

# **New York State Testing Program 2008: Mathematics, 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>COPYRIGHT .....</b>	<b>2</b>
<b>TABLE OF CONTENTS .....</b>	<b>0</b>
<b>LIST OF TABLES .....</b>	<b>3</b>
<b>SECTION I: INTRODUCTION AND OVERVIEW .....</b>	<b>2</b>
INTRODUCTION .....	2
TEST PURPOSE .....	2
TARGET POPULATION .....	2
TEST USE AND DECISIONS BASED ON ASSESSMENT .....	2
<i>Scale Scores</i> .....	2
<i>Proficiency Level Cut Score and Classification</i> .....	3
<i>Standard Performance Index Scores</i> .....	3
TESTING ACCOMMODATIONS .....	3
TEST TRANSCRIPTIONS .....	3
TEST TRANSLATIONS .....	4
<b>SECTION II: TEST DESIGN AND DEVELOPMENT .....</b>	<b>5</b>
TEST DESCRIPTION .....	5
TEST CONFIGURATION .....	5
TEST BLUEPRINT .....	6
2008 ITEM MAPPING BY NEW YORK STATE STANDARDS AND STRANDS .....	14
NEW YORK STATE EDUCATOR'S INVOLVEMENT IN TEST DEVELOPMENT .....	16
CONTENT RATIONALE .....	16
ITEM DEVELOPMENT .....	17
ITEM REVIEW .....	17
MATERIALS DEVELOPMENT .....	18
ITEM SELECTION AND TEST CREATION (CRITERIA AND PROCESS) .....	18
PROFICIENCY AND PERFORMANCE STANDARDS .....	19
<b>SECTION III: VALIDITY .....</b>	<b>20</b>
CONTENT VALIDITY .....	20
CONSTRUCT (INTERNAL STRUCTURE) VALIDITY .....	21
<i>Internal Consistency</i> .....	21
<i>Unidimensionality</i> .....	21
<i>Minimization of Bias</i> .....	23
<b>SECTION IV: TEST ADMINISTRATION AND SCORING .....</b>	<b>25</b>
TEST ADMINISTRATION .....	25
SCORING PROCEDURES OF OPERATIONAL TESTS .....	25
SCORING MODELS .....	25
SCORING OF CONSTRUCTED-RESPONSE ITEMS .....	26
SCORER QUALIFICATIONS AND TRAINING .....	27
QUALITY CONTROL PROCESS .....	27
<b>SECTION V: OPERATIONAL TEST DATA COLLECTION AND CLASSICAL ANALYSIS .....</b>	<b>28</b>
DATA COLLECTION .....	28
DATA PROCESSING .....	28

CLASSICAL ANALYSIS AND CALIBRATION SAMPLE CHARACTERISTICS .....	30
CLASSICAL DATA ANALYSIS .....	34
<i>Item Rescoring and Replacing</i> .....	35
<i>Item Difficulty and Response Distribution</i> .....	35
<i>Point-Biserial Correlation Coefficients</i> .....	42
<i>Distractor Analysis</i> .....	43
<i>Test Statistics and Reliability Coefficients</i> .....	43
<i>Speededness</i> .....	44
<i>Differential Item Functioning</i> .....	44
<b>SECTION VI: IRT SCALING AND EQUATING .....</b>	<b>46</b>
IRT MODELS AND RATIONALE FOR USE .....	46
CALIBRATION SAMPLE .....	47
CALIBRATION PROCESS .....	47
ITEM-MODEL FIT .....	48
LOCAL INDEPENDENCE .....	49
SCALING AND EQUATING .....	50
<i>Anchor Set Security</i> .....	52
<i>Anchor Item Evaluation</i> .....	52
ITEM PARAMETERS .....	58
TEST CHARACTERISTIC CURVES .....	64
SCORING PROCEDURE .....	68
RAW SCORE-TO-SCALE SCORE AND SEM CONVERSION TABLES .....	68
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 for MC Items</i> .....	84
<i>Reliability for CR Items</i> .....	84
<i>Test Reliability for NCLB Reporting Categories</i> .....	84
STANDARD ERROR OF MEASUREMENT .....	90
PERFORMANCE LEVEL CLASSIFICATION CONSISTENCY AND ACCURACY .....	90
<i>Consistency</i> .....	91
<i>Accuracy</i> .....	92
<b>SECTION VIII: SUMMARY OF OPERATIONAL TEST RESULTS .....</b>	<b>94</b>
SCALE SCORE DISTRIBUTION SUMMARY .....	94
<i>Grade 3</i> .....	95
<i>Grade 4</i> .....	96
<i>Grade 5</i> .....	98
<i>Grade 6</i> .....	99
<i>Grade 7</i> .....	100
<i>Grade 8</i> .....	102
PERFORMANCE LEVEL DISTRIBUTION SUMMARY .....	103
<i>Grade 3</i> .....	104
<i>Grade 4</i> .....	105
<i>Grade 5</i> .....	106
<i>Grade 6</i> .....	108
<i>Grade 7</i> .....	109
<i>Grade 8</i> .....	110
<b>SECTION IX: LONGITUDINAL COMPARISON OF RESULTS .....</b>	<b>112</b>

