Initial # of cases		198744
Out of grade	209	198535
No grade	4	198531
Duplicate record	0	198531
Non-public and out-of-district schools	2601	195930
Missing values for ALL items on OP form	2	195928
Out-of-range CR scores	0	195928

**Table 7d. NYSTP ELA Grade 6 Data Cleaning**

Exclusion Rule	# Deleted	# Cases Remain
Initial # of cases		208464
Out of grade	286	208178
No grade	47	208131
Duplicate record	0	208131
Non-public and out-of-district schools	9499	198632
Missing values for ALL items on OP form	4	198628
Out-of-range CR scores	0	198628

**Table 7e. NYSTP ELA Grade 7 Data Cleaning**

Exclusion Rule	# Deleted	# Cases Remain
Initial # of cases		201604
Out of grade	318	201286
No grade	38	201248
Duplicate record	0	201248
Non-public and out-of-district schools	2858	198390
Missing values for ALL items on OP form	0	198390
Out-of-range CR scores	0	198390

**Table 7f. NYSTP ELA Grade 8 Data Cleaning**

Exclusion Rule	# Deleted	# Cases Remain
Initial # of cases		227034
Out of grade	368	226666
No grade	70	226596
Duplicate record	0	226596
Non-public and out-of-district schools	20434	206162
Missing values for ALL items on OP form	9	206153
Out-of-range CR scores	0	206153

***Classical Analysis and Calibration Sample Characteristics***

The demographic characteristics of students in the cleaned calibration and equating datasets 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 this section and calibrations. 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 8a. Grade 3 Sample Characteristics (N = 193433)**

Demographic Category		N-count	% of Total N-count
NRC	NYC	68659	35.49
	Big cities	8052	4.16
	Urban-suburban	15046	7.78
	Rural	11417	5.90
	Average needs	57933	29.95
	Low needs	29362	15.18
	Charter	2964	1.53
Ethnicity	Asian	14256	7.37
	Black	36703	18.97
	Hispanic	40915	21.15
	American Indian	970	0.50
	Multi-Racial	229	0.12
	White	100295	51.85
	Unknown	65	0.03
LEP	No	177036	91.52
	Yes	16397	8.48
SWD	No	168045	86.88
	Yes	25388	13.12
SUA	No	155245	80.26
	Yes	38188	19.74

**Table 8b. Grade 4 Sample Characteristics (N = 195029)**

Demographic Category		N-count	% of Total N-count
NRC	NYC	69109	35.44
	Big cities	7663	3.93
	Urban-suburban	14841	7.61
	Rural	11426	5.86
	Average needs	59379	30.45
	Low needs	30254	15.51
	Charter	2357	1.21
Ethnicity	Asian	14136	7.25
	Black	37119	19.03
	Hispanic	40877	20.96
	American Indian	943	0.48
	Multi-Racial	184	0.09
	White	101684	52.14
	Unknown	86	0.04
LEP	No	181108	92.86
	Yes	13921	7.14
SWD	No	166916	85.59
	Yes	28113	14.41
SUA	No	154847	79.40
	Yes	40182	20.60

**Table 8c. Grade 5 Sample Characteristics (N = 195928)**

Demographic Category		N-count	% of Total N-count
NRC	NYC	68502	34.96
	Big cities	7493	3.82
	Urban-suburban	14480	7.39
	Rural	11323	5.78
	Average needs	60332	30.79
	Low needs	30535	15.58
	Charter	3263	1.67
Ethnicity	Asian	14503	7.40
	Black	37494	19.14
	Hispanic	40285	20.56
	American Indian	897	0.46
	Multi-Racial	164	0.08
	White	102510	52.32
	Unknown	75	0.04
LEP	No	184632	94.23
	Yes	11296	5.77
SWD	No	167059	85.27
	Yes	28869	14.73
SUA	No	156759	80.01
	Yes	39169	19.99

**Table 8d. Grade 6 Sample Characteristics (N = 198628)**

Demographic Category		N-count	% of Total N-count
NRC	NYC	68345	34.41
	Big cities	7490	3.77
	Urban-suburban	15249	7.68
	Rural	11760	5.92
	Average needs	61553	30.99
	Low needs	31518	15.87
	Charter	2713	1.37
Ethnicity	Asian	14481	7.29
	Black	37567	18.91
	Hispanic	40299	20.29
	American Indian	904	0.46
	Multi-Racial	156	0.08
	White	105140	52.93
	Unknown	81	0.04
LEP	No	188842	95.07
	Yes	9786	4.93
SWD	No	169031	85.10
	Yes	29597	14.90
SUA	No	161063	81.09
	Yes	37565	18.91

**Table 8e. Grade 7 Sample Characteristics (N = 198390)**

Demographic Category		N-count	% of Total N-count
NRC	NYC	70266	35.42
	Big cities	7785	3.92
	Urban-suburban	15688	7.91
	Rural	12430	6.27
	Average needs	60195	30.34
	Low needs	29742	14.99
	Charter	2284	1.15
Ethnicity	Asian	14227	7.17
	Black	38932	19.62
	Hispanic	40830	20.58
	American Indian	988	0.50
	Multi-Racial	104	0.05
	White	103245	52.04
	Unknown	64	0.03
LEP	No	189226	95.38
	Yes	9164	4.62
SWD	No	169830	85.60
	Yes	28560	14.40
SUA	No	162268	81.79
	Yes	36122	18.21

**Table 8f. Grade 8 Sample Characteristics (N = 206153)**

Demographic Category		N-count	% of Total N-count
NRC	NYC	70263	34.08
	Big cities	8025	3.89
	Urban-suburban	15945	7.73
	Rural	13067	6.34
	Average needs	65687	31.86
	Low needs	31747	15.40
	Charter	1419	0.69
Ethnicity	Asian	14227	6.90
	Black	39658	19.24
	Hispanic	40525	19.66
	American Indian	1036	0.50
	Multi-Racial	110	0.05
	White	110545	53.62
	Unknown	52	0.03
LEP	No	197593	95.85
	Yes	8560	4.15
SWD	No	177233	85.97
	Yes	28920	14.03
SUA	No	169511	82.23
	Yes	36642	17.77

## ***Classical Data Analysis***

Classical data analysis of the 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 response patterns, item difficulty (p-value), and item-test correlation (point biserial) is examined thoroughly. If any serious error were to occur with an item (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 (Table 9a–9f) illustrate student test performance, as observed from both MC and CR item responses. Omit rates signify the percentage of students who did not attempt the item. For MC items, “% at 0” represents the percentage of students who double-bubbled responses, and other “% SEL” categories represent the percentage of students who selected each answer response (without double marking). Proportions of students who selected the correct answer option are denoted with an asterisk (\*) and are repeated in the p-value field. For CR items, the “% at 0,” “% SEL,” and “% at 5” (only in Grades 6 and 8) categories depict the percentage of students who earned each valid score on the item, from zero to the maximum score.

Item difficulty is classically measured by the p-value statistic. It assesses the proportion of students who responded correctly to each MC item or the average percentage 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 0.31 to 0.99. For Grade 3, the item p-values were between 0.52 and 0.97 with a mean of 0.79. For Grade 4, the item p-values were between 0.58 and 0.94 with a mean of 0.73. For Grade 5, the item p-values were between 0.44 and 0.95 with a mean of 0.77. For Grade 6, the item p-values were between 0.51 and 0.95 with a mean of 0.78. For Grade 7, the item p-values were between 0.31 and 0.99 with a mean of 0.76. For Grade 8, the item p-values were between 0.44 and 0.93 with a mean of 0.77. These mean p-value statistics are also provided in Tables 9a-9f, along with other classical test summary statistics.

**Table 9a. P-values, Scored Response Distributions, and Point Biseriails, Grade 3**

Item	N-count	P-value	% Omit	% at 0	% Sel Option 1	% Sel Option 2	% Sel Option 3	% Sel Option 4	Pbis Option 1	Pbis Option 2	Pbis Option 3	Pbis Option 4	Pbis Key
1	193433	0.87	0.05	0.04	4.38	*87.25	4.17	4.10	-0.26	*0.43	-0.28	-0.16	0.43
2	193433	0.86	0.10	0.11	*86.42	6.94	3.22	3.21	*0.32	-0.14	-0.24	-0.15	0.32
3	193433	0.56	0.20	0.11	14.52	*56.18	19.43	9.56	-0.16	*0.25	0.01	-0.23	0.25
4	193433	0.86	0.30	0.10	4.63	5.51	*85.78	3.68	-0.25	-0.26	*0.47	-0.25	0.47
5	193433	0.86	0.40	0.27	4.49	5.55	3.08	*86.22	-0.31	-0.24	-0.26	*0.5	0.50
6	193433	0.65	0.18	0.11	8.76	*65.44	11.73	13.77	-0.17	*0.47	-0.16	-0.34	0.47
7	193433	0.67	0.20	0.18	*67.28	8.93	1.38	22.03	*0.32	-0.26	-0.21	-0.11	0.32
8	193433	0.72	0.30	0.08	21.58	*71.85	2.58	3.60	-0.28	*0.45	-0.24	-0.22	0.45
9	193433	0.78	0.35	0.12	*77.75	13.28	5.17	3.33	*0.3	-0.02	-0.28	-0.28	0.30
10	193433	0.81	0.46	0.13	5.60	6.95	5.58	*81.28	-0.29	-0.21	-0.22	*0.46	0.46
11	193433	0.81	0.11	0.05	11.54	4.10	*81.28	2.93	-0.41	-0.24	*0.55	-0.20	0.55
12	193433	0.88	0.10	0.04	*87.73	7.00	2.57	2.55	*0.48	-0.31	-0.25	-0.23	0.48
13	193433	0.69	0.15	0.07	15.39	*69.14	4.81	10.44	-0.28	*0.43	-0.23	-0.14	0.43
14	193433	0.94	0.19	0.11	*93.54	1.87	2.86	1.43	*0.47	-0.26	-0.28	-0.23	0.47
15	193433	0.78	0.22	0.11	6.82	6.62	*77.87	8.35	-0.25	-0.23	*0.49	-0.28	0.49
16	193433	0.69	0.23	0.08	17.81	*68.58	5.61	7.70	-0.15	*0.45	-0.23	-0.35	0.45
17	193433	0.83	0.28	0.13	*82.64	5.11	5.19	6.66	*0.48	-0.29	-0.17	-0.29	0.48
18	193433	0.52	0.37	0.19	14.12	9.71	23.51	*52.10	-0.15	-0.14	-0.08	*0.27	0.27
19	193433	0.72	0.47	0.13	*71.71	6.89	9.02	11.78	*0.43	-0.22	-0.24	-0.19	0.43
20	193433	0.67	0.97	0.04	4.82	*67.07	10.06	17.05	-0.28	*0.42	-0.25	-0.14	0.42
21	193433	0.93	0.94	3.86	5.32	89.88							
22	193433	0.91	0.07	0.02	4.01	*90.52	2.65	2.72	-0.22	*0.31	-0.09	-0.20	0.31
23	193433	0.95	0.08	0.10	1.06	1.31	2.89	*94.56	-0.19	-0.22	-0.26	*0.4	0.40
24	193433	0.97	0.11	0.08	*97.47	0.80	0.94	0.61	*0.3	-0.15	-0.20	-0.15	0.30
25	193433	0.87	0.18	0.01	5.87	4.30	*87.07	2.56	-0.14	-0.19	*0.31	-0.20	0.31
26	193433	0.80	0.22	6.57	27.16	66.04							
27	193433	0.53	0.47	33.79	24.95	40.79							
28	193433	0.88	0.27	5.46	4.09	10.46	79.72						

**Table 9b. P-values, Scored Response Distributions, and Point Biserials, Grade 4**

Item	N-count	P-value	% Omit	% at 0	% Sel Option 1	% Sel Option 2	% Sel Option 3	% Sel Option 4	Pbis Option 1	Pbis Option 2	Pbis Option 3	Pbis Option 4	Pbis Key
1	195029	0.79	0.02	0.01	*79.16	10.64	2.06	8.10	*0.4	-0.19	-0.22	-0.26	0.40
2	195029	0.69	0.06	0.03	11.18	15.30	*69.45	3.98	-0.28	-0.24	*0.47	-0.20	0.47
3	195029	0.76	0.05	0.09	11.54	7.85	4.96	*75.52	-0.28	-0.17	-0.26	*0.45	0.45
4	195029	0.65	0.10	0.10	20.27	6.75	7.84	*64.95	-0.15	-0.26	-0.23	*0.4	0.40
5	195029	0.63	0.08	0.07	1.56	*63.39	12.56	22.34	-0.19	*0.39	-0.15	-0.26	0.39
6	195029	0.66	0.08	0.07	16.38	13.38	*65.90	4.19	-0.20	-0.25	*0.42	-0.19	0.42
7	195029	0.60	0.09	0.07	20.76	*60.16	3.55	15.38	-0.25	*0.45	-0.22	-0.21	0.45
8	195029	0.82	0.06	0.09	7.28	*82.44	5.86	4.27	-0.27	*0.47	-0.26	-0.23	0.47
9	195029	0.83	0.12	0.09	4.05	7.46	5.21	*83.07	-0.19	-0.25	-0.23	*0.42	0.42
10	195029	0.92	0.09	0.04	2.41	2.45	*91.79	3.22	-0.28	-0.24	*0.44	-0.23	0.44
11	195029	0.58	0.11	0.05	16.36	12.82	*58.31	12.35	-0.13	-0.27	*0.4	-0.17	0.40
12	195029	0.94	0.10	0.03	3.27	*94.04	1.76	0.80	-0.14	*0.24	-0.12	-0.16	0.24
13	195029	0.77	0.14	0.06	13.88	3.27	*77.48	5.17	-0.25	-0.25	*0.43	-0.20	0.43
14	195029	0.66	0.15	0.07	4.83	26.45	2.25	*66.26	-0.23	-0.12	-0.23	*0.3	0.30
15	195029	0.79	0.16	0.05	*78.55	2.45	3.71	15.09	*0.39	-0.22	-0.28	-0.20	0.39
16	195029	0.80	0.18	0.07	6.37	2.93	10.29	*80.16	-0.33	-0.30	-0.25	*0.53	0.53
17	195029	0.68	0.32	0.04	8.06	21.55	*67.66	2.37	-0.20	-0.27	*0.44	-0.23	0.44
18	195029	0.79	0.38	0.08	*79.36	6.92	4.37	8.89	*0.47	-0.23	-0.25	-0.26	0.47
19	195029	0.65	0.45	0.08	13.23	5.85	14.92	*65.47	-0.14	-0.25	-0.23	*0.41	0.41
20	195029	0.74	0.49	0.04	7.74	14.49	*74.44	2.81	-0.17	-0.34	*0.47	-0.20	0.47
21	195029	0.89	0.55	0.03	2.29	*88.68	4.75	3.71	-0.25	*0.47	-0.28	-0.23	0.47
22	195029	0.86	0.68	0.07	7.15	3.21	*85.94	2.95	-0.26	-0.25	*0.45	-0.20	0.45
23	195029	0.85	0.77	0.07	*85.41	5.42	3.83	4.50	*0.48	-0.22	-0.28	-0.26	0.48
24	195029	0.86	1.46	0.06	4.98	*85.88	4.40	3.23	-0.24	*0.45	-0.26	-0.20	0.45
25	195029	0.58	1.67	0.06	23.05	*57.64	10.01	7.57	-0.21	*0.48	-0.22	-0.24	0.48
26	195029	0.60	1.90	0.11	5.42	18.97	13.69	*59.91	-0.27	-0.07	-0.35	*0.47	0.47
27	195029	0.60	2.17	0.05	*60.16	7.18	22.69	7.75	*0.4	-0.26	-0.09	-0.27	0.40
28	195029	0.68	2.36	0.02	12.34	5.04	*67.89	12.34	-0.23	-0.26	*0.5	-0.25	0.50
29	195029	0.71	0.07	0.46	5.26	26.30	46.13	21.78					
30	195029	0.67	0.12	1.21	8.64	29.73	43.00	17.30					
31	195029	0.71	0.09	1.71	18.01	46.36	33.83						

**Table 9c. P-values, Scored Response Distributions, and Point Biseriars, Grade 5**

Item	N-count	P-value	% Omit	% at 0	% Sel Option 1	% Sel Option 2	% Sel Option 3	% Sel Option 4	Pbis Option 1	Pbis Option 2	Pbis Option 3	Pbis Option 4	Pbis Key
1	195928	0.76	0.04	0.01	6.94	*76.41	3.16	13.44	-0.20	*0.39	-0.21	-0.23	0.39
2	195928	0.79	0.03	0.02	9.50	8.35	*79.10	3.00	-0.22	-0.21	*0.38	-0.18	0.38
3	195928	0.94	0.03	0.02	1.84	2.43	*93.64	2.04	-0.21	-0.26	*0.4	-0.20	0.40
4	195928	0.48	0.04	0.02	26.22	*48.27	15.79	9.65	-0.23	*0.29	0.02	-0.17	0.29
5	195928	0.95	0.03	0.07	*95.42	1.10	0.86	2.52	*0.35	-0.17	-0.19	-0.23	0.35
6	195928	0.90	0.05	0.09	3.98	2.24	3.37	*90.27	-0.27	-0.19	-0.19	*0.4	0.40
7	195928	0.78	0.08	0.03	5.21	*77.99	13.37	3.31	-0.22	*0.43	-0.26	-0.22	0.43
8	195928	0.89	0.08	0.03	1.09	1.59	*89.00	8.21	-0.21	-0.22	*0.44	-0.32	0.44
9	195928	0.66	0.08	0.02	13.48	*66.12	14.39	5.90	-0.06	*0.2	-0.09	-0.16	0.20
10	195928	0.63	0.08	0.03	28.44	5.52	*63.43	2.50	-0.17	-0.19	*0.33	-0.25	0.33
11	195928	0.75	0.11	0.03	*75.00	2.83	16.28	5.74	*0.36	-0.21	-0.20	-0.20	0.36
12	195928	0.75	0.11	0.04	*74.51	14.54	2.87	7.93	*0.31	-0.14	-0.20	-0.19	0.31
13	195928	0.73	0.15	0.09	6.71	5.69	14.29	*73.06	-0.28	-0.26	-0.12	*0.39	0.39
14	195928	0.85	0.18	0.03	2.57	*84.93	9.55	2.74	-0.21	*0.43	-0.27	-0.23	0.43
15	195928	0.78	0.19	0.04	6.83	10.09	*77.53	5.33	-0.24	-0.21	*0.46	-0.29	0.46
16	195928	0.78	0.41	0.03	2.85	15.74	*78.10	2.87	-0.21	-0.29	*0.43	-0.21	0.43
17	195928	0.70	0.44	0.05	*70.15	15.30	9.03	5.03	*0.33	-0.14	-0.10	-0.32	0.33
18	195928	0.88	0.45	0.04	2.41	*87.78	1.70	7.62	-0.20	*0.45	-0.22	-0.30	0.45
19	195928	0.83	0.54	0.06	2.00	*83.33	9.10	4.97	-0.20	*0.39	-0.17	-0.29	0.39
20	195928	0.72	0.76	0.02	11.98	4.83	10.49	*71.92	-0.24	-0.21	-0.23	*0.45	0.45
21	195928	0.59	1.10	13.10	53.97	31.83							
22	195928	0.93	0.06	0.00	*93.14	0.78	4.28	1.74	*0.3	-0.13	-0.23	-0.13	0.30
23	195928	0.90	0.07	0.02	*89.56	1.54	1.97	6.83	*0.36	-0.18	-0.19	-0.24	0.36
24	195928	0.76	0.08	0.01	0.78	21.68	*75.87	1.58	-0.15	-0.31	*0.4	-0.22	0.40
25	195928	0.79	0.13	0.01	2.75	8.41	*78.87	9.84	-0.11	-0.11	*0.24	-0.17	0.24
26	195928	0.82	0.13	3.67	29.03	67.17							
27	195928	0.44	0.29	29.66	25.55	27.27	17.22						

**Table 9d. P-values, Scored Response Distributions, and Point Biserials, Grade 6**

Item	N-count	P-value	% Omit	% at 0	% Sel Option 1	% Sel Option 2	% Sel Option 3	% Sel Option 4	% at 5	Pbis Option 1	Pbis Option 2	Pbis Option 3	Pbis Option 4	Pbis Key
1	198628	0.79	0.04	0.01	14.90	3.43	*79.45	2.16		-0.20	-0.21	*0.3	-0.07	0.30
2	198628	0.93	0.02	0.01	*92.74	3.61	0.70	2.91		*0.28	-0.19	-0.12	-0.17	0.28
3	198628	0.95	0.02	0.01	1.69	*95.17	2.29	0.81		-0.22	*0.35	-0.20	-0.18	0.35
4	198628	0.95	0.03	0.02	1.63	*94.66	1.67	1.98		-0.17	*0.35	-0.19	-0.24	0.35
5	198628	0.51	0.13	0.03	*50.71	8.65	27.70	12.78		*0.31	-0.07	-0.16	-0.19	0.31
6	198628	0.82	0.06	0.05	14.70	1.59	1.96	*81.64		-0.09	-0.18	-0.20	*0.22	0.22
7	198628	0.83	0.07	0.01	7.11	5.13	*83.37	4.30		-0.19	-0.25	*0.39	-0.19	0.39
8	198628	0.85	0.07	0.03	5.48	6.40	*85.24	2.79		-0.25	-0.34	*0.51	-0.24	0.51
9	198628	0.65	0.08	0.04	*64.53	9.14	6.95	19.26		*0.34	-0.20	-0.19	-0.15	0.34
10	198628	0.79	0.09	0.03	4.84	*79.44	3.92	11.69		-0.24	*0.45	-0.29	-0.22	0.45
11	198628	0.87	0.07	0.05	*86.69	6.45	4.34	2.40		*0.44	-0.21	-0.30	-0.22	0.44
12	198628	0.69	0.09	0.03	7.58	14.38	*69.17	8.76		-0.29	-0.20	*0.45	-0.21	0.45
13	198628	0.67	0.13	0.03	11.52	15.83	5.16	*67.34		-0.32	-0.19	-0.21	*0.47	0.47
14	198628	0.82	0.09	0.03	*82.43	6.91	4.49	6.05		*0.49	-0.22	-0.31	-0.28	0.49
15	198628	0.65	0.11	0.05	*64.92	20.67	5.63	8.61		*0.34	-0.09	-0.28	-0.21	0.34
16	198628	0.81	0.15	0.04	9.01	6.62	*80.67	3.50		-0.22	-0.28	*0.46	-0.25	0.46
17	198628	0.90	0.25	0.04	4.81	*89.59	2.30	3.00		-0.31	*0.49	-0.23	-0.26	0.49
18	198628	0.83	0.29	0.08	3.50	9.28	3.74	*83.11		-0.27	-0.28	-0.24	*0.48	0.48
19	198628	0.85	0.31	0.04	*84.85	3.47	8.21	3.12		*0.53	-0.24	-0.33	-0.28	0.53
20	198628	0.60	0.36	0.05	17.91	*60.13	9.03	12.53		-0.13	*0.29	-0.24	-0.06	0.29
21	198628	0.76	0.42	0.13	6.65	11.10	5.22	*76.48		-0.30	-0.25	-0.27	*0.52	0.52
22	198628	0.71	0.68	0.04	15.83	*70.87	6.75	5.84		-0.15	*0.36	-0.27	-0.14	0.36
23	198628	0.69	0.77	0.05	*68.87	4.61	20.29	5.41		*0.46	-0.26	-0.22	-0.27	0.46
24	198628	0.81	0.88	0.04	6.28	7.95	*81.38	3.46		-0.27	-0.22	*0.44	-0.19	0.44
25	198628	0.88	0.93	0.05	*88.14	4.25	3.42	3.21		*0.52	-0.28	-0.29	-0.27	0.52
26	198628	0.85	0.99	0.02	3.45	3.53	*84.99	7.03		-0.27	-0.24	*0.48	-0.26	0.48
27	198628	0.64	0.09	0.79	6.24	18.18	33.21	28.69	12.80					
28	198628	0.65	0.14	0.98	6.03	16.44	32.17	31.30	12.95					
29	198628	0.73	0.13	1.16	15.12	47.13	36.46							

**Table 9e. P-values, Scored Response Distributions, and Point Biseriars, Grade 7**

Item	N-count	P-value	% Omit	% at 0	% Sel Option 1	% Sel Option 2	% Sel Option 3	% Sel Option 4	Pbis Option 1	Pbis Option 2	Pbis Option 3	Pbis Option 4	Pbis Key
1	198390	0.99	0.01	0.01	*98.64	0.79	0.38	0.18	*0.22	-0.16	-0.13	-0.07	0.22
2	198390	0.88	0.06	0.01	5.15	2.04	*87.71	5.04	-0.12	-0.17	*0.31	-0.22	0.31
3	198390	0.82	0.09	0.03	6.11	1.68	10.47	*81.63	-0.17	-0.17	-0.14	*0.28	0.28
4	198390	0.77	0.06	0.02	4.67	2.94	15.73	*76.58	-0.23	-0.29	0.06	*0.18	0.18
5	198390	0.86	0.04	0.04	5.45	5.04	3.74	*85.68	-0.29	-0.38	-0.20	*0.54	0.54
6	198390	0.81	0.11	0.02	7.40	3.50	*80.83	8.15	-0.28	-0.23	*0.45	-0.22	0.45
7	198390	0.73	0.11	0.03	13.29	11.23	*72.84	2.50	-0.16	-0.26	*0.37	-0.18	0.37
8	198390	0.73	0.13	0.05	*72.75	5.26	4.54	17.28	*0.45	-0.22	-0.26	-0.25	0.45
9	198390	0.65	0.19	0.05	15.34	*65.11	14.19	5.12	-0.29	*0.39	-0.13	-0.17	0.39
10	198390	0.64	0.10	0.07	22.49	4.03	9.01	*64.30	-0.16	-0.18	-0.13	*0.3	0.30
11	198390	0.72	0.20	0.02	5.39	*72.07	13.53	8.79	-0.19	*0.37	-0.24	-0.14	0.37
12	198390	0.83	0.08	0.04	7.45	4.12	*82.92	5.38	-0.25	-0.17	*0.46	-0.31	0.46
13	198390	0.72	0.23	0.04	*71.75	6.71	7.36	13.91	*0.37	-0.22	-0.17	-0.19	0.37
14	198390	0.83	0.11	0.04	6.81	5.01	*82.56	5.47	-0.22	-0.24	*0.49	-0.33	0.49
15	198390	0.69	0.10	0.03	20.91	3.84	5.91	*69.21	-0.22	-0.22	-0.15	*0.36	0.36
16	198390	0.92	0.14	0.03	3.47	2.58	2.05	*91.73	-0.24	-0.24	-0.22	*0.42	0.42
17	198390	0.61	0.17	0.03	11.78	*61.19	19.98	6.84	-0.17	*0.3	-0.14	-0.13	0.30
18	198390	0.69	0.23	0.05	14.68	*69.14	6.95	8.96	-0.07	*0.33	-0.30	-0.17	0.33
19	198390	0.66	0.21	0.05	11.96	7.17	*66.13	14.48	-0.24	-0.23	*0.41	-0.15	0.41
20	198390	0.66	0.45	0.04	21.76	6.50	*65.90	5.34	-0.10	-0.31	*0.34	-0.18	0.34
21	198390	0.76	0.44	0.04	3.71	2.54	*76.22	17.05	-0.30	-0.26	*0.41	-0.18	0.41
22	198390	0.79	0.60	0.05	9.74	4.89	5.95	*78.77	-0.20	-0.26	-0.22	*0.43	0.43
23	198390	0.84	0.63	0.05	5.68	5.12	4.74	*83.77	-0.29	-0.31	-0.28	*0.56	0.56
24	198390	0.79	0.69	0.05	*79.02	11.01	4.83	4.41	*0.55	-0.33	-0.29	-0.22	0.55
25	198390	0.53	0.81	0.05	9.16	*53.41	17.43	19.13	-0.20	*0.31	-0.12	-0.10	0.31
26	198390	0.84	1.02	0.02	*84.02	4.94	6.25	3.76	*0.5	-0.26	-0.27	-0.26	0.50
27	198390	0.69	2.67	10.33	36.26	50.74							
28	198390	0.61	4.46	15.12	38.05	42.37							
29	198390	0.97	0.13	0.01	1.56	*96.60	0.92	0.78	-0.15	*0.26	-0.13	-0.18	0.26
30	198390	0.82	0.16	0.01	*82.07	7.60	3.66	6.49	*0.41	-0.25	-0.16	-0.24	0.41
31	198390	0.82	0.15	0.02	10.55	7.38	*81.63	0.26	-0.20	-0.09	*0.23	-0.09	0.23
32	198390	0.80	0.19	0.01	9.21	5.46	*80.50	4.63	-0.14	-0.25	*0.33	-0.16	0.33
33	198390	0.88	0.21	2.91	16.87	80.01							
34	198390	0.80	0.27	3.48	32.15	64.10							
35	198390	0.31	0.38	42.48	28.25	23.21	5.68						

**Table 9f. P-values, Scored Response Distributions, and Point Biserials, Grade 8**

Item	N-count	P-value	% Omit	% at 0	% Sel Option 1	% Sel Option 2	% Sel Option 3	% Sel Option 4	% at 5	Pbis Option 1	Pbis Option 2	Pbis Option 3	Pbis Option 4	Pbis Key
1	206153	0.83	0.04	0.01	2.10	*83.06	7.31	7.49		-0.06	*0.36	-0.21	-0.26	0.36
2	206153	0.92	0.03	0.02	*92.34	3.80	1.13	2.69		*0.39	-0.26	-0.21	-0.20	0.39
3	206153	0.91	0.05	0.01	4.99	1.36	*91.34	2.25		-0.22	-0.17	*0.36	-0.23	0.36
4	206153	0.90	0.05	0.03	*90.12	3.39	2.55	3.87		*0.45	-0.28	-0.22	-0.25	0.45
5	206153	0.83	0.04	0.02	13.40	1.62	2.11	*82.81		-0.17	-0.08	-0.20	*0.26	0.26
6	206153	0.89	0.08	0.02	2.25	2.45	*88.66	6.55		-0.22	-0.22	*0.32	-0.14	0.32
7	206153	0.83	0.05	0.02	*83.27	4.29	2.86	9.51		*0.35	-0.26	-0.23	-0.13	0.35
8	206153	0.52	0.14	0.03	9.14	20.48	*51.64	18.58		-0.20	-0.15	*0.33	-0.11	0.33
9	206153	0.82	0.09	0.02	3.74	*82.50	10.01	3.64		-0.25	*0.46	-0.28	-0.21	0.46
10	206153	0.72	0.06	0.03	*72.47	7.31	3.18	16.95		*0.37	-0.27	-0.21	-0.15	0.37
11	206153	0.91	0.05	0.03	*90.86	1.67	5.39	2.00		*0.38	-0.22	-0.25	-0.17	0.38
12	206153	0.70	0.12	0.02	8.60	*69.89	11.03	10.34		-0.25	*0.43	-0.22	-0.18	0.43
13	206153	0.78	0.07	0.02	6.71	*77.56	9.51	6.14		-0.11	*0.29	-0.22	-0.11	0.29
14	206153	0.73	0.06	0.03	*72.88	9.65	2.85	14.53		*0.29	-0.18	-0.11	-0.16	0.29
15	206153	0.67	0.07	0.03	15.35	11.84	*67.23	5.47		-0.11	-0.13	*0.24	-0.12	0.24
16	206153	0.78	0.14	0.02	7.51	6.26	7.90	*78.16		-0.25	-0.29	-0.18	*0.46	0.46
17	206153	0.84	0.12	0.02	5.49	4.39	*83.67	6.30		-0.30	-0.25	*0.48	-0.23	0.48
18	206153	0.44	0.14	0.03	8.50	*44.34	13.83	33.17		-0.17	*0.23	-0.12	-0.04	0.23
19	206153	0.76	0.18	0.03	*76.50	6.57	10.33	6.39		*0.44	-0.32	-0.20	-0.16	0.44
20	206153	0.93	0.15	0.03	1.49	3.54	*92.59	2.21		-0.24	-0.33	*0.49	-0.24	0.49
21	206153	0.85	0.28	0.03	10.23	2.72	*84.78	1.95		-0.34	-0.25	*0.48	-0.16	0.48
22	206153	0.59	0.35	0.04	23.00	7.98	*59.19	9.44		-0.14	-0.19	*0.29	-0.10	0.29
23	206153	0.63	0.39	0.03	27.08	4.30	5.44	*62.75		-0.10	-0.23	-0.17	*0.29	0.29
24	206153	0.79	0.42	0.03	*79.14	15.10	2.75	2.57		*0.31	-0.13	-0.20	-0.24	0.31
25	206153	0.81	0.50	0.03	*80.61	3.29	2.61	12.96		*0.49	-0.21	-0.22	-0.33	0.49
26	206153	0.72	0.55	0.03	6.66	4.36	16.05	*72.34		-0.21	-0.25	-0.27	*0.47	0.47
27	206153	0.70	0.18	0.71	4.44	12.87	28.10	31.85	21.85					
28	206153	0.75	0.14	0.43	2.67	8.38	24.72	35.64	28.01					
29	206153	0.77	0.20	1.13	11.20	43.63	43.84							

### **Point-Biserial Correlation Coefficients**

Point-biserial (pbis) statistics are used to examine item-test correlations or item discrimination for MC items. In the Tables 9a–9f, point-biserial correlation coefficients were computed for each answer option. Point biserials for the correct answer option are denoted with an asterisk (\*) and are repeated in the Pbis Key field. The point-biserial correlation is a measure of internal consistency that ranges between +/-1. It indicates a correlation of students' responses to an item relative to their performance on the rest of the test. 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. For incorrect answer options (distractors), the point biserial should be negative, which indicated that students who scored lower on the overall test had a tendency to pick a distractor. None of the grades had any item answer keys that were flagged for low point biserials. Point biserials for correct answer options (pbis\*) on the tests ranged 0.18–0.56. For Grade 3, the pbis\* were between 0.25 and 0.55. For Grade 4, the pbis\* were between 0.24 and 0.53. For Grade 5, the pbis\* were between 0.20 and 0.46. For Grade 6, pbis\* were between 0.22 and 0.53. For Grade 7, the pbis\* were between 0.18 and 0.56. For Grade 8, the pbis\* were between 0.23 and 0.49.

### **Distractor Analysis**

Item distractors provide additional information on student performance on test questions. Two types of information on item distractors are available from New York State test data: information on proportion of students selecting incorrect item response options and the point biserial coefficient of distractors (discrimination power of incorrect answer choices). The proportions of students selecting incorrect responses while responding to MC items are provided in Tables 9a–9f of this report. Distribution of student responses across answer choices was evaluated. It was expected that the proportion of students selecting the correct answer would be higher than proportions of students selecting any other answer choice. This was true for all New York State ELA items.

As mentioned in the “Point-Biserial Correlations Coefficients” subsection, items were flagged if the point biserial of any distractor was positive. The only item with a distractor that had a non-negative point biserial was item number 1 in Grade 7, which had a point biserial of 0. All other point biserials for distractors in each grade were negative.

### **Test Statistics and Reliability Coefficients**

Test statistics including raw-score mean and raw-score standard deviation are presented in Table 10. For both Grades 4 and 8, weighted and unweighted test statistics are provided. Grade 4 and 8 CR items were weighted by a 1.38 factor to increase proportion of score points obtainable from these items. Weighting CR items for these two grades resulted in better alignment of proportions of test raw-score points obtainable from MC and CR items between 2006 and 2008 ELA operational tests for these grades. More information on weighting CR items and the effect on test content is provided in Section VI, “IRT Scaling and Equating.” 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 0.84–0.90. Feldt-Raju reliability coefficients ranged 0.85–0.90. The lowest reliability was observed for the Grade 5 test, but since that test had the lowest number of score points it was reasonable that its reliability would not be as high as the other grades’ tests. The highest reliability was observed for the Grade 4 test. All reliabilities met or exceeded 0.80, 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 10. NYSTP ELA 2008 Test Form Statistics and Reliability**

Grade	Max RS	RS Mean	RS SD	P-value Mean	Cronbach’s alpha	Feldt-Raju
3	33	26.02	5.78	0.79	0.86	0.87
4	39 (43 WGT)	28.31 (31.20 WGT)	7.36 (8.02 WGT)	0.73	0.90	0.90
5	31	23.06	5.37	0.74	0.84	0.85
6	39	29.12	6.79	0.75	0.88	0.89
7	41	30.04	7.04	0.73	0.87	0.88
8	39 (44 WGT)	29.68 (33.32 WGT)	6.49 (7.34 WGT)	0.76	0.86	0.88

Note: WGT = weighted results

### **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 9a–9f 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 needs resource category (focal group: High Needs; reference group: Low Needs), gender (focal group: Female; reference group: Male), and ethnicity (focal groups: Black, Hispanic, and Asian; reference group: White). The DIF analyses were conducted using all cases from the clean data sets. Table 11 shows the number of cases for subgroups.

**Table 11. NYSTP ELA 2008 Classical DIF Sample N-Counts**

Grade	Ethnicity				Gender		Needs Resource Category	
	Black\ African American	Hispanic\ Latino	Asian	White	Female	Male	High	Low
3	36703	40915	14256	101559	94184	99249	103174	87295
4	37119	40877	14136	102897	95690	99339	103039	89633
5	37494	40285	14503	103646	95932	99996	101798	90867
6	37567	40299	14481	106281	96916	101712	102844	93071
7	38932	40830	14227	104401	97433	100957	106169	89937
8	39658	40525	14227	111743	100953	105200	107300	97434

Table 12 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 operational item selection for possible item bias. Only those items that were determined free of bias were included in the operational tests.

**Table 12. Number of Items Flagged by SMD and Mantel-Haenszel DIF Methods**

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

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  $\theta$  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-performing student.

For analysis of the CR items, the two-parameter partial credit (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  $\theta$  having a score  $(k - 1)$  at the  $k$ -th level of the  $j$ -th item is

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

where

$$Z_{jk} = A_{jk}\theta + C_{jk}; \text{ and}$$

$k$  is the item response category ( $k=1, 2, \dots, m$ ).

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 where

$\alpha_j$  and  $\gamma_{ji}$  are the free parameters to be estimated from the data.

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

### ***Calibration Sample***

The cleaned classical analysis and calibration sample data (as described in Section V, subsection, “Classical Analysis and Calibration Sample Characteristics”) was used for calibration and scaling of New York State ELA Tests. It should be noted that the scaling was done on nearly the total New York State population of students in public schools, and exclusion of some cases during the data cleaning had very minimal or no effect on parameter estimation.

### ***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 operational 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. These default values of the  $c$ -parameter were obtained during the field test calibration and were held unchanged between field test and operational administrations. For the Grades 3–8 ELA tests, all calibration estimation results are reasonable. The summary of calibration results is presented in Table 13.

**Table 13. NYSTP ELA 2008 Calibration Results**

Grade	Largest $a$ -parameter	$b$ -parameter Range		# Items with Default $c$ -parameter	Theta Mean	Theta Standard Deviation	# Students
3	2.496	-4.829	0.974	13	0.15	1.356	193433
4	2.169	-3.594	0.097	11	0.04	1.182	195029
5	2.063	-4.200	0.727	15	0.07	1.232	195928
6	2.770	-3.777	0.521	14	0.06	1.186	198628
7	3.173	-5.581	0.418	14	0.05	1.170	198390
8	3.073	-4.447	0.691	11	0.07	1.201	206153

### ***Item-Model Fit***

Item-fit statistics discern the appropriateness of using an item in the 3PL or 2PPC model. A procedure described by Yen (1981) was used to measure fit to the 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

$$Q_{i1} = \sum_{k=1}^{10} \frac{N_{ik} (O_{ik} - E_{ik})^2}{E_{ik} (1 - E_{ik})},$$

with

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

A modification of this procedure was used to measure fit to the two-parameter partial credit model. For the two-parameter partial credit model,  $Q_{lj}$  was assumed to have approximately a chi-square distribution with the following degree of freedom:

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

where

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

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

where

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

The value of  $Z$  still will increase with sample size, all else being 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 operational tests, which have large calibration sample sizes, the criterion  $Z_{Q_1}Crit$  used to flag items was calculated using the expression

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

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. It should be noted that all items in the NYSTP 2008 ELA Tests for Grades 4, 6, and 8 demonstrated good model fit. Item 9 in Grade 3, item 21 in Grade 5, and item 4 in Grade 7 exhibited poor item-model fit statistics. The fact that only three items were flagged for poor fit across all ELA tests further supports the use of the chosen models. Fit statistics and status for all items in the Grades 3–8 ELA Tests are presented in Appendix F.

### ***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

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

and

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

The generalization to items with multiple response categories uses

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

where

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

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 only one pair of items was found to be locally dependent. Grade 7 items 27 and 28 (both CR items) were found to be locally dependent ( $Q_3 = 0.282$ ). The magnitude of this statistic was not sufficient to warrant concern about inflating score reliability estimates.

### ***Scaling and Equating***

The 2008 Grades 3–8 ELA assessments were calibrated and equated to the associated 2007 assessments, using two separate equating procedures.

In the first equating procedure, the new 2008 OP forms were pre-equated to the corresponding 2007 assessments. Prior to pre-equating, the FT items administered in 2007 were placed onto the OP scales in each grade. The equating of 2007 FT items to the 2007 OP scales was conducted via common examinees. FT items that were eligible for future OP administrations were then included in the NYS item pool. Other items in the NYS item pool were items field tested in 2006, 2005, and (for Grades 4 and 8 only) 2003. All items field tested between 2003 and 2006 were also 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.

At the pre-equating stage, the pool of FT items administered in years 2003, 2005, 2006, and 2007 was used to select the 2008 OP test forms using the following criteria:

- Content coverage of each form matched the test blueprint
- Psychometric properties of the items:
  - item fit (see subsection “Item-Model Fit”)
  - differential item functioning (see subsections “Differential Item Functioning” and “IRT DIF Statistics”)
  - item difficulty (see subsection “Item Difficulty and Response Distribution”)
  - item discrimination (see subsection “Point-Biserial Correlation Coefficient”)
  - omit rates (see subsection “Speededness”)
- Test characteristic curve (TCC) and standard error (SE) curve alignment of the 2008 forms with the target 2007 OP forms (Note that the 2007 OP TCC and SE curves

were based on OP parameters and the 2008 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 2008 ELA OP data were re-calibrated after the 2008 OP administration. FT parameters for all MC items in OP tests were used as anchors to transform the 2008 OP item parameters to the OP scale. The CR items were not used as anchors in order to avoid potential error associated with rater effect. The MC items contained in the 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 and that relationship defined by OP anchor item parameter estimates. This places the transformed parameters for the OP test items onto the New York State OP scale.

In this procedure, new OP parameter estimates were obtained for all items. The  $a$ -parameters and  $b$ -parameters were allowed to be estimated freely while  $c$ -parameters of anchor items were fixed to their FT parameter 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:

$$M1 = A * MI_{Fi}$$
$$M2 = A * M2_{Fi} + B$$

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  $MI_{Fi}$  and  $M2_{Fi}$  are the transformation constants previously used to place the anchor item FT parameter estimates onto the NYS scale.

The  $A$  and  $B$  values are derived from the input (FT) and estimate (OP) values of anchor items. Anchor input or FT values are known item parameter estimates entered into equating. Anchor estimate or OP values are parameter estimates for the same anchor items re-estimated during the equating procedure. The input and estimate anchor parameter estimates are expected to have similar values. The  $A$  and  $B$  constants are computed as follows:

$$A = \frac{SD_{op}}{SD_{Fi}}$$

$$B = (Mean_{OP} - \frac{SD_{Op}}{SD_{Fi}} Mean_{Fi})$$

where

$SD_{Op}$  is the standard deviation of anchor estimates in scale score metric.

$SD_{Fi}$  is the standard deviation of anchor input values in scale score metric.

$Mean_{Op}$  is the mean of anchor estimates in scale score metric.

$Mean_{Fi}$  is the mean of anchor input in scale score metric.

The  $M1$  and  $M2$  transformation parameters obtained in the Stocking and Lord equating process were used to transform item parameters obtained in the calibration process into the final scale score metric. Table 14 presents the 2008 OP transformation parameters for New York State Grades 3–8 ELA.

**Table 14. NYSTP ELA 2008 Final Transformation Constants**

Grade	$M1$	$M2$
3	28.515	665.451
4	34.630	665.882
5	22.287	665.689
6	23.068	660.034
7	23.286	661.251
8	29.170	655.261

### 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. In the New York State Testing Program, different anchor sets are used each year to minimize item exposure that could adversely affect the accuracy of the equatings.

### Anchor Item Evaluation

Anchor items were evaluated using several procedures. Procedures 1 and 2 evaluate the overall anchor set, while procedures 3, 4, and 5 evaluate individual anchor items.

1. Anchor set input and estimate 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 FT and the 2008 OP administration.
2. Correlations of anchor input and estimates of  $a$ - and  $b$ -parameters and p-values. Correlations of anchor input and estimates of  $a$ - and  $b$ -parameters and p-values were evaluated for magnitude. Ideally, the correlations between anchor input and estimate for  $a$ -parameter should be at least 0.80 and the correlations for  $b$ -parameters and p-values 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 FT estimates and the other on transformed estimates from the 2008 OP calibration. 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.
4. Delta plots (differences in the standardized proportion correct value). The delta-plot method relies on the differences in the standardized proportion correct value (p-value). P-values of the anchor items based on the FT (years 2003, 2005, 2006, and/or 2007) and the 2008 OP administration were calculated. The p-values were then converted to z-scores that correspond to the (1-p)th percentiles. A rule to identify outlier items that are functioning differentially between the two groups with respect to the level of difficulty is to draw the perpendicular distance to the line-of-best-fit. The fitted line is chosen to minimize the sum of squared perpendicular distances of the points to the line. Items lying more than two standard deviations from the fitted line are flagged as outliers.
5. Lord's chi-square criterion. Lord's  $\chi^2$  criterion involves significance testing of both item difficulty and discrimination parameters simultaneously for each item and evaluating the results based on the chi-square distribution table. (For details see Divgi, 1985; Lord, 1980.) If the null hypothesis that the item difficulty and discrimination parameters are equal is true, the item is not flagged for differential performance. If the null hypothesis is rejected and the observed value for  $\chi^2$  is greater than the critical  $\chi^2$  value, the items are flagged for performance differences between the two item administrations.

Table 15 provides a summary of anchor item evaluation and item flags.

**Table 15. ELA Anchor Evaluation Summary**

Grade	Number of Anchors	Anchor Input/ Estimate Correlation			Flagged Anchors (item numbers)			
		<i>a</i> -par	<i>b</i> -par	p-value	RMSD <i>a</i> -par	RMSD <i>b</i> -par	Delta	Lord's Chi-Square
3	24	0.877	0.963	0.974	14	22	11, 24	
4	28	0.789	0.946	0.911	10	1, 2, 3	2, 3	2
5	24	0.919	0.902	0.960	6, 18	25		
6	26	0.823	0.909	0.965	3	2, 3		3
7	30	0.901	0.877	0.921	19	1, 30	1	1
8	26	0.821	0.716	0.878	3, 4		11	2,3, 4,5,6,7,9,11

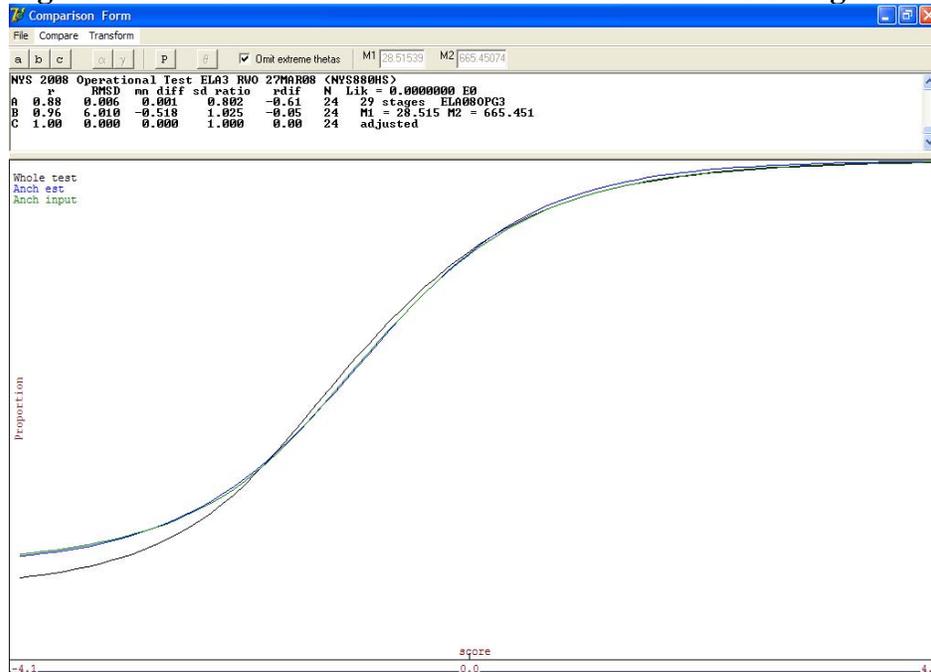
It should be noted that in all cases the overall TCC alignment for anchor set input and estimate was good. The correlations for input and estimated p-values were over 0.90 for all grades except Grade 8, for which the correlation for input and estimate p-values was 0.88. Correlations for *b*-parameter input and estimates ranged from 0.72 for Grade 8 to 0.96 for Grade 3. Correlations for *a*-parameter input and estimate ranged from 0.79 for Grade 4 to 0.92 for Grade 5. Correlations between *a*-parameter input and estimates for Grade 4 and correlations between *b*-parameter input and estimates for Grades 7 and 8 were slightly below the NYS criterion.

It was found that the overall TCC alignment for anchor set input and estimate was very good (see Figures 1–5). In addition, correlations between parameter input and estimates were satisfactory for Grades 3–7. Therefore, despite the fact that some individual items were flagged by multiple methods in Grades 4, 6, and 7, no anchors were removed from any of the anchor sets. It was determined that removal of flagged anchors from Grades 3–7 anchor sets had only minimal effect on item parameter estimates and minimal or no effect on the scoring tables.

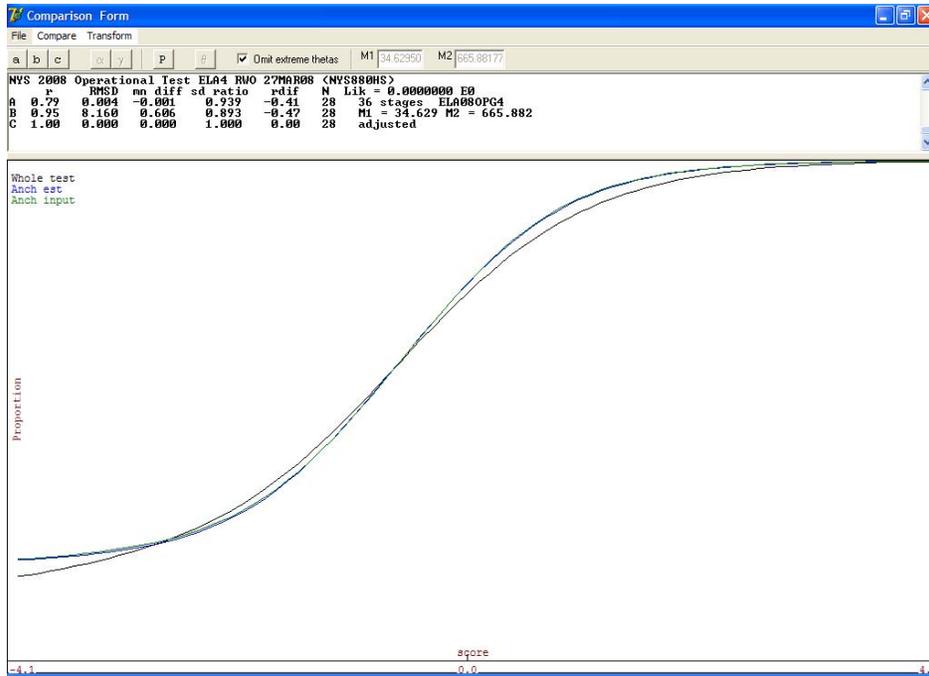
An investigation of lower than expected correlations for Grade 8 revealed that under OP administration conditions students performed better on anchor items 1–5, field tested in 2007, while performing slightly worse on items 6–15, which were field tested in 2003 or 2005. It is possible that familiarity of items field tested in 2007 could have contributed to better performance on these items when they were administered operationally. Also, several factors (including population changes, curriculum changes, instruction method changes, etc.) might have contributed to item parameter changes between 2003/2005 FT and 2008 OP administration. Three Grade 8 items (items 3, 4, and 11) were flagged by more than one anchor evaluation method. It was determined that removal of these items would not negatively affect the anchor set content coverage. A test run of Grade 8 equating without anchors 3, 4, and 11 revealed that there would be only a minimal impact from removing these

items from the anchor set on the estimated parameters and no impact on the scoring table and student scores. However, because the overall anchor set TCC alignment for Grade 8 was good (see Figure 6), all anchor items were retained for this grade. Similarly, all anchor items were retained for the remaining grades. Retaining all anchor items in all grades allowed for adequate anchor item content coverage and maintaining anchor set reliability.

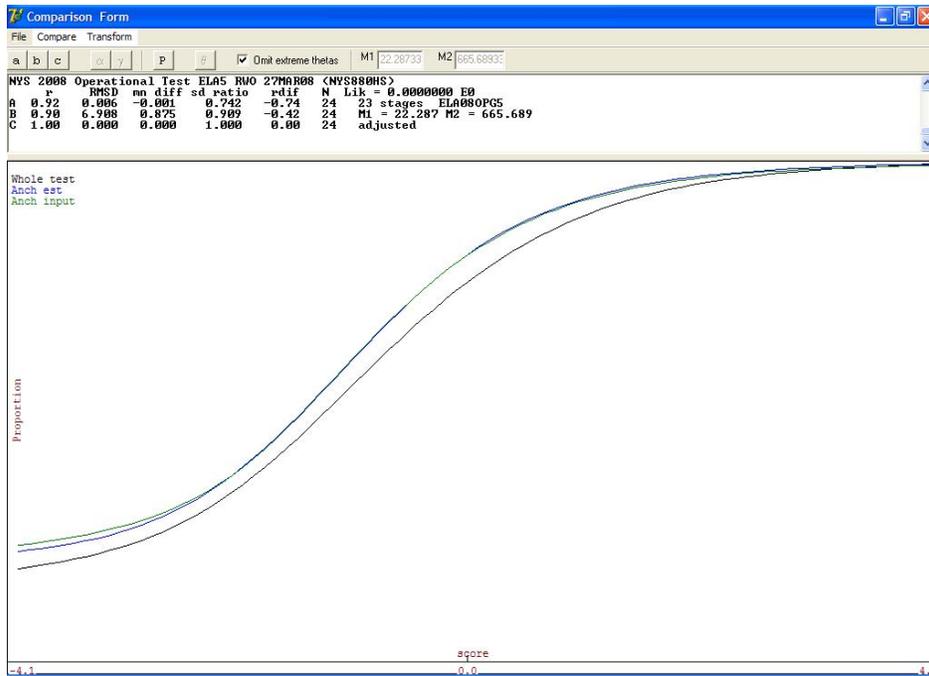
**Figure 1. ELA Grade 3 Anchor Set and Whole Test TCC Alignment**



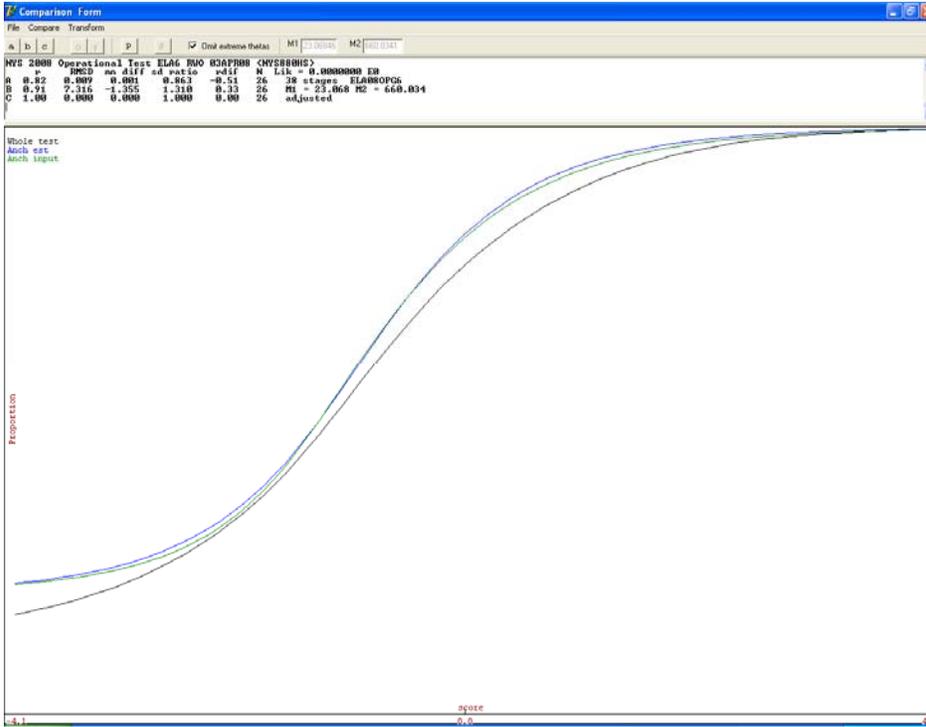
**Figure 2. ELA Grade 4 Anchor Set and Whole Test TCC Alignment**



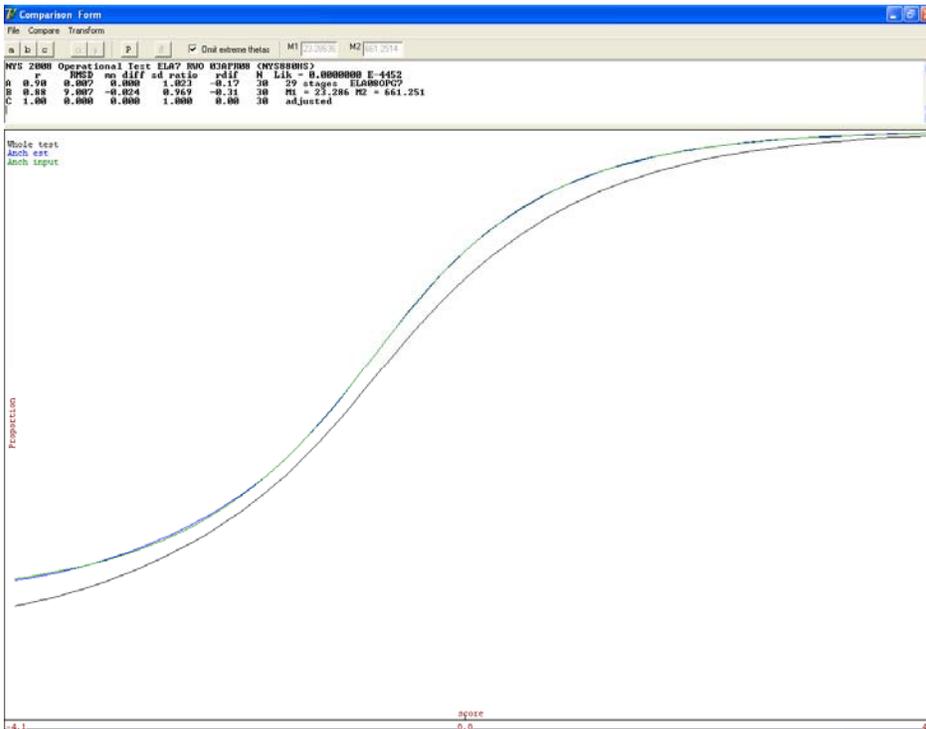
**Figure 3. ELA Grade 5 Anchor Set and Whole Test TCC Alignment**



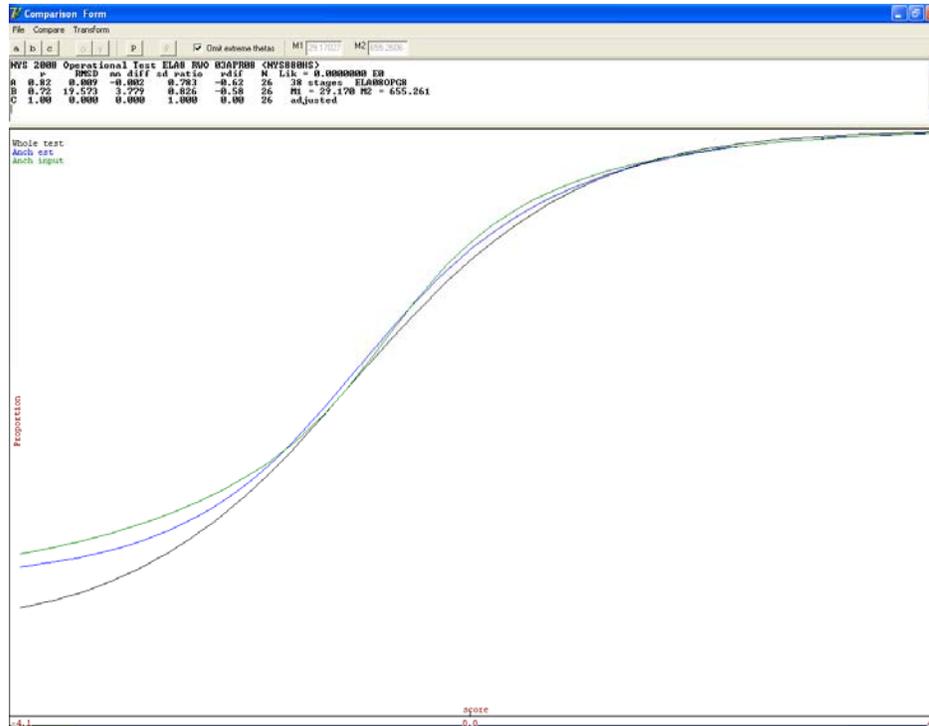
**Figure 4. ELA Grade 6 Anchor Set and Whole Test TCC Alignment**



**Figure 5. ELA Grade 7 Anchor Set and Whole Test TCC Alignment**



**Figure 6. ELA Grade 8 Anchor Set and Whole Test TCC Alignment**



Note that in Figures 1–6 anchor input parameters are represented by a green TCC, anchor parameter estimates are represented by a blue TCC, and the whole test (OP parameters for all items) is represented by a black TCC. As seen in all the figures the alignment of anchor input and estimated parameters is good, indicating overall good stability of anchor parameters between FT and OP test administrations.

It should be noted that in some cases the TCC for the whole test was not well aligned with the anchor set TCC. Such discrepancies between the anchor set TCC and whole test TCC are due to differences between anchor set difficulty and total test difficulty. The anchor set contains only MC items while the total test contains both MC and CR items. If the CR items are overall less difficult than MC item set, then the total test TCC will tend to be shifted to the left side of the anchor TCC. If the CR items are more difficult than MC items, then the total test TCC will likely be shifted to the right side of the anchor TCC (for example, Grade 5). The anchor sets used to equate new OP assessments to the NYS scale are MC items only, and these items are representative of the test blueprint. However, the difficulty of the anchor set does not always reflect the total test difficulty. (For example, the MC portion of the test may be somewhat less or more difficult than CR portion of the test.) If the difficulty of the anchor set does not reflect well the difficulty of the total test, some discrepancies in anchor set and whole test TCCs will likely occur. As stated before, the CR items were not included in anchor sets in order to avoid potential error associated with possible rater effects.

### ***Item Parameters***

The final item parameters in scale score metric obtained via linear transformation of theta metric parameters using the final *M1* and *M2* transformation constants, which are shown in Table 16, are presented in Tables 16a–16f. Descriptions of what each of the parameter variables mean is presented in the subsection depicting the IRT models and rationale.

**Table 16a. 2007 Operational Item Parameter Estimates, Grade 3**

Item	Max Pts	<i>a</i> -par/ $\alpha$	<i>b</i> -par/ $\gamma_1$	<i>c</i> -par/ $\gamma_2$	$\gamma_3$
1	1	0.034	624.927	0.219	
2	1	0.023	615.684	0.200	
3	1	0.016	667.886	0.146	
4	1	0.038	627.946	0.147	
5	1	0.043	628.419	0.138	
6	1	0.042	659.163	0.200	
7	1	0.021	653.082	0.200	
8	1	0.044	655.412	0.273	
9	1	0.017	627.410	0.200	
10	1	0.033	633.402	0.152	
11	1	0.051	638.501	0.136	
12	1	0.045	628.268	0.200	
13	1	0.031	653.149	0.200	
14	1	0.051	615.568	0.200	
15	1	0.038	639.756	0.125	
16	1	0.032	651.236	0.138	
17	1	0.037	634.490	0.200	
18	1	0.030	684.686	0.283	
19	1	0.029	645.908	0.138	
20	1	0.033	656.506	0.200	
21	2	0.064	39.952	39.122	
22	1	0.022	601.426	0.200	
23	1	0.039	606.195	0.200	
24	1	0.036	587.582	0.200	
25	1	0.022	611.552	0.200	
26	2	0.029	17.540	18.636	
27	2	0.019	12.661	12.096	
28	3	0.030	19.155	18.386	17.830

**Table 16b. 2007 Operational Item Parameter Estimates, Grade 4**

Item	Max Pts	<i>a</i> -par/ alpha	<i>b</i> -par/ gamma1	<i>c</i> -par/ gamma2	gamma3	gamma4
1	1	0.022	627.520	0.153		
2	1	0.029	651.541	0.186		
3	1	0.025	638.072	0.166		
4	1	0.023	656.721	0.177		
5	1	0.023	659.828	0.186		
6	1	0.023	652.808	0.151		
7	1	0.032	665.632	0.177		
8	1	0.031	629.516	0.200		
9	1	0.025	623.659	0.200		
10	1	0.037	608.498	0.200		
11	1	0.022	665.521	0.155		
12	1	0.017	565.243	0.200		
13	1	0.023	632.901	0.170		
14	1	0.013	646.885	0.170		
15	1	0.025	644.180	0.367		
16	1	0.035	634.310	0.170		
17	1	0.027	655.044	0.211		
18	1	0.030	635.442	0.200		
19	1	0.025	657.794	0.200		
20	1	0.028	641.696	0.175		
21	1	0.036	617.218	0.200		
22	1	0.030	620.472	0.200		
23	1	0.034	624.355	0.200		
24	1	0.031	621.503	0.200		
25	1	0.036	667.461	0.158		
26	1	0.029	663.206	0.149		
27	1	0.026	666.592	0.200		
28	1	0.034	654.703	0.183		
29	4	0.034	18.200	19.929	21.910	24.065
30	4	0.042	23.602	25.532	27.548	29.905
31	3	0.038	20.943	23.512	25.823	

**Table 16c. 2007 Operational Item Parameter Estimates, Grade 5**

Item	Max Pts	<i>a</i> -par/ alpha	<i>b</i> -par/ gamma1	<i>c</i> -par/ gamma2	gamma3
1	1	0.034	644.242	0.147	
2	1	0.033	641.558	0.191	
3	1	0.054	624.475	0.237	
4	1	0.028	680.944	0.200	
5	1	0.050	616.386	0.200	
6	1	0.044	627.774	0.200	
7	1	0.040	643.754	0.133	
8	1	0.049	631.777	0.185	
9	1	0.013	651.074	0.200	
10	1	0.028	659.416	0.164	
11	1	0.032	647.538	0.200	
12	1	0.025	644.803	0.200	
13	1	0.031	647.679	0.163	
14	1	0.044	637.345	0.200	
15	1	0.048	648.244	0.200	
16	1	0.040	645.505	0.200	
17	1	0.026	651.295	0.200	
18	1	0.050	634.737	0.200	
19	1	0.037	637.241	0.200	
20	1	0.048	654.110	0.190	
21	2	0.048	30.073	32.795	
22	1	0.033	612.989	0.200	
23	1	0.037	625.743	0.200	
24	1	0.044	653.968	0.312	
25	1	0.019	630.268	0.200	
26	2	0.038	22.570	24.404	
27	3	0.046	30.333	30.727	31.823

**Table 16d. 2007 Operational Item Parameter Estimates, Grade 6**

Item	Max Pts	<i>a</i> -par/ alpha	<i>b</i> -par/ gamma1	<i>c</i> -par/ gamma2	gamma3	gamma4	gamma5
1	1	0.022	626.993	0.200			
2	1	0.029	603.481	0.200			
3	1	0.045	610.917	0.305			
4	1	0.044	609.590	0.200			
5	1	0.028	671.071	0.181			
6	1	0.015	609.604	0.200			
7	1	0.033	629.208	0.200			
8	1	0.052	630.926	0.119			
9	1	0.026	651.016	0.149			
10	1	0.040	638.070	0.200			
11	1	0.043	628.121	0.200			
12	1	0.038	647.071	0.124			
13	1	0.050	655.567	0.265			
14	1	0.050	635.180	0.147			
15	1	0.033	658.978	0.279			
16	1	0.049	641.784	0.304			
17	1	0.059	627.410	0.200			
18	1	0.052	636.120	0.200			
19	1	0.065	635.589	0.200			
20	1	0.023	659.993	0.200			
21	1	0.052	642.236	0.140			
22	1	0.029	646.068	0.200			
23	1	0.049	651.989	0.201			
24	1	0.042	636.224	0.200			
25	1	0.071	632.449	0.230			
26	1	0.051	633.255	0.200			
27	5	0.051	29.478	31.241	32.463	33.909	35.281
28	5	0.060	35.426	37.153	38.400	39.965	41.726
29	3	0.061	35.345	37.991	40.560		

**Table 16e. 2007 Operational Item Parameter Estimates, Grade 7**

Item	Max Pts	<i>a</i> -par/ $\alpha$	<i>b</i> -par/ $\gamma_1$	<i>c</i> -par/ $\gamma_2$	$\gamma_3$
1	1	0.044	587.274	0.200	
2	1	0.026	614.903	0.200	
3	1	0.021	623.308	0.200	
4	1	0.012	614.341	0.200	
5	1	0.066	636.465	0.241	
6	1	0.056	645.332	0.336	
7	1	0.028	642.004	0.155	
8	1	0.037	644.612	0.146	
9	1	0.033	654.334	0.179	
10	1	0.023	655.022	0.200	
11	1	0.031	646.460	0.200	
12	1	0.043	634.242	0.173	
13	1	0.031	645.613	0.171	
14	1	0.050	637.266	0.203	
15	1	0.030	649.424	0.194	
16	1	0.048	620.809	0.200	
17	1	0.023	657.854	0.173	
18	1	0.024	647.663	0.200	
19	1	0.034	653.008	0.172	
20	1	0.029	656.624	0.248	
21	1	0.033	640.865	0.200	
22	1	0.040	641.904	0.258	
23	1	0.080	640.454	0.241	
24	1	0.065	642.856	0.161	
25	1	0.027	670.246	0.200	
26	1	0.058	636.966	0.194	
27	2	0.051	31.665	33.312	
28	2	0.045	28.650	29.980	
29	1	0.036	595.143	0.200	
30	1	0.035	633.159	0.200	
31	1	0.018	617.575	0.200	
32	1	0.027	631.040	0.200	
33	2	0.040	23.981	24.902	
34	2	0.033	19.032	21.030	
35	3	0.040	26.735	27.081	28.814

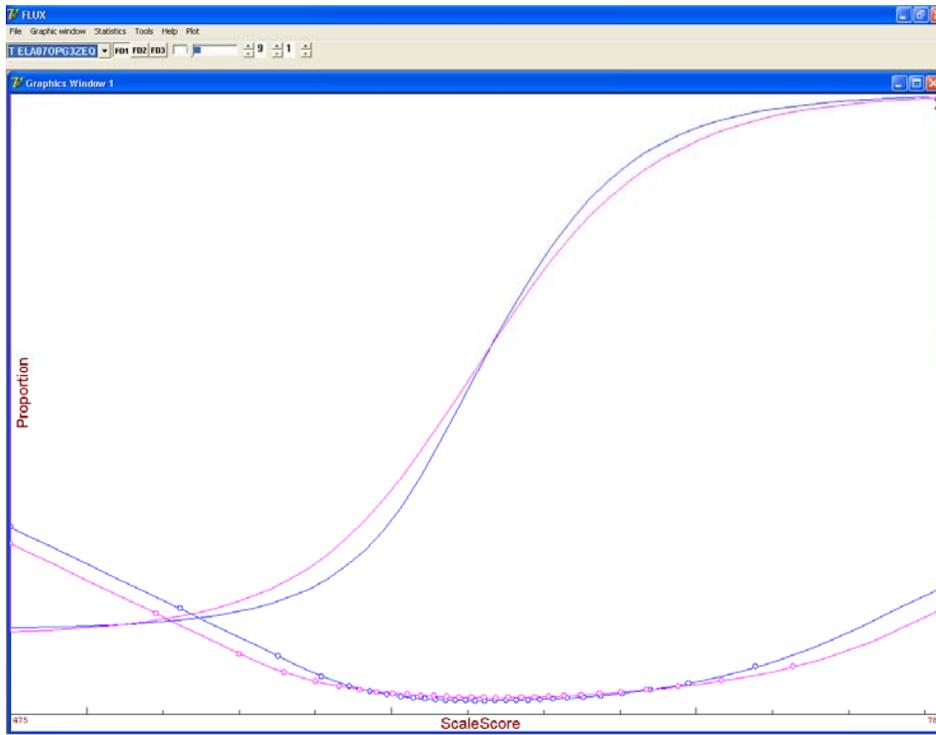
**Table 16f. 2007 Operational Item Parameter Estimates, Grade 8**

Item	Max Pts	<i>a</i> -par/ alpha	<i>b</i> -par/ gamma1	<i>c</i> -par/ gamma2	gamma3	gamma4	gamma5
1	1	0.023	613.618	0.200			
2	1	0.035	601.119	0.200			
3	1	0.032	608.333	0.384			
4	1	0.041	611.045	0.227			
5	1	0.014	596.867	0.200			
6	1	0.022	597.517	0.200			
7	1	0.022	611.113	0.200			
8	1	0.025	668.887	0.207			
9	1	0.033	622.633	0.200			
10	1	0.023	634.750	0.200			
11	1	0.031	602.023	0.200			
12	1	0.038	647.549	0.279			
13	1	0.017	617.045	0.200			
14	1	0.017	629.472	0.200			
15	1	0.021	664.865	0.424			
16	1	0.030	624.544	0.108			
17	1	0.049	630.781	0.338			
18	1	0.015	683.108	0.143			
19	1	0.028	626.805	0.110			
20	1	0.062	613.052	0.313			
21	1	0.047	629.005	0.354			
22	1	0.024	661.773	0.253			
23	1	0.016	645.753	0.159			
24	1	0.018	615.409	0.200			
25	1	0.046	633.587	0.292			
26	1	0.034	637.029	0.156			
27	5	0.052	29.386	31.048	32.339	33.937	35.441
28	5	0.054	30.163	31.693	32.845	34.601	36.307
29	3	0.054	30.722	33.019	35.783		

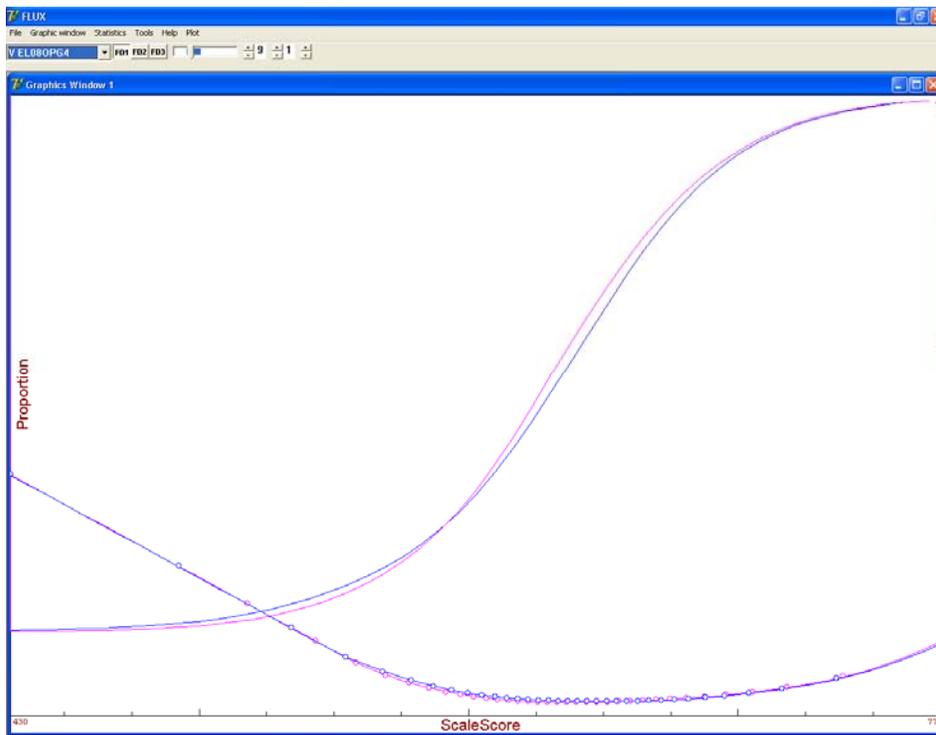
## Test Characteristic Curves

Test characteristic curves (TCCs) provide an overview of the tests in the IRT scale score metric. The 2007 and 2008 TCCs were generated using final OP item parameters for all test items administered in 2007 and 2008. 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 2007 and 2008 TCCs and SE curves are presented in Figures 7–12. Following the adoption of the chain equating method by New York State, the TCCs for new OP test forms are compared to the previous year's TCCs rather than to the baseline 2006 test form TCCs. Therefore, the 2007 OP curves are considered to be target curves for the 2008 OP test TCCs. This equating process enables the comparisons of impact results (i.e., percentages of examinees at and above each proficiency level) between adjacent test administrations. Note that in all figures the pink TCCs and SE curves represent the 2007 OP test and blue TCCs and SE curves represent the 2008 OP test.

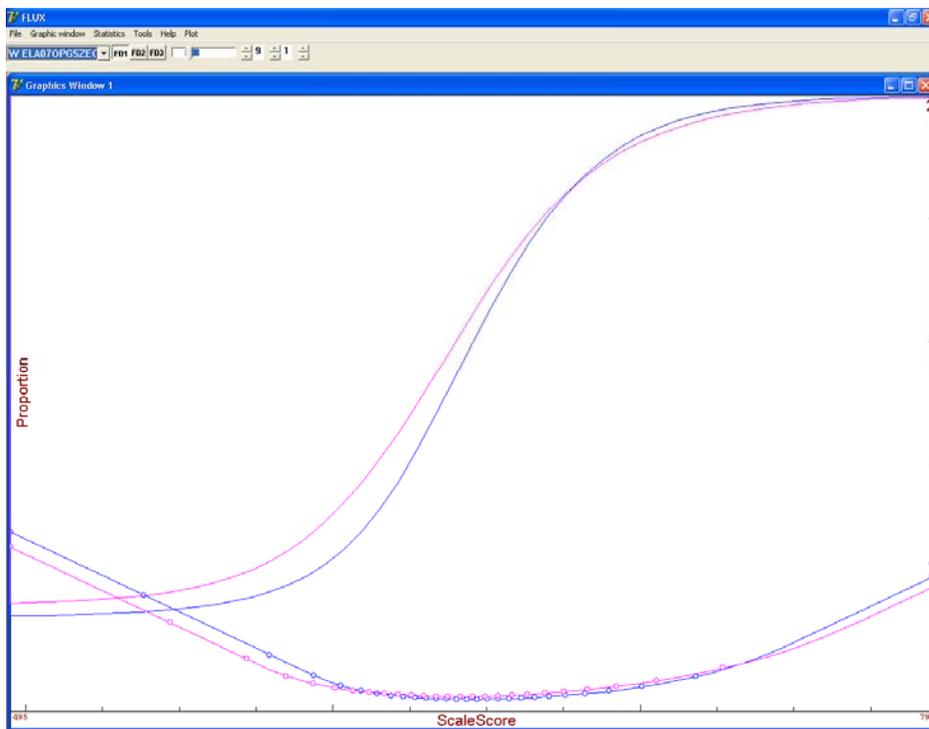
**Figure 7. Grade 3 ELA 2007 and 2008 OP TCCs and SE curves**



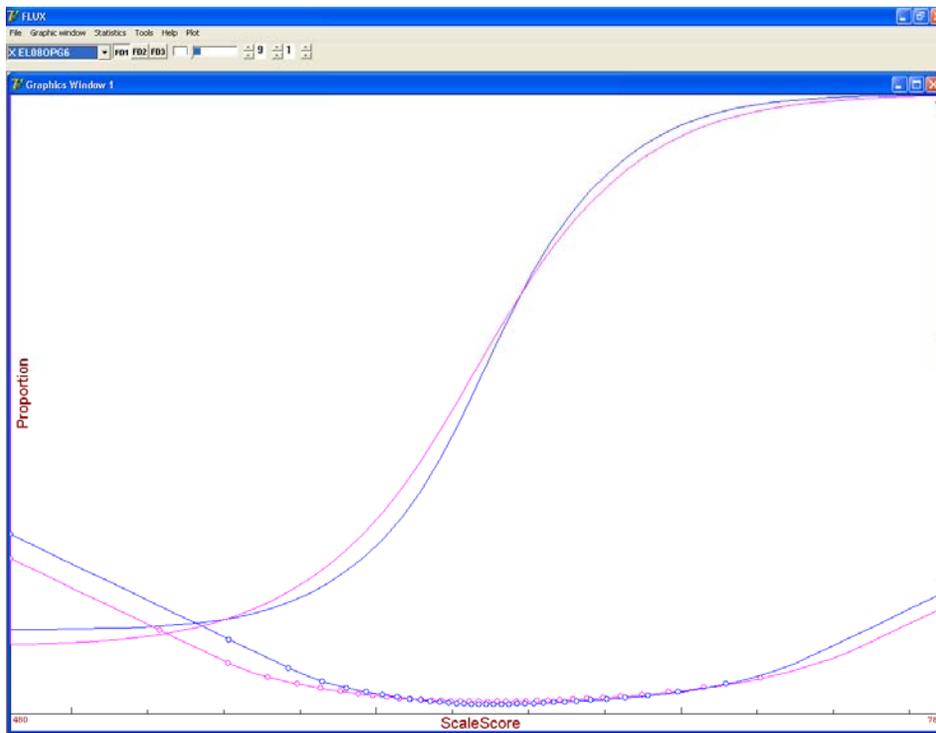
**Figure 8. Grade 4 ELA 2007 and 2008 OP TCCs and SE curves**



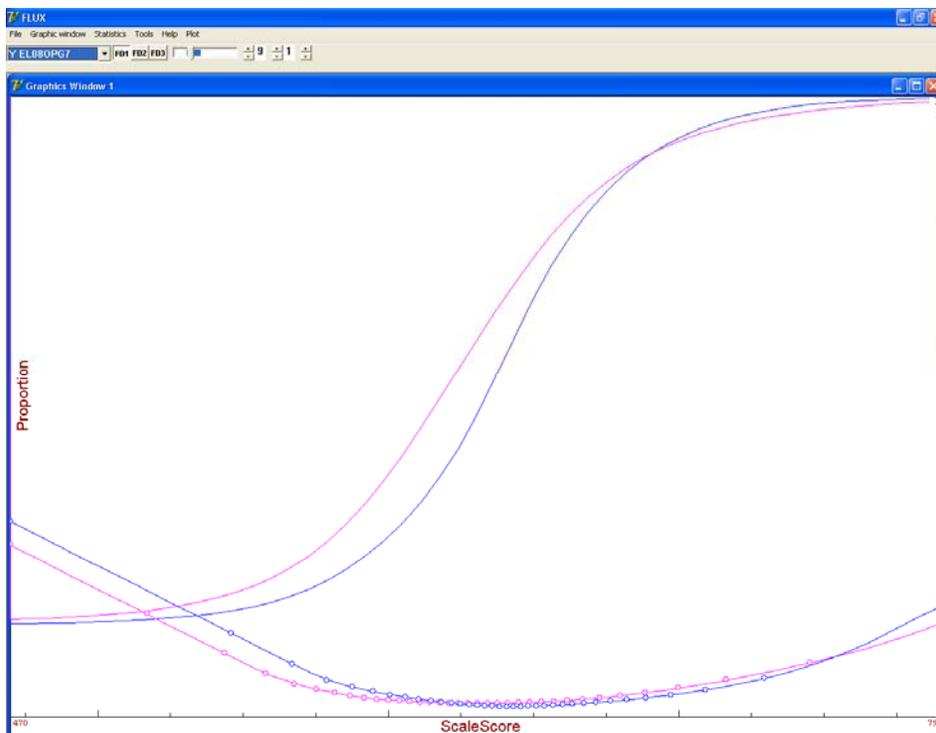
**Figure 9. Grade 5 ELA 2007 and 2008 OP TCCs and SE curves**



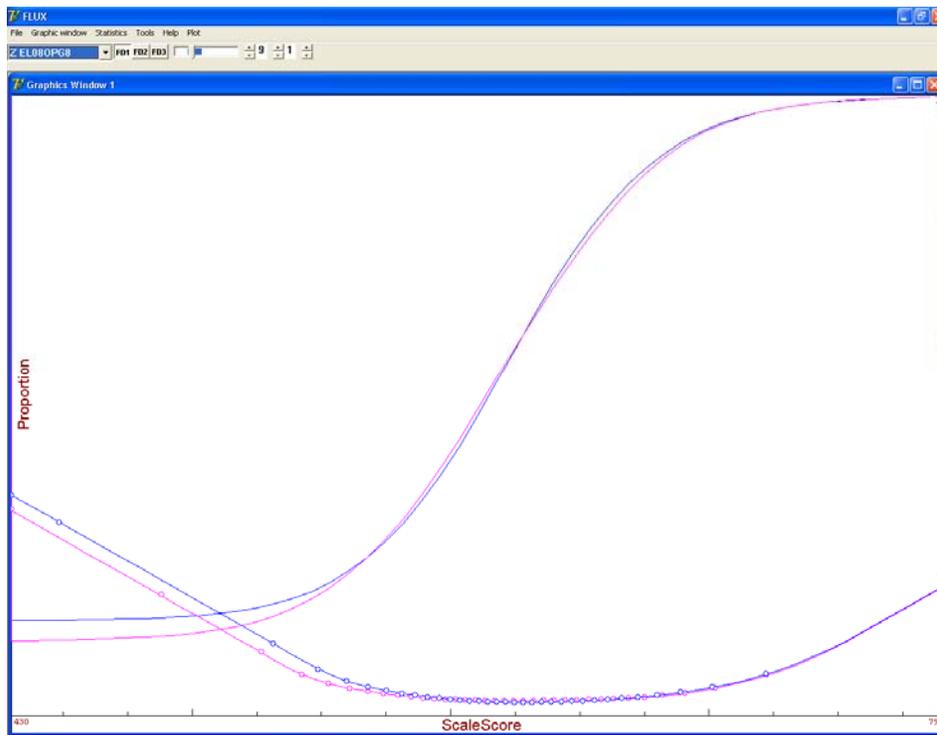
**Figure 10. Grade 6 ELA 2007 and 2008 TCCs and SE curves**



**Figure 11. Grade 7 ELA 2007 and 2007 TCCs and SE curves**



**Figure 12. Grade 8 ELA 2007 and 2008 TCCs and SE curves**



As seen in Figures 7–12, good alignments of 2007 and 2008 TCCs and SE curves were found for Grades 3, 4, 6, and 8. The TCCs for Grade 5 were somewhat less well aligned at the lower end of the scale (indicating that the 2008 form tended to be slightly more difficult for lower-ability students), and the TCCs for Grade 7 were less well aligned at the lower and middle parts of the ability scales (indicating that the 2008 test form tended to be slightly more difficult for lower and middle-ability 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 weight specified in a scoring process ( $v_i=1$  if no weights are specified).

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

### **Weighting Constructed-Response Items in Grades 4 and 8**

Consistently with 2006 and 2007 scoring procedures, a weight factor of 1.38 was applied to all CR items in Grades 4 and 8. The CR items were weighted in order to align proportions of raw score points obtainable from MC and CR items on 2008 with past ELA Grade 4 and 8 tests. Weighting CR items in Grades 4 and 8 had no substantial effect on the coverage of content standards in the test blueprint.

The inverse TCC scoring method was extended to incorporate weights for CR items for Grades 4 and 8 and weights of 1.38 were specified for these items. It should be noted that when weights are applied, the statistical characteristics of the trait estimates (i.e., bias and standard errors) will depend on the weights that are specified and the statistical characteristics of the items.

### ***Raw Score-to-Scale Score and SEM Conversion Tables***

The scale score (SS) is the basic score for the New York State tests. It is used to derive other scores that describe test performance, such as the four performance levels and 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 17. Grade 3 Raw Score to Scale Score (with Standard Error)**

Raw Score	Scale Score	Standard Error
0	475	128
1	475	128
2	475	128
3	475	128
4	475	128
5	531	72
6	563	40
7	577	26
8	586	19
9	593	16
10	599	13
11	603	12
12	607	11
13	611	10
14	614	10
15	618	10
16	621	9
17	624	9
18	627	9
19	631	9
20	634	9
21	638	9
22	641	10
23	645	10
24	649	10
25	653	10
26	658	11

*(Continued on next page)*

**Table 17. Grade 3 Raw Score to Scale Score (with Standard Error) (cont.)**

Raw Score	Scale Score	Standard Error
27	663	12
28	669	13
29	676	14
30	685	17
31	697	21
32	720	32
33	780	93

**Table 18. Grade 4 Raw Score to Scale Score (with Standard Error)**

Weighted Raw Score	Scale Score	Standard Error
0	430	146
1	430	146
2	430	146
3	430	146
4	430	146
5	430	146
6	496	80
7	525	50
8	542	37
9	555	30
10	566	26
11	574	22
12	582	20
13	589	18
14	595	16
15	600	15
16	605	14
17	609	13
18	613	12
19	617	12
20	621	11
21	625	11
22	629	11
23	632	10
24	636	10
25	639	10
26	643	10
27	646	10

*(Continued on next page)*

**Table 18. Grade 4 Raw Score to Scale Score (with Standard Error) (cont.)**

Weighted Raw Score	Scale Score	Standard Error
28	650	10
29	653	10
30	657	10
31	660	10
32	664	10
33	668	11
34	673	11
35	677	11
36	682	12
37	688	13
38	694	14
39	702	15
40	711	18
41	723	21
42	742	28
43	775	51

**Table 19. Grade 5 Raw Score to Scale Score (with Standard Error)**

Raw Score	Scale Score	Standard Error
0	495	123
1	495	123
2	495	123
3	495	123
4	495	123
5	538	80
6	579	39
7	594	24
8	603	17
9	609	14
10	614	12
11	619	11
12	623	10
13	627	9
14	630	9
15	634	9
16	637	8
17	640	8
18	644	8
19	647	8
20	650	8

*(Continued on next page)*

**Table 19. Grade 5 Raw Score to Scale Score (with Standard Error) (cont.)**

Raw Score	Scale Score	Standard Error
21	654	8
22	657	9
23	661	9
24	666	9
25	670	10
26	676	11
27	682	12
28	690	14
29	701	17
30	718	24
31	795	101

**Table 20. Grade 6 Raw Score to Scale Score (with Standard Error)**

Raw Score	Scale Score	Standard Error
0	480	122
1	480	122
2	480	122
3	480	122
4	480	122
5	480	122
6	552	51
7	571	31
8	582	22
9	590	18
10	597	15
11	602	13
12	607	11
13	611	10
14	615	9
15	618	9
16	621	8
17	624	8
18	626	7
19	629	7
20	631	7
21	634	7
22	636	7
23	639	7
24	641	7
25	644	7

*(Continued on next page)*

**Table 20. Grade 6 Raw Score to Scale Score (with Standard Error) (cont.)**

Raw Score	Scale Score	Standard Error
26	646	7
27	649	7
28	652	7
29	655	7
30	658	8
31	662	8
32	666	9
33	670	9
34	676	10
35	682	11
36	689	13
37	699	15
38	715	21
39	785	91

**Table 21. Grade 7 Raw Score to Scale Score (with Standard Error)**

Raw Score	Scale Score	Standard Error
0	470	133
1	470	133
2	470	133
3	470	133
4	470	133
5	470	133
6	470	133
7	546	57
8	567	36
9	579	25
10	588	20
11	595	17
12	601	15
13	606	14
14	611	12
15	615	11
16	618	10
17	622	9
18	625	9
19	628	8
20	631	8
21	633	8
22	636	7

*(Continued on next page)*

**Table 21. Grade 7 Raw Score to Scale Score (with Standard Error) (cont.)**

Raw Score	Scale Score	Standard Error
23	638	7
24	641	7
25	643	7
26	646	7
27	648	7
28	651	7
29	654	8
30	657	8
31	660	8
32	664	9
33	667	9
34	672	10
35	676	11
36	682	12
37	689	13
38	697	15
39	709	19
40	729	27
41	790	83

**Table 22. Grade 8 Raw Score to Scale Score (with Standard Error)**

Weighted Raw Score	Scale Score	Standard Error
0	430	145
1	430	145
2	430	145
3	430	145
4	430	145
5	430	145
6	447	128
7	527	48
8	544	31
9	555	23
10	563	20
11	569	17
12	575	16
13	580	14
14	585	13
15	589	12
16	593	12
17	597	11

*(Continued on next page)*

**Table 22. Grade 8 Raw Score to Scale Score (with Standard Error) (cont.)**

Weighted Raw Score	Scale Score	Standard Error
18	600	11
19	604	10
20	607	10
21	610	10
22	613	9
23	616	9
24	619	9
25	622	9
26	625	9
27	629	9
28	632	9
29	635	9
30	638	9
31	642	9
32	645	10
33	649	10
34	653	10
35	657	11
36	661	11
37	666	12
38	671	12
39	677	13
40	684	15
41	693	17
42	705	20
43	726	30
44	790	94

***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 2008 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 (scale score of 650 for all grades). Table 23 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 23. SPI Target Ranges**

Grade	Objective	# Items	Total Points	Level III Cut SPI Target Range
3	1	9	10	60–77
	2	13	15	69–81
	3	5	5	59–76
4	1	12	12	56–72
	2	13	16	60–73
	3	5	8	59–71
5	1	12	13	63–76
	2	10	10	63–78
	3	4	5	51–68
6	1	11	11	63–79
	2	12	16	67–77
	3	5	9	59–72
7	1	16	17	71–81
	2	12	13	58–74
	3	6	8	61–74
8	1	9	13	72–84
	2	14	14	74–84
	3	5	9	61–75

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 of greatest value to improving student achievement. Instruction focused on the identified needs of students has the best chance of helping those students increase their skills in the areas measured by the test. SPI reports provide students, parents, and educators the opportunity to identify and target specific areas within the broader content domain 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 ( $\theta$ ) for each examinee were estimated for the three-parameter logistic model or the two-parameter partial credit model in the case of CR items. The item parameters were based on data from the total population of examinees. Then the population was divided into NRC, gender, or ethnic groups, and the members in each group are sorted into 10 equal score categories (deciles) based upon their location on the scale score ( $\theta$ ) 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

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

where

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

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

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

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,

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

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 given by:

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

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

$$D_i = O_{i\cdot} - P_i.$$

These indices are indicators of the degree to which members of a specific subgroup perform better or worse than expected on each item. Differences for decile groups provide an index for each of the ten regions on the scale score ( $\theta$ ) 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), and Low Needs districts (by NRC code). Applying the Linn-Harnisch method revealed that no items were flagged for DIF on the Grade 3, 6 and 7 tests; one item was flagged on the Grade 4 test and one item was flagged on Grade 5 test; and two items were flagged on the Grade 8 test, as is shown in Table 24. As indicated in the classical DIF analysis section, items flagged for DIF do not necessarily display bias.

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

**Table 24. Number of Items Flagged for DIF by the Linn-Harnisch Method**

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

## Section VII: Reliability and Standard Error of Measurement

---

This section presents specific information on various test reliability statistics (RS) and standard errors of measurement (SEM), as well as the results from a study of performance level classification accuracy and consistency. The dataset 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 25 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 25. 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	195695	28	33	0.86	2.18	0.87	2.12
4	196915	31	39	0.90	2.39	0.90	2.29
5	197901	27	31	0.84	2.18	0.85	2.10
6	200352	29	39	0.88	2.41	0.90	2.21
7	206871	35	41	0.88	2.50	0.88	2.43
8	208959	29	39	0.86	2.45	0.88	2.24

All the coefficients for total test reliability are in the range of 0.84–0.90, 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 26 presents reliabilities for the MC subsets.

**Table 26 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	195695	24	0.84	1.74	0.85	1.72
4	196915	28	0.88	2.01	0.88	2.00
5	197901	24	0.82	1.78	0.82	1.77
6	200352	26	0.86	1.82	0.86	1.80
7	206871	30	0.85	2.04	0.86	2.03
8	208959	26	0.83	1.88	0.83	1.88

### Reliability of CR Items

Reliability coefficients were also computed for the subsets of CR items. It should be noted that the Grades 3–8 ELA Tests include only three to five CR items, depending on grade level, and the results presented in Table 27 should be interpreted with caution.

**Table 27 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	195695	4	9	0.57	1.23	0.58	1.21
4	196915	3	11	0.78	0.99	0.78	0.98
5	197901	3	7	0.51	1.19	0.56	1.12
6	200352	3	13	0.78	1.20	0.80	1.13
7	206871	5	11	0.66	1.34	0.68	1.31
8	208959	3	13	0.82	1.11	0.84	1.03

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), limited English proficiency (LEP) status, all students with disabilities (SWD), and all students using test accommodations (SUA). As shown in Tables 28a–28f, 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 or equal to 0.80, with the exception of Grade 5 NRC = 6 (Low Needs districts). Feldt-Raju reliability coefficients, which tend to be larger than the Cronbach alpha estimates for the same group, were all larger than or equal to 0.80 with the exception of Grade 5

NRC=6 (Low Needs districts). All other test reliability alpha statistics were in the 0.81–0.92 range, indicating very good test internal consistency (reliability) for analyzed subgroups of examinees.

**Table 28a. 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	195695	0.86	2.18	0.87	2.12
Gender	Female	95081	0.85	2.11	0.85	2.06
	Male	100614	0.87	2.23	0.88	2.16
Ethnicity	Asian	14302	0.83	1.99	0.84	1.95
	Black	37585	0.85	2.38	0.86	2.31
	Hispanic	41422	0.86	2.40	0.87	2.33
	American Indian	978	0.83	2.34	0.84	2.29
	Multi-Racial	233	0.81	2.09	0.82	2.05
	White	101110	0.84	2.00	0.84	1.95
	Unknown	65	0.84	1.92	0.85	1.85
NRC	New York City	69150	0.87	2.32	0.87	2.25
	Big 4 Cites	8103	0.86	2.47	0.86	2.39
	High Needs Urban/Suburban	15830	0.85	2.28	0.86	2.22
	High Needs Rural	11469	0.85	2.21	0.85	2.15
	Average Needs	58520	0.83	2.05	0.84	2.00
	Low Needs	29390	0.80	1.84	0.80	1.80
	Charter	3023	0.80	2.20	0.80	2.16
SWD	All Codes	26011	0.87	2.65	0.88	2.55
SUA	All Codes	39019	0.86	2.63	0.87	2.53
LEP	LEP = Y	16568	0.84	2.59	0.85	2.51

**Table 28b. 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	196915	0.90	2.39	0.90	2.29
Gender	Female	96399	0.89	2.33	0.90	2.24
	Male	100516	0.90	2.43	0.91	2.33
Ethnicity	Asian	14174	0.88	2.23	0.89	2.14
	Black	37975	0.88	2.55	0.89	2.46
	Hispanic	41178	0.88	2.56	0.89	2.46
	American Indian	952	0.90	2.49	0.90	2.39
	Multi-Racial	188	0.89	2.35	0.90	2.26
	White	102361	0.89	2.24	0.90	2.15
	Unknown	87	0.85	2.26	0.86	2.20
	NRC	New York City	69692	0.89	2.52	0.90
Big 4 Cites		7697	0.89	2.59	0.90	2.50
High Needs Urban/Suburban		15655	0.89	2.46	0.90	2.37
High Needs Rural		11490	0.89	2.40	0.90	2.32
Average Needs		59451	0.89	2.27	0.89	2.19
Low Needs		30278	0.86	2.09	0.87	2.01
Charter		2406	0.86	2.46	0.87	2.40
SWD	All Codes	28885	0.88	2.68	0.89	2.60
SUA	All Codes	41079	0.88	2.68	0.88	2.59
LEP	LEP = Y	14008	0.85	2.70	0.86	2.61

**Table 28c. 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	197901	0.84	2.18	0.85	2.10
Gender	Female	96664	0.83	2.17	0.84	2.07
	Male	101237	0.85	2.19	0.86	2.11
Ethnicity	Asian	14572	0.83	2.05	0.84	1.96
	Black	38296	0.83	2.29	0.83	2.24
	Hispanic	40577	0.84	2.30	0.85	2.24
	American Indian	902	0.83	2.25	0.84	2.19
	Multi-Racial	168	0.83	2.19	0.84	2.10
	White	103311	0.81	2.07	0.83	1.98
	Unknown	75	0.82	2.06	0.84	1.94

*(Continued on next page)*

**Table 28c. 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	69157	0.84	2.27	0.85	2.19
	Big 4 Cites	7525	0.84	2.33	0.85	2.29
	High Needs Urban/Suburban	15199	0.84	2.24	0.85	2.18
	High Needs Rural	11366	0.82	2.18	0.84	2.12
	Average Needs	60375	0.81	2.10	0.82	2.02
	Low Needs	30708	0.76	1.94	0.78	1.87
	Charter	3309	0.80	2.27	0.81	2.21
SWD	All Codes	29635	0.84	2.41	0.85	2.38
SUA	All Codes	40120	0.84	2.40	0.85	2.37
LEP	LEP = Y	11372	0.82	2.43	0.82	2.40

**Table 28d. 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	200352	0.88	2.41	0.90	2.21
Gender	Female	97505	0.87	2.36	0.89	2.16
	Male	102847	0.88	2.44	0.90	2.24
Ethnicity	Asian	14521	0.87	2.29	0.89	2.08
	Black	37970	0.87	2.56	0.88	2.40
	Hispanic	40576	0.87	2.59	0.89	2.41
	American Indian	914	0.88	2.53	0.90	2.34
	Multi-Racial	160	0.86	2.27	0.88	2.09
	White	106129	0.86	2.23	0.88	2.05
	Unknown	82	0.89	2.35	0.92	2.03
NRC	New York City	69102	0.88	2.57	0.89	2.37
	Big 4 Cites	7544	0.88	2.60	0.89	2.43
	High Needs Urban/Suburban	15305	0.87	2.46	0.89	2.29
	High Needs Rural	11776	0.86	2.36	0.88	2.19
	Average Needs	61976	0.85	2.25	0.87	2.09
	Low Needs	31556	0.83	2.07	0.85	1.91
	Charter	2768	0.85	2.38	0.86	2.27
SWD	All Codes	30391	0.87	2.70	0.88	2.56
SUA	All Codes	38535	0.87	2.70	0.88	2.56
LEP	LEP = Y	9816	0.85	2.77	0.86	2.62

**Table 28e. 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	206871	0.88	2.50	0.88	2.43
Gender	Female	101183	0.86	2.45	0.87	2.38
	Male	105688	0.88	2.55	0.89	2.48
Ethnicity	Asian	14498	0.87	2.35	0.88	2.27
	Black	39865	0.86	2.66	0.87	2.61
	Hispanic	41370	0.87	2.66	0.88	2.60
	American Indian	1020	0.86	2.62	0.87	2.57
	Multi-Racial	128	0.87	2.43	0.88	2.36
	White	109922	0.86	2.37	0.86	2.30
	Unknown	68	0.88	2.43	0.89	2.35
	NRC	New York City	71310	0.88	2.62	0.88
Big 4 Cites		7855	0.88	2.71	0.88	2.66
High Needs Urban/Suburban		15718	0.87	2.58	0.87	2.53
High Needs Rural		12493	0.86	2.51	0.87	2.45
Average Needs		64994	0.85	2.40	0.86	2.34
Low Needs		31790	0.82	2.23	0.83	2.17
Charter		2294	0.83	2.58	0.83	2.52
SWD	All Codes	30308	0.87	2.78	0.87	2.74
SUA	All Codes	38066	0.87	2.78	0.87	2.74
LEP	LEP = Y	9245	0.84	2.84	0.85	2.80

**Table 28f. 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	208959	0.86	2.45	0.88	2.24
Gender	Female	101842	0.85	2.35	0.88	2.17
	Male	107117	0.86	2.51	0.89	2.30
Ethnicity	Asian	14272	0.86	2.34	0.89	2.12
	Black	40399	0.84	2.60	0.86	2.43
	Hispanic	40911	0.86	2.64	0.88	2.44
	American Indian	1059	0.86	2.56	0.88	2.38
	Multi-Racial	117	0.85	2.25	0.87	2.06
	White	112147	0.84	2.27	0.87	2.09
	Unknown	54	0.87	2.29	0.90	1.98

*(Continued on next page)*

**Table 28f. Grade 8 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	71488	0.86	2.61	0.88	2.40
	Big 4 Cites	8337	0.86	2.69	0.88	2.50
	High Needs Urban/Suburban	15970	0.86	2.52	0.88	2.33
NRC	High Needs Rural	13148	0.85	2.42	0.87	2.23
	Average Needs	66177	0.84	2.29	0.86	2.12
	Low Needs	31770	0.81	2.06	0.83	1.92
	Charter	1423	0.81	2.49	0.83	2.35
SWD	All Codes	30218	0.85	2.77	0.86	2.61
SUA	All Codes	38035	0.85	2.79	0.87	2.61
LEP	LEP = Y	8618	0.83	2.89	0.85	2.70

### ***Standard Error of Measurement***

The standard errors of measurement (SEM), as computed from Cronbach's alpha and the Feldt-Raju reliability statistics, are presented in Table 25. SEMs ranged 2.10–2.50, which is reasonable and small. In other words, the error of measurement from the observed test score ranged from approximately +/- 2–3 raw score points. SEMs are directly related to reliability: the higher the reliability, the lower the standard error. As discussed, the reliability of these tests is relatively high, so it was expected that the SEMs would be very low.

The SEMs for subpopulations, as computed from Cronbach's alpha and the Feldt-Raju reliability statistics, are presented in Tables 28a–28f. The SEMs associated with all reliability estimates for all subpopulations are in the range of 1.80–2.89, 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 I.

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 29 and 30 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 29 depicts the consistency study results based on the range of performance levels for all grades. Overall, between 71% and 82% 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 0.54–0.65.

**Table 29. Decision Consistency (All Cuts)**

Grade	N-count	Agreement	Inconsistency	Kappa
3	195695	0.7083	0.2917	0.5374
4	196915	0.7676	0.2324	0.5991
5	197901	0.7801	0.2199	0.5583
6	200352	0.8080	0.1920	0.6451
7	206871	0.8168	0.1832	0.6322
8	208959	0.7502	0.2498	0.5942

Table 30 depicts the consistency study results based on two performance levels (passing and not passing) as defined by the Level III cut. Overall, about 85%–90% of the classifications of individual students are estimated to remain stable with a second administration. Kappa coefficients for classification consistency based on one cut ranged 0.70–0.75.

**Table 30. Decision Consistency (Level III Cut)**

Grade	N-count	Agreement	Inconsistency	Kappa
3	195695	0.8723	0.1277	0.6983
4	196915	0.8980	0.1020	0.7532
5	197901	0.8890	0.1110	0.6903
6	200352	0.8833	0.1167	0.7373
7	206871	0.8845	0.1155	0.7283
8	208959	0.8535	0.1465	0.7021

### Accuracy

The results of classification accuracy are presented in Table 31. Included in the table are case counts (N-count), classification accuracy (Accuracy) for all performance levels (All Cuts) and for the Level III cut score 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 for the true variable to be located in, instead of four. The accuracy rates indicate that the categorization of a student’s observed performance is in agreement with the location of their true ability approximately 78%–86% of the time across all performance levels and approximately 89%–93% of the time in regards to the Level III cut score.

**Table 31. 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	195695	<b>0.7804</b>	0.1575	0.0622	<b>0.9080</b>	0.0500	0.0420
4	196915	<b>0.8268</b>	0.1222	0.0510	<b>0.9252</b>	0.0466	0.0282
5	197901	<b>0.8289</b>	0.1304	0.0408	<b>0.9172</b>	0.0511	0.0317
6	200352	<b>0.8592</b>	0.0965	0.0442	<b>0.9150</b>	0.0508	0.0341
7	206871	<b>0.8642</b>	0.0902	0.0457	<b>0.9173</b>	0.0450	0.0377
8	208959	<b>0.8115</b>	0.1369	0.0516	<b>0.8915</b>	0.0702	0.0383

## Section VIII: Summary of Operational Test Results

---

This section summarizes the distribution of OP scale score results on the New York State 2008 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), limited English proficiency (LEP), 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. Data include examinees with valid scores from all public and charter schools. Note that complete scale score frequency distribution tables are located in Appendix I.

### *Scale Score Distribution Summary*

Scale score distribution summary tables are presented and discussed in Tables 32–38. In Table 32, scale score statistics for total populations of students from public and charter schools are presented. In Tables 33–38, 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 LEP, SWD and/or SUA achieved below the State aggregate (All Students) in every percentile. This pattern of achievement was consistent across all grades.

**Table 32. ELA Grades 3–8 Scale Score Distribution Summary**

Grade	N-count	SS Mean	SS Std Dev	10 <sup>th</sup> %tile	25 <sup>th</sup> %tile	50 <sup>th</sup> %tile	75 <sup>th</sup> %tile	90 <sup>th</sup> %tile
3	195695	668.87	39.50	627	645	669	685	720
4	196915	666.19	40.18	617	643	668	688	711
5	197901	667.22	31.02	634	650	666	682	701
6	200352	661.29	30.22	629	644	662	676	689
7	206871	662.09	29.54	631	646	664	676	697
8	208959	656.91	37.96	616	635	657	677	705

### **Grade 3**

Scale score statistics and N-counts of demographic groups for Grade 3 are presented in Table 33. The population scale score mean was 668.87 with a standard deviation of 39.50. By gender subgroup, Females outperformed Males, but the difference was less than six 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 White ethnic group had the highest average scale score mean (678.80). 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 standard deviation below the population mean. SWD, SUA, and LEP subgroups scored, on average, approximately three-

fourths of one standard deviation below the mean scale score for the population. The SWD subgroup, which had a scale score mean about 35 scale score units below the population mean, was the lowest performing group analyzed. At the 50<sup>th</sup> percentile, the following groups exceeded the population score of 669: Asian (676), White (676), and Low Needs districts (685).

**Table 33. Scale Score Distribution Summary, by Subgroup, Grade 3**

Demographic Category (Subgroup)		N-count	SS Mean	SS Std Dev	10 <sup>th</sup> %tile	25 <sup>th</sup> %tile	50 <sup>th</sup> %tile	75 <sup>th</sup> %tile	90 <sup>th</sup> %tile
State	All Students	195695	668.87	39.50	627	645	669	685	720
Gender	Female	95081	671.60	38.61	631	649	669	685	720
	Male	100614	666.29	40.16	621	645	663	685	720
Ethnicity	Asian	14302	677.89	38.64	638	658	676	697	720
	Black	37585	655.96	35.67	618	638	653	676	697
	Hispanic	41422	653.46	36.38	614	634	653	669	697
	American Indian	978	657.76	32.18	621	638	658	676	697
	Multi-Racial	233	674.14	36.64	634	653	669	685	720
	White	101110	678.80	38.85	638	658	676	697	720
	Unknown	65	678.02	34.94	641	663	676	685	720
NRC	New York City	69150	659.23	38.76	618	638	658	676	697
	Big 4 Cities	8103	650.50	35.69	611	631	649	669	685
	High Needs Urban/Suburban	15830	662.54	36.14	621	641	663	685	697
	High Needs Rural	11469	665.89	36.42	627	645	663	685	697
	Average Needs	58520	675.89	37.85	634	653	669	697	720
	Low Needs	29390	687.89	38.62	649	663	685	697	720
	Charter	3023	666.07	32.10	631	645	663	685	697
LEP	LEP = Y	16568	638.05	32.67	603	624	641	658	669
SWD	All Codes	26011	634.24	38.74	593	614	638	658	676
SUA	All Codes	39019	637.06	36.05	599	618	638	658	676

#### Grade 4

Scale score statistics and N-counts of demographic groups for Grade 4 are presented in Table 34. The Grade 4 population (All Students) mean was 666.19, with a standard deviation of 40.18. By gender subgroup, Females outperformed Males, but the difference was less than twelve 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 (680.23). 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 standard deviation below the population mean. The SWD subgroup had a scale score mean nearly 40 scale score units below the population mean and were at or below the scale score of any given percentile for

any other subgroup. At the 50<sup>th</sup> percentile, the following groups exceeded the population score of 668: Female (673), Asian (677), White (677), Average Needs districts (673), and Low Needs districts (688).

**Table 34. Scale Score Distribution Summary, by Subgroup, Grade 4**

Demographic Category (Subgroup)		N-count	SS Mean	SS Std Dev	10 <sup>th</sup> %tile	25 <sup>th</sup> %tile	50 <sup>th</sup> %tile	75 <sup>th</sup> %tile	90 <sup>th</sup> %tile
State	All Students	196915	666.19	40.18	617	643	668	688	711
Gender	Female	96399	671.61	38.94	625	650	673	694	723
	Male	100516	660.98	40.66	613	639	664	688	702
Ethnicity	Asian	14174	680.23	38.33	636	657	677	702	723
	Black	37975	651.36	38.09	609	629	653	673	694
	Hispanic	41178	651.25	38.51	605	629	653	673	694
	American Indian	952	655.29	39.02	605	636	657	682	702
	Multi-Racial	188	666.26	36.55	621	643	668	688	711
	White	102361	675.84	38.17	632	653	677	694	723
	Unknown	87	677.77	33.83	636	657	673	694	723
NRC	New York City	69692	656.46	40.68	609	632	657	682	702
	Big 4 Cities	7697	646.85	39.77	600	625	650	673	694
	High Needs Urban/Suburban	15655	658.99	38.52	613	639	660	682	702
	High Needs Rural	11490	662.54	37.38	617	643	664	688	702
	Average Needs	59451	672.96	37.13	629	653	673	694	711
	Low Needs	30278	686.50	35.59	646	668	688	702	742
	Charter	2406	659.23	31.91	621	639	660	677	702
LEP	LEP = Y	14008	630.52	35.69	589	613	636	653	668
SWD	All Codes	28885	625.21	42.60	574	605	629	653	673
SUA	All Codes	41079	629.85	40.31	582	609	636	657	673

## Grade 5

Scale score summary statistics for Grade 5 students are in Table 35. Overall, the scale score mean was 667.22, with a standard deviation of 31.02. The difference between mean scale scores by gender groups was very small (about four 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 a half of standard deviation below the population mean. SWD, SUA, and LEP subgroups scored approximately one standard deviation below the mean scale score for the population. The LEP subgroup, which had a scale score mean nearly thirty-one scale score units (one standard deviation) below the population mean, was the lowest performing group analyzed.

At the 50<sup>th</sup> percentile, the following groups exceeded the population score of 666: Asian (676), White (676), Average Needs districts (670), and Low Needs districts (676).

**Table 35. Scale Score Distribution Summary, by Subgroup, Grade 5**

Demographic Category (Subgroup)		N-count	SS Mean	SS Std Dev	10 <sup>th</sup> %tile	25 <sup>th</sup> %tile	50 <sup>th</sup> %tile	75 <sup>th</sup> %tile	90 <sup>th</sup> %tile
State	All Students	197901	667.22	31.02	634	650	666	682	701
Gender	Female	96664	669.29	30.98	637	654	666	682	701
	Male	101237	665.24	30.93	634	650	666	682	701
Ethnicity	Asian	14572	676.47	33.23	644	657	676	690	718
	Black	38296	655.96	27.27	627	644	657	670	682
	Hispanic	40577	656.19	29.19	627	644	657	670	690
	American Indian	902	659.15	28.14	630	644	657	676	690
	Multi-Racial	168	669.48	28.91	637	650	666	690	701
	White	103311	674.47	30.13	644	657	676	690	701
	Unknown	75	676.53	34.26	640	657	670	690	718
	NRC	New York City	69157	660.50	31.32	627	644	661	676
	Big 4 Cities	7525	651.76	29.47	619	637	654	666	682
	High Needs Urban/Suburban	15199	661.08	29.09	630	647	661	676	690
	High Needs Rural	11366	665.24	28.09	634	650	666	682	690
	Average Needs	60375	672.08	28.92	644	657	670	690	701
	Low Needs	30708	681.53	29.84	654	666	676	690	718
	Charter	3309	659.35	25.05	630	644	657	676	690
LEP	LEP = Y	11372	636.27	29.59	609	623	640	654	666
SWD	All Codes	29635	640.24	30.83	609	627	644	657	670
SUA	All Codes	40120	641.58	30.10	609	627	644	657	670

## Grade 6

Scale score summary statistics for Grade 6 students are in Table 36. The scale score mean was 661.29, with a standard deviation of 30.22. 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 standard deviation below the population mean. SWD, SUA, and LEP subgroups scored over one standard deviation below the mean scale score for the population. The LEP subgroup, which had a scale score mean slightly more than thirty-two scale score units below the population mean, was the lowest performing group analyzed. At the 50<sup>th</sup> percentile, the following groups exceeded the population score of 662: Asian (670), White (666), Average Needs districts (666), and Low Needs districts (676).

**Table 36. Scale Score Distribution Summary, by Subgroup, Grade 6**

Demographic Category (Subgroup)		N-count	SS Mean	SS Std Dev	10 <sup>th</sup> %tile	25 <sup>th</sup> %tile	50 <sup>th</sup> %tile	75 <sup>th</sup> %tile	90 <sup>th</sup> %tile
State	All Students	200352	661.29	30.22	629	644	662	676	689
Gender	Female	97505	665.06	30.41	634	649	662	682	699
	Male	102847	657.71	29.60	626	641	658	676	689
Ethnicity	Asian	14521	669.96	31.87	636	652	670	682	699
	Black	37970	649.90	25.79	621	636	652	666	676
	Hispanic	40576	648.53	27.56	618	634	649	666	676
	American Indian	914	653.12	29.36	621	639	655	666	682
	Multi-Racial	160	667.18	31.18	634	649	662	682	699
	White	106129	669.10	29.43	639	652	666	682	699
	Unknown	82	671.01	39.25	629	646	664	689	715
NRC	New York City	69102	651.80	28.78	621	636	652	666	682
	Big 4 Cities	7544	648.51	27.41	618	634	649	666	676
	High Needs Urban/Suburban	15305	655.66	27.30	626	641	655	670	682
	High Needs Rural	11776	660.44	26.89	631	646	658	676	689
	Average Needs	61976	667.16	28.12	639	652	666	682	699
	Low Needs	31556	677.63	30.68	646	662	676	689	715
	Charter	2768	655.57	22.28	629	641	655	670	682
LEP	LEP = Y	9816	627.34	28.47	597	615	631	644	655
SWD	All Codes	30391	633.18	27.48	602	621	636	649	662
SUA	All Codes	38535	634.39	27.90	602	621	636	652	662

**Grade 7**

Scale score statistics and N-counts of demographic groups for Grade 7 are presented in Table 37. The population scale score mean was 662.09 and the population standard deviation was 29.54. By gender subgroup, Females outperformed Males, but the difference was about one quarter 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 (671.98). The students from the Big 4 Cities scored below their peers from schools with other NRC designations and about a half of standard deviation below the population mean. SWD and SUA subgroups scored approximately one standard deviation below the mean scale score for the population. The LEP subgroup, which had a scale score mean slightly more than thirty-five scale score units below the population mean, was the lowest performing group analyzed. At the 50<sup>th</sup> percentile, the following groups exceeded the population score of 664: Female (667), Asian (672), White (667), Average Needs districts (667), and Low Needs districts (676).

**Table 37. Scale Score Distribution Summary, by Subgroup, Grade 7**

Demographic Category (Subgroup)		N-count	SS Mean	SS Std Dev	10 <sup>th</sup> %tile	25 <sup>th</sup> %tile	50 <sup>th</sup> %tile	75 <sup>th</sup> %tile	90 <sup>th</sup> %tile
State	All Students	206871	662.09	29.54	631	646	664	676	697
Gender	Female	101183	666.28	28.64	636	651	667	682	697
	Male	105688	658.08	29.83	625	643	660	676	689
Ethnicity	Asian	14498	671.98	31.58	638	654	672	689	709
	Black	39865	650.96	26.94	622	638	651	667	682
	Hispanic	41370	650.82	28.10	618	636	651	667	682
	American Indian	1020	654.18	26.81	625	641	654	667	689
	Multi-Racial	128	665.33	30.20	638	648	664	682	697
	White	109922	669.13	28.04	638	654	667	682	697
	Unknown	68	668.72	30.32	631	648	670	686	709
NRC	New York City	71310	654.89	29.80	622	638	654	672	689
	Big 4 Cities	7855	644.66	29.72	611	631	646	660	676
	High Needs Urban/Suburban	15718	655.69	27.44	625	641	657	672	689
	High Needs Rural	12493	660.57	27.17	631	646	660	676	689
	Average Needs	64994	667.22	27.08	638	651	667	682	697
	Low Needs	31790	676.67	26.97	648	660	676	689	709
	Charter	2294	659.40	24.10	633	646	657	672	689
LEP	LEP = Y	9245	626.94	30.14	595	615	631	646	657
SWD	All Codes	30308	634.29	30.53	601	622	638	654	664
SUA	All Codes	38066	635.18	30.66	601	622	638	654	667

**Grade 8**

Scale score statistics and N-counts of demographic groups for Grade 8 are presented in Table 38. The population scale score mean was 656.91 with a standard deviation of 37.96. By gender subgroup, Females outperformed Males, but the difference was slightly larger than a quarter 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. The students from the Big 4 Cities scored below their peers from schools with other NRC designations and about a half of standard deviation below the population mean. SWD and SUA subgroups scored approximately one standard deviation below the mean scale score for the population. The LEP subgroup, which had a scale score mean just below fifty scale score units below the population mean, was the lowest performing group analyzed. At the 50<sup>th</sup> percentile, the following groups exceeded the population score of 657: Female (661), Asian (666), Multi-racial (666), White (666), Average Needs districts (661), and Low Needs districts (671).

**Table 38. Scale Score Distribution Summary, by Subgroup, Grade 8**

Demographic Category (Subgroup)		N-count	SS Mean	SS Std Dev	10 <sup>th</sup> %tile	25 <sup>th</sup> %tile	50 <sup>th</sup> %tile	75 <sup>th</sup> %tile	90 <sup>th</sup> %tile
State	All Students	208959	656.91	37.96	616	635	657	677	705
Gender	Female	101842	663.26	38.06	622	642	661	684	705
	Male	107117	650.88	36.86	610	632	649	671	693
Ethnicity	Asian	14272	667.87	40.36	625	645	666	693	705
	Black	40399	642.44	31.86	607	625	642	661	677
	Hispanic	40911	641.09	34.86	604	622	642	661	677
	American Indian	1059	643.65	35.07	607	625	645	661	677
	Multi-Racial	117	662.69	31.15	629	649	666	677	705
	White	112147	666.62	37.10	629	645	666	684	705
	Unknown	54	672.39	42.41	622	649	671	693	726
NRC	New York City	71488	646.13	36.39	607	625	645	666	684
	Big 4 Cities	8337	636.86	35.34	597	619	638	657	677
	High Needs Urban/Suburban	15970	648.28	34.50	610	629	649	666	684
	High Needs Rural	13148	654.48	34.82	616	635	653	671	693
	Average Needs	66177	664.07	35.33	625	645	661	684	705
	Low Needs	31770	678.28	37.06	638	657	671	693	726
	Charter	1423	649.15	28.72	616	632	649	666	684
LEP	LEP = Y	8618	608.96	36.70	569	593	613	632	645
SWD	All Codes	30218	620.26	33.71	585	604	622	642	657
SUA	All Codes	38035	621.10	34.92	585	604	625	642	657

***Performance Level Distribution Summary***

Percentage of students in each performance level was computed based on performance levels scale score ranges established during the 2006 Standard Setting for all grades except Grade 3, Level IV. For the 2008 Grade 3 English Language Arts Test only, a scale score of 720 (corresponding to a raw score of 32 out of 33) has been accepted as a lowest possible scale score that student needs to earn to be classified in Level IV. Although the original cut score for Level IV established during the Standard Setting for Grade 3 was 730, there was no such score in 2008 raw score-to-scale score conversion table and the next higher scale score in the table was the highest obtainable scale score of 780 associated with the perfect Grade 3 test raw score. NYS Technical Advisory Group endorsed a policy decision by NYSED to adjust the Level IV cut for Grade 3 in 2008 test administration so that students were not required to earn a perfect raw score (33 out of 33) in order to achieve a Level IV. If a perfect raw score were to be required to achieve a Level IV for any of the future Grades 3–8 English Language Arts Tests, a similar adjustment will be made. Information on the cut score adjustment for Grade 3 was posted on the NYSED web site at <http://www.emsc.nysed.gov/irts/ela-math/ela-math-08/2008ELAScaleScoretoPerformanceLevels.html>

Table 39 shows the ELA cut scores used for classification of students to the four performance level categories in 2008.

**Table 39. ELA Grades 3–8 Performance Level Cut Scores**

Grade	Level II cut	Level III cut	Level IV cut
3	616	650	720
4	612	650	716
5	608	650	711
6	598	650	705
7	600	650	712
8	602	650	715

Tables 40–46 show the performance level distribution for all examinees from public and charter school with valid scores. Table 40 presents performance level data for total populations of students in Grades 3–8. Tables 41–46 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 LEP students, SWD, and SUA 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 40. 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	195695	5.94	23.94	57.75	12.38	70.13
4	196915	7.48	21.39	62.71	8.42	71.13
5	197901	1.85	20.53	71.70	5.92	77.62
6	200352	1.71	31.30	62.33	4.66	66.99
7	206871	1.85	28.03	67.58	2.55	70.12
8	208959	5.16	38.62	50.52	5.69	56.21

### Grade 3

Performance level distributions and N-counts of demographic groups for Grade 3 are presented in Table 41. Statewide, 70.13% of third-graders were Level III or Level IV. 7.42% of Male students were Level I, as compared to only 4.37% of Female students. The percentage of students in Levels III and IV varied widely by ethnicity and NRC subgroups. About 88% of Low Needs district students and about 80% of Multi-Racial students were

classified in Levels III and IV; whereas the American Indian, Black, Charter, and/or Big 4 Cities had a range of about 30%–50% of students who were in Level I or Level II. About one-fifth to one-quarter of students with LEP, SWD, or SUA status were in Level I and only about 1% are 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 41. 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	195695	5.94	23.94	57.75	12.38	70.13
Gender	Female	95081	4.37	22.69	59.63	13.31	72.94
	Male	100614	7.42	25.11	55.98	11.50	67.47
Ethnicity	Asian	14302	2.69	17.47	63.67	16.17	79.84
	Black	37585	9.36	34.36	50.51	5.77	56.28
	Hispanic	41422	10.71	35.59	48.44	5.25	53.69
	American Indian	978	7.06	34.76	52.76	5.42	58.18
	Multi-Racial	233	3.43	17.17	66.52	12.88	79.40
	White	101110	3.16	16.12	63.44	17.28	80.72
	Unknown	65	3.08	10.77	72.31	13.85	86.15
NRC	New York City	69150	9.02	31.22	51.75	8.01	59.76
	Big 4 Cities	8103	12.40	39.24	43.54	4.81	48.35
	High Needs Urban/Suburban	15830	6.99	29.08	55.57	8.36	63.93
	High Needs Rural	11469	5.79	25.17	59.56	9.48	69.04
	Average Needs	58520	3.38	18.14	63.39	15.10	78.48
	Low Needs	29390	1.51	10.29	65.10	23.11	88.21
	Charter	3023	2.58	29.14	59.74	8.53	68.28
LEP	LEP = Y	16568	18.35	48.71	31.86	1.07	32.93
SWD	All Codes	26011	26.67	42.25	28.87	2.21	31.08
SUA	All Codes	39019	22.06	44.91	31.15	1.88	33.03

#### Grade 4

Performance level distributions and N-counts of demographic groups for Grade 4 are presented in Table 42. Across New York, approximately 71% 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 (88.93%), and the SWD subgroup had the lowest (29.97%). Students in the Black, and Hispanic subgroups had percentage classified in Levels III and IV just above 50% which was more than 20% below the other ethnic subgroups. Over twice as many Big 4 City students were in Level I than the population. Nearly a quarter of students with LEP, SWD, or SUA status were in Level I (over three times the amount of the Statewide rate of 7.48%) 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, White, Average Needs districts, and Low Needs districts.

**Table 42. 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	196915	7.48	21.39	62.71	8.42	71.13
Gender	Female	96399	5.35	19.36	64.45	10.85	75.30
	Male	100516	9.52	23.34	61.05	6.09	67.14
Ethnicity	Asian	14174	3.85	13.07	67.75	15.33	83.08
	Black	37975	12.01	31.52	53.15	3.33	56.47
	Hispanic	41178	12.14	31.06	53.55	3.26	56.81
	American Indian	952	11.66	26.68	57.98	3.68	61.66
	Multi-Racial	188	6.91	22.34	65.43	5.32	70.74
	White	102361	4.39	14.85	69.28	11.47	80.76
	Unknown	87	2.30	11.49	73.56	12.64	86.21
	NRC	New York City	69692	10.71	28.18	55.37	5.74
	Big 4 Cities	7697	15.30	33.87	47.47	3.35	50.83
	High Needs Urban/Suburban	15655	9.66	25.17	59.98	5.19	65.17
	High Needs Rural	11490	7.74	22.92	63.88	5.46	69.34
	Average Needs	59451	4.67	16.48	69.32	9.53	78.85
	Low Needs	30278	2.03	9.04	71.99	16.94	88.93
	Charter	2406	5.74	29.18	61.51	3.57	65.09
LEP	LEP = Y	14008	23.71	44.65	31.33	0.32	31.65
SWD	All Codes	28885	31.65	38.39	29.45	0.52	29.97
SUA	All Codes	41079	26.58	39.69	33.17	0.56	33.73

### Grade 5

Performance level distributions and N-counts of demographic groups for Grade 5 are presented in Table 43. About 68% 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 (92.53%). Fewer Male students were in the Level I category than was observed with Grades 3 and 4, by a few percentage points. Students in the American Indian, Black, and Hispanic subgroups had rates around 65% of students classified in Levels III and IV, approximately 20% less than other ethnic subgroups. Over twice as many Big 4 City students were in Level I than the population's rate. Close to a quarter of the students with LEP, SWD, or SUA status were in Level I (approximately four to five times as many as the Statewide rate of 1.85%), yet only about 40% 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, High Needs Rural, Average Needs districts, and Low Needs districts.

**Table 43. 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	197901	1.85	20.53	71.70	5.92	77.62
Gender	Female	96664	1.37	19.00	72.86	6.77	79.63
	Male	101237	2.31	21.99	70.59	5.11	75.70
Ethnicity	Asian	14572	1.13	12.92	75.39	10.55	85.95
	Black	38296	2.63	33.21	62.15	2.01	64.16
	Hispanic	40577	3.45	31.28	62.90	2.37	65.27
	American Indian	902	2.22	30.16	64.30	3.33	67.63
	Multi-Racial	168	1.19	18.45	73.81	6.55	80.36
	White	103311	1.04	12.60	78.24	8.13	86.36
	Unknown	75	0.00	17.33	69.33	13.33	82.67
NRC	New York City	69157	2.82	28.17	64.73	4.28	69.00
	Big 4 Cities	7525	4.37	37.78	56.17	1.67	57.85
	High Needs Urban/Suburban	15199	2.61	25.94	67.90	3.56	71.46
	High Needs Rural	11366	1.62	20.62	73.61	4.14	77.76
	Average Needs	60375	0.96	14.44	77.86	6.74	84.60
	Low Needs	30708	0.43	7.05	81.23	11.30	92.53
	Charter	3309	1.21	29.98	66.42	2.39	68.81
LEP	LEP = Y	11372	9.77	56.14	33.94	0.15	34.09
SWD	All Codes	29635	8.81	50.27	40.38	0.53	40.92
SUA	All Codes	40120	8.01	49.18	42.25	0.56	42.81

**Grade 6**

Performance level distributions and N-counts of demographic groups for Grade 6 are presented in Table 44. Statewide, 66.99% 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 (87.41%), and the LEP, SWD, and SUA subgroups had the fewest. Students in the American Indian, Black, and Hispanic subgroups had around 50% of students classified in Level III and above. Students from Low Needs districts outperformed students in all other subgroups, across demographic categories as in the previous grades. Over twice the percentage of Big 4 City students were placed in Level I than the percentage of the general population with the exception of NYC, and about 49% of students from those districts were classified in Levels III and IV (with 1.31% in Level IV). The majority of students with LEP, 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, High Needs Rural, Average Needs districts, and Low Needs districts.

**Table 44. 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	200352	1.71	31.30	62.33	4.66	66.99
Gender	Female	97505	1.08	27.80	65.05	6.07	71.12
	Male	102847	2.31	34.61	59.74	3.33	63.08
Ethnicity	Asian	14521	1.17	21.07	69.60	8.16	77.76
	Black	37970	2.43	47.56	48.77	1.24	50.01
	Hispanic	40576	3.39	48.62	46.66	1.33	47.99
	American Indian	914	2.30	40.48	54.70	2.52	57.22
	Multi-Racial	160	0.00	26.88	64.38	8.75	73.13
	White	106129	0.88	20.19	72.24	6.69	78.93
	Unknown	82	1.22	26.83	58.54	13.41	71.95
NRC	New York City	69102	2.92	44.54	50.35	2.19	52.54
	Big 4 Cities	7544	3.45	47.57	47.67	1.31	48.98
	High Needs Urban/Suburban	15305	1.91	38.25	57.43	2.42	59.84
	High Needs Rural	11776	1.19	30.88	64.45	3.48	67.93
	Average Needs	61976	0.81	22.09	71.56	5.54	77.10
	Low Needs	31556	0.33	12.26	76.39	11.02	87.41
	Charter	2768	0.58	39.85	58.09	1.48	59.57
LEP	LEP = Y	9816	10.91	74.17	14.88	0.03	14.91
SWD	All Codes	30391	8.07	67.90	23.89	0.14	24.03
SUA	All Codes	38535	7.72	66.45	25.58	0.25	25.83

**Grade 7**

Performance level distributions and N-counts of demographic groups for Grade 7 are presented in Table 45. In Grade 7, 70.12% 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 60% of Big 4 Cities students were in Levels I and II. Almost 90% of Low Needs students were in Levels III and IV. Fewer than 20% of LEP students were in Levels III and IV. The LEP, SWD, and SUA subgroups were well below the performance achievement of the general population, with around 80% 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 45. 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	206871	1.85	28.03	67.58	2.55	70.12
Gender	Female	101183	1.18	23.35	72.19	3.29	75.47
	Male	105688	2.49	32.51	63.16	1.84	65.00
Ethnicity	Asian	14498	1.28	18.20	74.84	5.67	80.51
	Black	39865	2.74	42.91	53.74	0.61	54.35
	Hispanic	41370	3.35	42.04	53.89	0.72	54.61
	American Indian	1020	2.06	39.41	57.45	1.08	58.53
	Multi-Racial	128	2.34	24.22	69.53	3.91	73.44
	White	109922	1.04	18.55	76.87	3.54	80.41
	Unknown	68	1.47	25.00	70.59	2.94	73.53
NRC	New York City	71310	2.83	37.90	57.68	1.60	59.28
	Big 4 Cities	7855	5.19	50.21	44.04	0.56	44.60
	High Needs Urban/Suburban	15718	2.07	36.23	60.54	1.16	61.71
	High Needs Rural	12493	1.54	29.02	67.62	1.81	69.43
	Average Needs	64994	0.96	20.43	75.66	2.95	78.60
	Low Needs	31790	0.39	10.90	83.25	5.45	88.71
	Charter	2294	0.78	31.82	66.26	1.13	67.39
LEP	LEP = Y	9245	11.52	70.81	17.65	0.02	17.67
SWD	All Codes	30308	8.58	61.81	29.47	0.14	29.61
SUA	All Codes	38066	8.37	60.88	30.59	0.17	30.76

**Grade 8**

Performance level distributions and N-counts of demographic groups for Grade 8 are presented in Table 46. In Grade 8, 56.21% of the students were in Levels III and IV. Over 10% more Female than Male students were in Levels III or IV. Over 55% of American Indian, Black, and Hispanic students were in Levels I and II. Over 80% of Low Needs students were in Levels III and IV, while fewer than 7% of LEP students were in Levels III and IV. The LEP, SWD, and SUA subgroups were well below the performance achievement of the general population, with over 80% 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 46. 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	208959	5.16	38.62	50.52	5.69	56.21
Gender	Female	101842	3.46	33.62	55.22	7.70	62.92
	Male	107117	6.78	43.38	46.05	3.79	49.84
Ethnicity	Asian	14272	4.16	25.78	60.76	9.30	70.07
	Black	40399	7.57	55.32	35.57	1.54	37.11
	Hispanic	40911	9.78	52.94	35.65	1.64	37.28
	American Indian	1059	7.55	50.42	39.85	2.17	42.02
	Multi-Racial	117	3.42	28.21	63.25	5.13	68.38
	White	112147	2.72	28.93	60.11	8.24	68.35
	Unknown	54	3.70	22.22	55.56	18.52	74.07
NRC	New York City	71488	7.84	49.51	39.77	2.87	42.64
	Big 4 Cities	8337	11.92	56.40	30.25	1.43	31.68
	High Needs Urban/Suburban	15970	7.02	47.40	42.53	3.06	45.59
	High Needs Rural	13148	4.40	42.32	48.95	4.34	53.29
	Average Needs	66177	2.82	31.16	59.28	6.75	66.03
	Low Needs	31770	1.06	18.15	67.66	13.13	80.79
	Charter	1423	3.23	52.07	42.45	2.25	44.69
LEP	LEP = Y	8618	34.00	59.57	6.39	0.03	6.43
SWD	All Codes	30218	22.57	64.12	13.08	0.23	13.31
SUA	All Codes	38035	22.59	62.34	14.73	0.34	15.07

## Section IX: Longitudinal Comparison of Results

This section provides longitudinal comparison of operational scale score results on the New York State 2006-2008 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 47.

**Table 47. 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	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	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	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	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	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
8	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

As seen in Table 47 an increase in scale score means was observed for Grades 5, 6, 7, and 8 between 2006 and 2008 test administrations. Relatively steady yearly gain was noticed for Grades 5, 7 and 8 with the overall population mean scale score increase of 5 or more scale points between years 2006 and 2008. An increase of approximately 5 scale score points was observed for Grade 6 between years 2006 and 2007. No score change was noticed for Grade 6 between administration years 2007 and 2008. For Grades 3 and 4, a slight mean scale score decline (1 to 2 scale score points) was observed between years 2006 and 2007 and again, a small increase (approximately 2 points) was observed between years 2007 and 2008. Overall, the mean scale score increase for Grades 3 and 4 was less than 1 scale score point between administration years 2006 and 2008.

The variability of scale score distribution was uniform across years for Grades 3, 4, and 8. The scale score standard deviation was around 40 scale score points for these grades in all three administration years. For Grades 5 and 7, the variability of scale score distribution decreased in 2008. The standard deviations for these grades decreased from about 40 scale score points in 2006 and 2007 administrations to approximately 30 points in 2008 administration. The standard deviation for Grade 6 decreased from approximately 40 scale score points in 2006 to about 35 scale score points in 2007 and again to approximately 30 scale score points in 2008.

Following evaluation of the pattern of means scale score change between 2006 and 2008 test administration, a longitudinal trend of proficiency score distribution was evaluated. The percentage of students classified in proficiency Levels III and IV increased only slightly for Grades 3 and 4 between years 2006 and 2008. Grades 5 and 7 proficiency score trend indicates slight increase (approximately 1%) in the percentage of students classified in Levels III and IV between years 2006 and 2007 and a larger increase in the percentage of students classified in Levels III and IV between years 2007 and 2008. The percentage of Grade 5 students classified in Levels III and IV increased from approximately 68% to 78% between years 2007 and 2008 and the percentage of Grade 7 students classified in Levels III and IV increased from approximately 58% to 70% between years 2007 and 2008. The percentage of Grade 6 students classified in Levels III and IV was observed to be steadily increasing (approximately 3% each year) between 2006 and 2008 test administrations. It was also observed that while approximately 8% more Grade 8 students were classified in Levels III and IV in 2007 than in 2006, no increase in the percentage of students classified in the two highest proficiency levels was observed between years 2007 and 2008.

In summary, the mean scale score change and the change in percentage of students classified in Levels III and IV was not uniform across grades during the three years of test administration. As expected, the mean scale score change was found to be in alignment with the performance levels score trend between years 2006 and 2008.

## 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 readability concept and summary statistics for passages selected for the 2008 operational administration are 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.

**Table A1. Readability Summary Information for 2008 Operational Test Passages**

Title	Passage Type	Word Count	Avg. Prose Score	Fry Graph	Spache/Dale-Chall*	Flesch-Kincaid Formula
<b>GRADE 3</b>						
<b>Book 1 (Reading)</b>						
Fishing for Dinner	Lit-Fiction	310	4.20	3.30	3.26	2.68
George Washington, Our First President	Info-Bio	325	5.73	4.91	3.03	4.13
How Shall We Go to Grandma's House?	Lit-Poem	90	n/a	n/a	n/a	n/a
Balloon Volleyball	Info-How to	325	5.46	4.89	3.13	4.08
<b>Readability Averages*</b>			<b>5.13</b>	<b>4.37</b>	<b>3.14</b>	<b>3.63</b>
<b>Book 2 (Listening)</b>						
The Tired Chipmunk	Lit-Fiction	465	3.31	2.03	2.80	1.27

*(Continued on next page)*

**Table A1. Readability Summary Information for 2008 Operational Test Passages (cont.)**

Title	Passage Type	Word Count	Avg. Prose Score	Fry Graph	Spache/Dale-Chall*	Flesch-Kincaid Formula
<b>GRADE 4</b>						
<b>Book 1 (Reading)</b>						
Follow That Horse	Info-Article	465	6.44	5.57	3.90	4.58
My Hand Was in the Cookie Jar	Lit-Poem	135	n/a	n/a	n/a	n/a
First in Line	Lit-Fiction	565	5.47	4.45	2.77	3.71
Why Do We Need Sleep?	Info-Article	290	6.87	6.66	4.00	5.51
Betsy Brandon Meets the President	Lit-Fiction	475	5.85	4.96	3.22	4.11
<b>Book 2 (Listening)</b>						
The Voice of Rigo	Lit-Fiction	450	4.16	3.04	2.41	2.51
<b>Book 3 (Reading pair)</b>						
What Kinds of Jobs Can Dogs Do?	Info-Article	300	5.21	4.03	3.22	3.39
Well Done, York	Lit-Fiction	480	3.94	2.40	3.41	1.72
<b>Readability Averages*</b>			<b>5.63</b>	<b>4.68</b>	<b>3.42</b>	<b>3.84</b>
<b>GRADE 5</b>						
<b>Book 1 (Reading)</b>						
Will My Toy Car Survive a Croc Attack?	Info-Article	545	6.46	6.42	9–10	5.25
A Tree Needs a Special Place	Lit-Fiction	590	4.70	3.68	5–6	3.06
Trapped by a King Cobra	Info-Article	485	6.60	6.25	9–10	4.98
A Spaghetti Tale	Lit-Essay	415	6.12	5.85	7–8	4.71
<b>Readability Averages*</b>			<b>5.97</b>	<b>5.55</b>	<b>7–8</b>	<b>4.50</b>
<b>Book 2 (Listening)</b>						
The Courage of Molly Pitcher	Info-Bio	460	7.49	7.45	7–8	6.20
<b>GRADE 6</b>						
<b>Book 1 (Reading)</b>						
Mira Sees the Light	Lit-Fiction	530	5.92	4.93	5–6	4.15
Nadia Begay	Info-Article	600	8.33	7.78	7–8	6.86
Galapagos Island Vacation	Info-Article	585	7.93	7.61	7–8	6.66
Soccer Cinderella	Lit-Fiction	710	4.38	3.31	7–8	2.67

*(Continued on next page)*

**Table A1. Readability Summary Information for 2008 Operational Test Passages (cont.)**

Title	Passage Type	Word Count	Avg. Prose Score	Fry Graph	Spache/Dale-Chall*	Flesch-Kincaid Formula
<b>GRADE 6</b>						
<b>Book 1 (Reading)</b>						
This Land Is Your Land	Info-Article	525	7.35	7.10	7–8	6.73
<b>Book 2 (Listening)</b>						
The Bat in the Refrigerator	Lit-Fiction	545	5.37	4.45	5–6	3.75
<b>Book 3 (Reading pair)</b>						
Final Approach	Info-Article	460	7.43	7.02	7–8	5.83
Dreams of Flying	Info-Article	600	7.98	7.89	7–8	6.90
<b>Readability Averages*</b>			<b>7.05</b>	<b>6.52</b>	<b>7–8</b>	<b>5.69</b>
<b>GRADE 7</b>						
<b>Book 1 (Reading)</b>						
Conversations with Apes	Info-Article	525	8.72	8.66	7–8	7.38
The Ride Home	Lit-Fiction	680	5.11	4.52	7–8	3.79
Once Upon a Time	Info-Article	590	6.78	6.49	7–8	5.34
Freaky Farm	Info-Article	350	8.71	8.43	11–12	7.53
The Island	Lit-Fiction	505	7.45	7.24	5–6	7.24
<b>Readability Averages*</b>			<b>7.35</b>	<b>7.07</b>	<b>7–8</b>	<b>6.26</b>
<b>Book 2 (Listening)</b>						
Flights of Fancy	Info-Article	460	6.49	6.56	7–8	5.41
<b>GRADE 8</b>						
<b>Book 1 (Reading)</b>						
Millicent Min, Girl Genius	Lit-Fiction	505	8.44	7.94	7–8	7.75
Living the Wild Life	Info-Article	840	8.80	7.95	9–10	7.81
Joan Benoit	Lit-Poem	105	n/a	n/a	n/a	n/a
Tech-Trash Tragedy	Info-Article	800	9.76	9.13	9–10	8.44
A Fork in the Road	Lit-Fiction	855	4.67	4.12	5–6	3.45
<b>Book 2 (Listening)</b>						
Classic Jazz Artist	Info-Article	540	9.35	9.22	9–10	7.74

*(Continued on next page)*

**Table A1. Readability Summary Information for 2008 Operational Test Passages (cont.)**

Title	Passage Type	Word Count	Avg. Prose Score	Fry Graph	Spache/Dale-Chall*	Flesch-Kincaid Formula
<b>GRADE 8</b>						
<b>Book 3 (Reading pair)</b>						
The Youngest of Them All	Info-Article	405	8.52	8.26	7–8	7.63
Helping Hand	Info-Article	420	8.69	8.15	9–10	7.13
<b>Readability Averages*</b>			<b>8.15</b>	<b>7.59</b>	<b>8–9</b>	<b>7.03</b>

**Table A2. Number, Type, and Length of Passages**

Grade	# of Listening Passages	Approximate Word Length	# of Reading Passages	Passage Types	Approximate Word Length	Passage Types
3	8	200–400	20 (includes 5 sets of short paired-passages)	Literary	200–600	50% Literary; 50% Informational
4	5	250–450	20 (includes 8 sets of short paired-passages)	Literary	250–600	50% Literary; 50% Informational
5	12	300–500	20 (includes 5 sets of short paired-passages)	Literary	250–600	50% Literary; 50% Informational
6	8	350–550	24 (includes 5 sets of short paired-passages)	Informational	300–650	50% Literary; 50% Informational

*(Continued on next page)*

**Table A2. Number, Type, and Length of Passages (cont.)**

Grade	# of Listening Passages	Approximate Word Length	# of Reading Passages	Passage Types	Approximate Word Length	Passage Types
7	8	400–600	24 (includes 5 sets of short paired-passages)	Informational (May be 2 short paired-pieces)	350–700	50% Literary; 50% Informational
8	5	450–650	20 (includes 8 sets of short paired-passages)	Informational (May be 2 short paired-pieces)	350–800	50% Literary; 50% Informational

## 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 2008 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 item pool. The pools of items eligible for 2008 item selection will include 2005, 2006, and 2007 FT items for Grades 3, 5, 6, and 7 and 2003, 2005, 2006, and 2007 FT items for Grades 4 and 8. 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 it is 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 2008 OP forms and the 2007 OP forms.
- From the ITEMWIN output evaluate expected percentage of maximum raw score at each scale score and difference between reference set (2007) and working set (2008)—we want the difference to be no more than 0.01, which is unfortunately sometimes hard to achieve, but please try your best.
  - It is especially important to get a good curve alignment at and around proficiency level cut scores. Good alignment will help preserve the impact data from the previous year of testing.
- Try to get the best scale coverage—make sure that 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, 2008 full selection, 2008 MC items only)

- Item number in 2008 OP book
- Item unique identification number, item type, FT year, FT form, and FT item number
- Item classical statistics (p-values, point biserial, 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: limited English proficiency (LEP), students with disabilities (SWD), and students using accommodations (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	LEP	<b>1</b>	<b>5.57</b>	<b>19.88</b>	<b>19.88</b>
		2	1.30	4.63	24.50
		3	1.06	3.78	28.28
	SWD	<b>1</b>	<b>6.42</b>	<b>22.93</b>	<b>22.93</b>
		2	1.43	5.12	28.05
		3	1.06	3.78	31.83
	SUA	<b>1</b>	<b>6.02</b>	<b>21.50</b>	<b>21.50</b>
		2	1.35	4.82	26.32
		3	1.10	3.94	30.26
4	LEP	<b>1</b>	<b>5.82</b>	<b>18.78</b>	<b>18.78</b>
		2	1.34	4.33	23.10
		3	1.04	3.36	26.46
		4	1.01	3.25	29.71
	SWD	<b>1</b>	<b>6.89</b>	<b>22.22</b>	<b>22.22</b>
		2	1.38	4.44	26.65
		3	1.03	3.31	29.97
	SUA	<b>1</b>	<b>6.66</b>	<b>21.48</b>	<b>21.48</b>
		2	1.39	4.47	25.94
3		1.03	3.32	29.27	
5	LEP	<b>1</b>	<b>4.83</b>	<b>17.88</b>	<b>17.88</b>
		2	1.15	4.27	22.15
		3	1.07	3.98	26.12
		4	1.04	3.84	29.96
	SWD	<b>1</b>	<b>5.43</b>	<b>20.11</b>	<b>20.11</b>
		2	1.14	4.21	24.33
		3	1.08	4.00	28.32
		4	1.01	3.74	32.06
	SUA	<b>1</b>	<b>5.34</b>	<b>19.77</b>	<b>19.77</b>
		2	1.14	4.23	24.00
		3	1.08	3.99	27.99
		4	1.01	3.73	31.71

*(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	LEP	<b>1</b>	<b>5.72</b>	<b>19.73</b>	<b>19.73</b>
		2	1.18	4.07	23.80
		3	1.05	3.60	27.40
		4	1.02	3.52	30.92
		5	1.01	3.47	34.39
	SWD	<b>1</b>	<b>6.38</b>	<b>22.01</b>	<b>22.01</b>
		2	1.19	4.12	26.13
		3	1.07	3.69	29.82
	SUA	<b>1</b>	<b>6.45</b>	<b>22.25</b>	<b>22.25</b>
2		1.17	4.04	26.29	
3		1.04	3.60	29.89	
7	LEP	<b>1</b>	<b>5.48</b>	<b>15.66</b>	<b>15.66</b>
		2	1.36	3.88	19.55
		3	1.17	3.34	22.88
		4	1.11	3.17	26.06
		5	1.07	3.06	29.12
		6	1.02	2.93	32.04
		7	1.01	2.87	34.91
	SWD	<b>1</b>	<b>6.35</b>	<b>18.14</b>	<b>18.14</b>
		2	1.29	3.70	21.84
		3	1.15	3.30	25.14
		4	1.07	3.06	28.20
		5	1.03	2.93	31.13
	SUA	<b>1</b>	<b>6.45</b>	<b>18.43</b>	<b>18.43</b>
		2	1.30	3.72	22.15
		3	1.13	3.24	25.39
4		1.07	3.06	28.45	
5		1.02	2.91	31.35	
6		1.01	2.87	34.22	
8	LEP	<b>1</b>	<b>5.12</b>	<b>17.67</b>	<b>17.67</b>
		2	1.19	4.11	21.78
		3	1.13	3.88	25.67
		4	1.07	3.70	29.36
		5	1.02	3.52	32.88
	SWD	<b>1</b>	<b>5.68</b>	<b>19.58</b>	<b>19.58</b>
		2	1.15	3.97	23.55
		3	1.09	3.77	27.31
		4	1.04	3.57	30.88
		5	1.00	3.45	34.33
	SUA	<b>1</b>	<b>5.88</b>	<b>20.26</b>	<b>20.26</b>
		2	1.12	3.85	24.11
		3	1.07	3.70	27.81
4		1.03	3.57	31.37	

## Appendix E—Items Flagged for DIF

These tables support the DIF information in Section V, “Operational Test Data Collection and Classical Analyses,” 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 2008 Classical DIF Item Flags**

Grade	Item #	Subgroup	DIF	SMD	Mantel-Haenszel	Delta
3	27	Black	In favor	0.11	n/a	n/a
4	7	Black	Against	-0.11	No Flag	No Flag
4	7	Hispanic	Against	-0.11	No Flag	No Flag
4	7	Asian	Against	-0.12	1314.40	-1.83
4	29	Asian	In favor	0.11	n/a	n/a
4	29	Female	In favor	0.13	n/a	n/a
4	31	Asian	In favor	0.18	n/a	n/a
4	31	Female	In favor	0.10	n/a	n/a
5	3	Asian	Against	No Flag	257.55	-1.69
5	21	Asian	In favor	0.13	n/a	n/a
5	26	Female	Against	-0.11	n/a	n/a
5	27	Black	Against	-0.19	n/a	n/a
5	27	Hispanic	Against	-0.14	n/a	n/a
5	27	Female	In favor	0.10	n/a	n/a
5	27	High Needs	Against	-0.14	n/a	n/a
6	20	Asian	Against	-0.11	No Flag	No Flag
6	27	Black	Against	-0.10	n/a	n/a
6	28	Asian	In favor	0.10	n/a	n/a
6	29	Asian	In favor	0.12	n/a	n/a
6	29	Female	In favor	0.12	n/a	n/a
7	19	Asian	Against	-0.10	905.35	-1.52
7	35	Black	Against	-0.11	n/a	n/a
7	35	Hispanic	Against	-0.12	n/a	n/a
8	23	Asian	Against	-0.12	No Flag	No Flag
8	28	Hispanic	In favor	0.10	n/a	n/a
8	28	Asian	In favor	0.18	n/a	n/a
8	28	Female	In favor	0.11	n/a	n/a
8	29	Female	In favor	0.11	n/a	n/a

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 ( $D_{ig}$ )
4	31	Asian	In Favor	0.141
5	27	High Needs	Against	-0.105
5	27	Black	Against	-0.164
5	27	Hispanic	Against	-0.104
8	23	Asian	Against	-0.115
8	28	Asian	In Favor	0.120

## Appendix F—Item-Model Fit Statistics

These tables support the item-model fit information in Section VI, “IRT Scaling and Equating.” The item number, calibration model, chi-square, degrees of freedom (DF), N-count, obtained-Z fit statistic, and critical-Z fit statistic are presented for each item. Fit for all items in the Grades 3–8 ELA Tests was acceptable (critical  $Z >$  obtained  $Z$ ).

**Table F1. ELA Item Fit Statistics, Grade 3**

Item Number	Model	Chi-Square	DF	N-count	Obtained $Z$	Critical $Z$	Fit Ok?
1	3PL	89.13	7	184554	21.95	492.14	Y
2	3PL	392.71	7	184554	103.08	492.14	Y
3	3PL	415.10	7	184554	109.07	492.14	Y
4	3PL	317.28	7	184554	82.93	492.14	Y
5	3PL	157.65	7	184554	40.26	492.14	Y
6	3PL	932.21	7	184554	247.27	492.14	Y
7	3PL	175.20	7	184554	44.95	492.14	Y
8	3PL	722.62	7	184554	191.26	492.14	Y
9	3PL	2484.41	7	184554	662.12	492.14	N
10	3PL	284.43	7	184554	74.15	492.14	Y
11	3PL	337.27	7	184554	88.27	492.14	Y
12	3PL	217.43	7	184554	56.24	492.14	Y
13	3PL	525.64	7	184554	138.61	492.14	Y
14	3PL	185.36	7	184554	47.67	492.14	Y
15	3PL	216.62	7	184554	56.02	492.14	Y
16	3PL	352.60	7	184554	92.37	492.14	Y
17	3PL	379.66	7	184554	99.60	492.14	Y
18	3PL	635.90	7	184554	168.08	492.14	Y
19	3PL	204.42	7	184554	52.76	492.14	Y
20	3PL	718.57	7	184554	190.17	492.14	Y
21	2PPC	698.31	17	184554	116.84	492.14	Y
22	3PL	176.69	7	184554	45.35	492.14	Y
23	3PL	98.11	7	184554	24.35	492.14	Y
24	3PL	322.57	7	184554	84.34	492.14	Y
25	3PL	89.91	7	184554	22.16	492.14	Y
26	2PPC	715.59	17	184554	119.81	492.14	Y
27	2PPC	1643.22	17	184554	278.89	492.14	Y
28	2PPC	748.71	26	184554	100.22	492.14	Y

**Table F2. ELA Item Fit Statistics, Grade 4**

Item Number	Model	Chi-Square	DF	N-count	Obtained Z	Critical Z	Fit Ok?
1	3PL	190.80	7	192011	49.12	512.03	Y
2	3PL	223.18	7	192011	57.78	512.03	Y
3	3PL	32.08	7	192011	6.70	512.03	Y
4	3PL	409.12	7	192011	107.47	512.03	Y
5	3PL	181.04	7	192011	46.51	512.03	Y
6	3PL	132.74	7	192011	33.6	512.03	Y
7	3PL	916.43	7	192011	243.06	512.03	Y
8	3PL	55.45	7	192011	12.95	512.03	Y
9	3PL	57.81	7	192011	13.58	512.03	Y
10	3PL	160.64	7	192011	41.06	512.03	Y
11	3PL	122.49	7	192011	30.87	512.03	Y
12	3PL	167.34	7	192011	42.85	512.03	Y
13	3PL	189.53	7	192011	48.78	512.03	Y
14	3PL	1293.14	7	192011	343.73	512.03	Y
15	3PL	264.17	7	192011	68.73	512.03	Y
16	3PL	384.28	7	192011	100.83	512.03	Y
17	3PL	97.26	7	192011	24.12	512.03	Y
18	3PL	86.26	7	192011	21.18	512.03	Y
19	3PL	122.13	7	192011	30.77	512.03	Y
20	3PL	258.82	7	192011	67.3	512.03	Y
21	3PL	402.85	7	192011	105.79	512.03	Y
22	3PL	177.75	7	192011	45.63	512.03	Y
23	3PL	120.85	7	192011	30.43	512.03	Y
24	3PL	84.55	7	192011	20.73	512.03	Y
25	3PL	646.57	7	192011	170.93	512.03	Y
26	3PL	324.41	7	192011	84.83	512.03	Y
27	3PL	161.06	7	192011	41.17	512.03	Y
28	3PL	221.92	7	192011	57.44	512.03	Y
29	2PPC	1535.25	35	192011	179.31	512.03	Y
30	2PPC	2517.82	35	192011	296.75	512.03	Y
31	2PPC	1146.13	26	192011	155.33	512.03	Y

**Table F3. ELA Item Fit Statistics, Grade 5**

Item Number	Model	Chi-Square	DF	N-count	Obtained Z	Critical Z	Fit Ok?
1	3PL	497.90	7	192873	131.20	514.33	Y
2	3PL	136.42	7	192873	34.59	514.33	Y
3	3PL	184.41	7	192873	47.42	514.33	Y
4	3PL	1111.64	7	192873	295.23	514.33	Y
5	3PL	228.32	7	192873	59.15	514.33	Y
6	3PL	79.73	7	192873	19.44	514.33	Y
7	3PL	382.97	7	192873	100.48	514.33	Y
8	3PL	137.06	7	192873	34.76	514.33	Y
9	3PL	1243.8	7	192873	330.55	514.33	Y
10	3PL	224.09	7	192873	58.02	514.33	Y
11	3PL	447.41	7	192873	117.71	514.33	Y
12	3PL	320.92	7	192873	83.90	514.33	Y
13	3PL	1140.72	7	192873	303.00	514.33	Y
14	3PL	90.26	7	192873	22.25	514.33	Y
15	3PL	283.37	7	192873	73.86	514.33	Y
16	3PL	512.94	7	192873	135.22	514.33	Y
17	3PL	973.56	7	192873	258.32	514.33	Y
18	3PL	152.00	7	192873	38.75	514.33	Y
19	3PL	154.77	7	192873	39.49	514.33	Y
20	3PL	704.53	7	192873	186.42	514.33	Y
21	2PPC	4144.35	17	192873	707.84	514.33	N
22	3PL	139.11	7	192873	35.31	514.33	Y
23	3PL	263.96	7	192873	68.67	514.33	Y
24	3PL	550.38	7	192873	145.22	514.33	Y
25	3PL	50.10	7	192873	11.52	514.33	Y
26	2PPC	601.20	17	192873	100.19	514.33	Y
27	2PPC	2874.33	26	192873	394.99	514.33	Y

**Table F4. ELA Item Fit Statistics, Grade 6**

Item Number	Model	Chi-Square	DF	N-count	Obtained Z	Critical Z	Fit Ok?
1	3PL	333.97	7	195574	87.39	521.53	Y
2	3PL	182.04	7	195574	46.78	521.53	Y
3	3PL	213.25	7	195574	55.12	521.53	Y
4	3PL	89.92	7	195574	22.16	521.53	Y
5	3PL	139.85	7	195574	35.51	521.53	Y
6	3PL	1315.73	7	195574	349.77	521.53	Y
7	3PL	19.50	7	195574	3.34	521.53	Y
8	3PL	98.16	7	195574	24.36	521.53	Y
9	3PL	42.37	7	195574	9.45	521.53	Y
10	3PL	55.48	7	195574	12.96	521.53	Y
11	3PL	34.80	7	195574	7.43	521.53	Y
12	3PL	193.24	7	195574	49.77	521.53	Y
13	3PL	724.51	7	195574	191.76	521.53	Y
14	3PL	224.40	7	195574	58.10	521.53	Y
15	3PL	153.91	7	195574	39.26	521.53	Y
16	3PL	238.46	7	195574	61.86	521.53	Y
17	3PL	432.03	7	195574	113.6	521.53	Y
18	3PL	118.82	7	195574	29.89	521.53	Y
19	3PL	219.46	7	195574	56.78	521.53	Y
20	3PL	255.90	7	195574	66.52	521.53	Y
21	3PL	207.15	7	195574	53.49	521.53	Y
22	3PL	144.49	7	195574	36.75	521.53	Y
23	3PL	432.06	7	195574	113.60	521.53	Y
24	3PL	197.59	7	195574	50.94	521.53	Y
25	3PL	239.53	7	195574	62.15	521.53	Y
26	3PL	161.18	7	195574	41.21	521.53	Y
27	2PPC	3000.32	44	195574	315.14	521.53	Y
28	2PPC	3653.47	44	195574	384.77	521.53	Y
29	2PPC	1323.33	26	195574	179.91	521.53	Y

**Table F5. ELA Item Fit Statistics, Grade 7**

Item Number	Model	Chi-Square	DF	N-count	Obtained Z	Critical Z	Fit Ok?
1	3PL	292.59	7	197260	76.33	526.03	Y
2	3PL	112.18	7	197260	28.11	526.03	Y
3	3PL	63.54	7	197260	15.11	526.03	Y
4	3PL	2682.34	7	197260	715.01	526.03	N
5	3PL	347.76	7	197260	91.07	526.03	Y
6	3PL	300.51	7	197260	78.44	526.03	Y
7	3PL	138.16	7	197260	35.05	526.03	Y
8	3PL	263.94	7	197260	68.67	526.03	Y
9	3PL	102.06	7	197260	25.41	526.03	Y
10	3PL	308.88	7	197260	80.68	526.03	Y
11	3PL	408.00	7	197260	107.17	526.03	Y
12	3PL	53.92	7	197260	12.54	526.03	Y
13	3PL	624.81	7	197260	165.12	526.03	Y
14	3PL	238.37	7	197260	61.84	526.03	Y
15	3PL	180.54	7	197260	46.38	526.03	Y
16	3PL	90.75	7	197260	22.38	526.03	Y
17	3PL	142.91	7	197260	36.32	526.03	Y
18	3PL	904.63	7	197260	239.90	526.03	Y
19	3PL	171.83	7	197260	44.05	526.03	Y
20	3PL	115.51	7	197260	29.00	526.03	Y
21	3PL	875.44	7	197260	232.10	526.03	Y
22	3PL	322.39	7	197260	84.29	526.03	Y
23	3PL	449.51	7	197260	118.27	526.03	Y
24	3PL	353.97	7	197260	92.73	526.03	Y
25	3PL	107.33	7	197260	26.81	526.03	Y
26	3PL	348.82	7	197260	91.35	526.03	Y
27	2PPC	779.51	17	197260	130.77	526.03	Y
28	2PPC	621.12	17	197260	103.61	526.03	Y
29	3PL	280.14	7	197260	73.00	526.03	Y
30	3PL	212.49	7	197260	54.92	526.03	Y
31	3PL	449.20	7	197260	118.18	526.03	Y
32	3PL	121.35	7	197260	30.56	526.03	Y
33	2PPC	1189.88	17	197260	201.15	526.03	Y
34	2PPC	402.57	17	197260	66.12	526.03	Y
35	2PPC	1884.71	26	197260	257.76	526.03	Y

**Table F6. ELA Item Fit Statistics, Grade 8**

Item Number	Model	Chi-Square	DF	N-count	Obtained Z	Critical Z	Fit Ok?
1	3PL	216.15	7	202479	55.90	539.94	Y
2	3PL	144.71	7	202479	36.80	539.94	Y
3	3PL	77.18	7	202479	18.76	539.94	Y
4	3PL	102.16	7	202479	25.43	539.94	Y
5	3PL	1491.78	7	202479	396.83	539.94	Y
6	3PL	44.03	7	202479	9.90	539.94	Y
7	3PL	995.17	7	202479	264.10	539.94	Y
8	3PL	407.08	7	202479	106.93	539.94	Y
9	3PL	167.22	7	202479	42.82	539.94	Y
10	3PL	107.39	7	202479	26.83	539.94	Y
11	3PL	55.61	7	202479	12.99	539.94	Y
12	3PL	299.15	7	202479	78.08	539.94	Y
13	3PL	434.50	7	202479	114.25	539.94	Y
14	3PL	72.39	7	202479	17.48	539.94	Y
15	3PL	134.71	7	202479	34.13	539.94	Y
16	3PL	368.84	7	202479	96.71	539.94	Y
17	3PL	309.00	7	202479	80.71	539.94	Y
18	3PL	487.06	7	202479	128.30	539.94	Y
19	3PL	406.83	7	202479	106.86	539.94	Y
20	3PL	285.46	7	202479	74.42	539.94	Y
21	3PL	319.91	7	202479	83.63	539.94	Y
22	3PL	584.56	7	202479	154.36	539.94	Y
23	3PL	115.64	7	202479	29.04	539.94	Y
24	3PL	468.25	7	202479	123.27	539.94	Y
25	3PL	322.87	7	202479	84.42	539.94	Y
26	3PL	291.69	7	202479	76.09	539.94	Y
27	2PPC	4052.58	44	202479	427.32	539.94	Y
28	2PPC	3520.35	44	202479	370.58	539.94	Y
29	2PPC	1451.31	26	202479	197.65	539.94	Y

## Appendix G—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  $\beta(r_j, s_j)$ :

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

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

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

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

where

$T_i$  is the expected value of the score for item  $i$  in standard  $j$  for a given  $\theta$ .

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

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

with

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

and

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

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

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

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

### ***Estimation of the Prior Distribution of $T_j$***

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

The 3PL model is

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

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 constructed-response 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 constructed-response item with  $l_i$  score levels, integer scores are assigned that ranged from 0 to  $l_i - 1$ :

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

where

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

and

$$\gamma_{i0} = 0$$

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

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  $\theta$  is the common trait value to which the items are scaled:

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

where

$l_i$  is the number of score levels in item  $i$ , including 0.

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

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

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

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

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

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

and

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

Solving these equations for  $r_j$  and  $s_j$  produces

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

and

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

where

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

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

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

Because  $T_j$  is a monotonic transformation of  $\theta$  (Lord, 1980, p.71):

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

where

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

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

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

and

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

Thus,

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

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 two-parameter partial-credit models. Furthermore, the parameters of the posterior distribution of  $T_j$  also can be expressed in terms of the IRT parameters:

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

and

$$q_j = [1 - \hat{T}_j] n_j^* + n_j - x_j. \quad (19)$$

The OPI is

$$\tilde{T}_j = \frac{p_j}{p_j + q_j} \quad (20)$$

$$= \frac{\hat{T}_j n_j^* + x_j}{n_j^* + n_j}. \quad (21)$$

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

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

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

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

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

### ***Check on Consistency and Adjustment of Weight Given to Prior***

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:

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

If  $Q \leq \chi^2(J, .10)$ , 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. Yen working with tests containing exclusively MC items, 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 H—Derivation of Classification Consistency and Accuracy

---

### *Classification Consistency*

Assume that  $\theta$  is a single latent trait measured by a test and denote  $\Phi$  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

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

where

$g(\theta)$  is the density of  $\theta$ .

In this report, the marginal distribution  $P(X = x)$  is denoted as  $f(x)$ , and the conditional error distribution  $P(X = x | \Phi = \theta)$  is denoted as  $f(x | \theta)$ . It is assumed that examinees are classified into one of  $H$  mutually exclusive categories on the basis of predetermined  $H-1$  observed score cutoffs,  $C_1, C_2, \dots, C_{H-1}$ . Let  $L_h$  represent the  $h^{\text{th}}$  category into which examinees with  $C_{h-1} \leq X \leq C_h$  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:

$$P(X \in L_h | \theta) = \sum_{x=C_{h-1}}^{C_h} f(x | \theta), \quad h = 1, 2, \dots, H.$$

$$P(X \in L_h) = \int \sum_{x=C_{h-1}}^{C_h} f(x | \theta)g(\theta)d\theta, \quad h = 1, 2, \dots, H.$$

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  $H \times H$  contingency table can be constructed. The elements of  $H \times H$  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  $\theta$ ,  $X_1$  and  $X_2$  are independent and identically distributed. Consequently, the conditional bivariate distribution of  $X_1$  and  $X_2$  is

$$f(x_1, x_2 | \theta) = f(x_1 | \theta)f(x_2 | \theta).$$

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

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

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

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

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

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

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.

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

Then, the coefficient kappa (Cohen, 1960) is

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

### ***Classification Accuracy***

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

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

where

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

## Appendix I—Scale Score Frequency Distributions

---

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

**Table I1. Grade 3 ELA 2008 SS Frequency Distribution, State**

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
475	458	0.23	458	0.23
531	396	0.20	854	0.44
563	582	0.30	1436	0.73
577	760	0.39	2196	1.12
586	905	0.46	3101	1.58
593	1020	0.52	4121	2.11
599	1157	0.59	5278	2.70
603	1348	0.69	6626	3.39
607	1476	0.75	8102	4.14
611	1626	0.83	9728	4.97
614	1891	0.97	11619	5.94
618	2176	1.11	13795	7.05
621	2526	1.29	16321	8.34
624	2826	1.44	19147	9.78
627	3314	1.69	22461	11.48
631	3812	1.95	26273	13.43
634	4475	2.29	30748	15.71
638	5292	2.70	36040	18.42
641	6164	3.15	42204	21.57
645	7280	3.72	49484	25.29
649	8975	4.59	58459	29.87
653	10634	5.43	69093	35.31
658	12482	6.38	81575	41.68
663	14988	7.66	96563	49.34
669	17103	8.74	113666	58.08

*(Continued on next page)*

**Table I1. Grade 3 ELA 2008 SS Frequency Distribution, State (cont.)**

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
676	19074	9.75	132740	67.83
685	19862	10.15	152602	77.98
697	18872	9.64	171474	87.62
720	15291	7.81	186765	95.44
780	8930	4.56	195695	100.00

**Table I2. Grade 4 ELA 2008 SS Frequency Distribution, State**

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
430	343	0.17	343	0.17
496	247	0.13	590	0.30
525	367	0.19	957	0.49
542	507	0.26	1464	0.74
555	632	0.32	2096	1.06
566	788	0.40	2884	1.46
574	1056	0.54	3940	2.00
582	1239	0.63	5179	2.63
589	1449	0.74	6628	3.37
595	1656	0.84	8284	4.21
600	1879	0.95	10163	5.16
605	2182	1.11	12345	6.27
609	2377	1.21	14722	7.48
613	2648	1.34	17370	8.82
617	2957	1.50	20327	10.32
621	3348	1.70	23675	12.02
625	3672	1.86	27347	13.89
629	3880	1.97	31227	15.86
632	4222	2.14	35449	18.00
636	4714	2.39	40163	20.40
639	5103	2.59	45266	22.99
643	5564	2.83	50830	25.81
646	6013	3.05	56843	28.87

*(Continued on next page)*

**Table I2. Grade 4 ELA 2008 SS Frequency Distribution, State (cont.)**

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
650	6528	3.32	63371	32.18
653	7084	3.60	70455	35.78
657	7666	3.89	78121	39.67
660	8211	4.17	86332	43.84
664	8917	4.53	95249	48.37
668	9532	4.84	104781	53.21
673	10364	5.26	115145	58.47
677	10804	5.49	125949	63.96
682	11576	5.88	137525	69.84
688	12001	6.09	149526	75.93
694	11699	5.94	161225	81.88
702	10559	5.36	171784	87.24
711	8552	4.34	180336	91.58
723	6975	3.54	187311	95.12
742	6574	3.34	193885	98.46
775	3030	1.54	196915	100.00

**Table I3. Grade 5 ELA 2008 SS Frequency Distribution, State**

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
495	452	0.23	452	0.23
538	447	0.23	899	0.45
579	680	0.34	1579	0.80
594	948	0.48	2527	1.28
603	1138	0.58	3665	1.85
609	1538	0.78	5203	2.63
614	1668	0.84	6871	3.47
619	1982	1.00	8853	4.47
623	2257	1.14	11110	5.61
627	2704	1.37	13814	6.98
630	3071	1.55	16885	8.53
634	3755	1.90	20640	10.43

*(Continued on next page)*

**Table I3. Grade 5 ELA 2008 SS Frequency Distribution, State (cont.)**

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
637	4536	2.29	25176	12.72
640	5366	2.71	30542	15.43
644	6325	3.20	36867	18.63
647	7425	3.75	44292	22.38
650	8639	4.37	52931	26.75
654	10169	5.14	63100	31.88
657	11649	5.89	74749	37.77
661	13422	6.78	88171	44.55
666	15266	7.71	103437	52.27
670	16744	8.46	120181	60.73
676	17503	8.84	137684	69.57
682	17984	9.09	155668	78.66
690	16771	8.47	172439	87.13
701	13747	6.95	186186	94.08
718	8653	4.37	194839	98.45
795	3062	1.55	197901	100.00

**Table I4. Grade 6 ELA 2008 SS Frequency Distribution, State**

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
480	393	0.20	393	0.20
552	313	0.16	706	0.35
571	426	0.21	1132	0.57
582	567	0.28	1699	0.85
590	766	0.38	2465	1.23
597	961	0.48	3426	1.71
602	1144	0.57	4570	2.28
607	1285	0.64	5855	2.92
611	1521	0.76	7376	3.68
615	1755	0.88	9131	4.56
618	1919	0.96	11050	5.52
621	2147	1.07	13197	6.59
624	2481	1.24	15678	7.83

*(Continued on next page)*

**Table I4. Grade 6 ELA 2008 SS Frequency Distribution, State (cont.)**

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
626	2763	1.38	18441	9.20
629	3099	1.55	21540	10.75
631	3534	1.76	25074	12.51
634	3975	1.98	29049	14.50
636	4485	2.24	33534	16.74
639	5040	2.52	38574	19.25
641	5670	2.83	44244	22.08
644	6521	3.25	50765	25.34
646	7178	3.58	57943	28.92
649	8191	4.09	66134	33.01
652	9261	4.62	75395	37.63
655	10493	5.24	85888	42.87
658	11891	5.94	97779	48.80
662	13216	6.60	110995	55.40
666	14509	7.24	125504	62.64
670	14977	7.48	140481	70.12
676	14876	7.42	155357	77.54
682	13887	6.93	169244	84.47
689	12163	6.07	181407	90.54
699	9601	4.79	191008	95.34
715	6279	3.13	197287	98.47
785	3065	1.53	200352	100.00

**Table I5. Grade 7 ELA 2008 SS Frequency Distribution, State**

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
470	435	0.21	435	0.21
546	351	0.17	786	0.38
567	484	0.23	1270	0.61
579	662	0.32	1932	0.93
588	820	0.40	2752	1.33

*(Continued on next page)*

**Table I5. Grade 7 ELA 2008 SS Frequency Distribution, State (cont.)**

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
595	1074	0.52	3826	1.85
601	1263	0.61	5089	2.46
606	1396	0.67	6485	3.13
611	1718	0.83	8203	3.97
615	1902	0.92	10105	4.88
618	2177	1.05	12282	5.94
622	2372	1.15	14654	7.08
625	2730	1.32	17384	8.40
628	2873	1.39	20257	9.79
631	3403	1.64	23660	11.44
633	3774	1.82	27434	13.26
636	4282	2.07	31716	15.33
638	4724	2.28	36440	17.61
641	5303	2.56	41743	20.18
643	5930	2.87	47673	23.04
646	6621	3.20	54294	26.25
648	7516	3.63	61810	29.88
651	8292	4.01	70102	33.89
654	9287	4.49	79389	38.38
657	10412	5.03	89801	43.41
660	11731	5.67	101532	49.08
664	12962	6.27	114494	55.35
667	13698	6.62	128192	61.97
672	14181	6.85	142373	68.82
676	14322	6.92	156695	75.75
682	14050	6.79	170745	82.54
689	12590	6.09	183335	88.62
697	10634	5.14	193969	93.76
709	7634	3.69	201603	97.45
729	4080	1.97	205683	99.43
790	1188	0.57	206871	100.00

**Table I6. Grade 8 ELA 2008 SS Frequency Distribution, State**

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
430	264	0.13	264	0.13
447	177	0.08	441	0.21
527	251	0.12	692	0.33
544	334	0.16	1026	0.49
555	425	0.20	1451	0.69
563	574	0.27	2025	0.97
569	696	0.33	2721	1.30
575	740	0.35	3461	1.66
580	808	0.39	4269	2.04
585	979	0.47	5248	2.51
589	1157	0.55	6405	3.07
593	1286	0.62	7691	3.68
597	1434	0.69	9125	4.37
600	1665	0.80	10790	5.16
604	1864	0.89	12654	6.06
607	2087	1.00	14741	7.05
610	2424	1.16	17165	8.21
613	2852	1.36	20017	9.58
616	3107	1.49	23124	11.07
619	3708	1.77	26832	12.84
622	4148	1.99	30980	14.83
625	4926	2.36	35906	17.18
629	5527	2.65	41433	19.83
632	6199	2.97	47632	22.79
635	7166	3.43	54798	26.22
638	7852	3.76	62650	29.98
642	8811	4.22	71461	34.20
645	9579	4.58	81040	38.78
649	10460	5.01	91500	43.79
653	11214	5.37	102714	49.16
657	11647	5.57	114361	54.73
661	12311	5.89	126672	60.62
666	12403	5.94	139075	66.56
671	12653	6.06	151728	72.61

*(Continued on next page)*

**Table I6. Grade 8 ELA 2008 SS Frequency Distribution, State (cont.)**

SS	N-count	Percent	Cumulative Frequency	Cumulative Percent
677	12301	5.89	164029	78.50
684	11623	5.56	175652	84.06
693	11049	5.29	186701	89.35
705	10361	4.96	197062	94.31
726	8218	3.93	205280	98.24
790	3679	1.76	208959	100.00

## References

---

- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 1999. *Standards for educational and psychological testing*. Washington, D.C.: American Psychological Association, Inc.
- Bock, R. D. 1972. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika* 37: 29–51.
- Bock, R. D. and M. Aitkin. 1981. Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika* 46: 443–459.
- Burket, G. R. 1988. *ITEMWIN* [Computer program].
- Burket, G. R. 2002. *PARDUX* [Computer program].
- Cattell, R.B. 1966. The Scree Test for the Number of Factors. *Multivariate Behavioral Research* 1: 245–276.
- Cronbach, L. J. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* 16: 297–334.
- 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 Measurement* 18: 109–118.

- Livingston, S. A. and C. Lewis. 1995. Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement* 32: 179–197.
- Lord, F. M. 1980. *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Lord, F. M. and M. R. Novick. 1968. *Statistical theories of mental test scores*. Menlo Park, CA: Addison-Wesley.
- Mehrens, W. A. and I. J. Lehmann. 1991. *Measurement and Evaluation in Education and Psychology, 3rd ed.* New York: Holt, Rinehart, and Winston.
- Muraki, E. 1992. A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement* 16: 159–176.
- Muraki, E., and R. D. Bock. 1991. *PARSCALE: Parameter Scaling of Rating Data* [Computer program]. Chicago, 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, IL, 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.