<b>APPENDIX A—CRITERIA FOR ITEM ACCEPTABILITY .....</b>	<b>114</b>
<b>APPENDIX B—PSYCHOMETRIC GUIDELINES FOR OPERATIONAL ITEM SELECTION .....</b>	<b>116</b>
<b>APPENDIX C—FACTOR ANALYSIS RESULTS.....</b>	<b>118</b>
<b>APPENDIX D—ITEMS FLAGGED FOR DIF.....</b>	<b>121</b>
<b>APPENDIX E—ITEM-MODEL FIT STATISTICS.....</b>	<b>125</b>
<b>APPENDIX F—DERIVATION OF THE GENERALIZED SPI PROCEDURE ..</b>	<b>131</b>
<b>APPENDIX G—DERIVATION OF CLASSIFICATION CONSISTENCY AND ACCURACY .....</b>	<b>137</b>
CLASSIFICATION CONSISTENCY.....	137
CLASSIFICATION ACCURACY.....	138
<b>APPENDIX H—SCALE SCORE FREQUENCY DISTRIBUTIONS.....</b>	<b>139</b>
<b>REFERENCES.....</b>	<b>149</b>

## List of Tables

---

TABLE 1. NYSTP MATHEMATICS 2008 TEST CONFIGURATION.....	5
TABLE 2. NYSTP MATHEMATICS 2008 TEST BLUEPRINT .....	6
TABLE 3A. NYSTP MATHEMATICS 2008 OPERATIONAL TEST MAP, GRADE 3.....	8
TABLE 3B. NYSTP MATHEMATICS 2008 OPERATIONAL TEST MAP, GRADE 4.....	9
TABLE 3C. NYSTP MATHEMATICS 2008 OPERATIONAL TEST MAP, GRADE 5.....	10
TABLE 3D. NYSTP MATHEMATICS 2008 OPERATIONAL TEST MAP, GRADE 6.....	11
TABLE 3E. NYSTP MATHEMATICS 2008 OPERATIONAL TEST MAP, GRADE 7 .....	12
TABLE 3F. NYSTP MATHEMATICS 2008 OPERATIONAL TEST MAP, GRADE 8.....	13
TABLE 4. NYSTP MATHEMATICS 2008 STRAND COVERAGE .....	14
TABLE 5. FACTOR ANALYSIS RESULTS FOR MATHEMATICS TESTS (TOTAL POPULATION) .....	22
TABLE 6A. NYSTP MATHEMATICS DATA CLEANING, GRADE 3.....	28
TABLE 6B. NYSTP MATHEMATICS DATA CLEANING, GRADE 4.....	29
TABLE 6C. NYSTP MATHEMATICS DATA CLEANING, GRADE 5.....	29
TABLE 6D. NYSTP MATHEMATICS DATA CLEANING, GRADE 6.....	29
TABLE 6E. NYSTP MATHEMATICS DATA CLEANING, GRADE 7.....	30
TABLE 6F. NYSTP MATHEMATICS DATA CLEANING, GRADE 8.....	30
TABLE 7A. GRADE 3 SAMPLE CHARACTERISTICS (N = 193566) .....	31
TABLE 7B. GRADE 4 SAMPLE CHARACTERISTICS (N = 195350) .....	31
TABLE 7C. GRADE 5 SAMPLE CHARACTERISTICS (N = 196251) .....	32
TABLE 7D. GRADE 6 SAMPLE CHARACTERISTICS (N = 198436) .....	32
TABLE 7E. GRADE 7 SAMPLE CHARACTERISTICS (N = 205232) .....	33
TABLE 7F. GRADE 8 SAMPLE CHARACTERISTICS (N = 206444).....	34
TABLE 8A. P-VALUES, SCORED RESPONSE DISTRIBUTIONS, AND POINT BISERIALS, GRADE 3.....	36
TABLE 8B. P-VALUES, SCORED RESPONSE DISTRIBUTIONS, AND POINT BISERIALS, GRADE 4.....	37

<b>TABLE 8C. P-VALUES, SCORED RESPONSE DISTRIBUTIONS, AND POINT BISERIALS, GRADE 5.....</b>	<b>38</b>
<b>TABLE 8D. P-VALUES, SCORED RESPONSE DISTRIBUTIONS, AND POINT BISERIALS, GRADE 6.....</b>	<b>39</b>
<b>TABLE 8E. P-VALUES, SCORED RESPONSE DISTRIBUTIONS, AND POINT BISERIALS, GRADE 7.....</b>	<b>40</b>
<b>TABLE 8F. P-VALUES, SCORED RESPONSE DISTRIBUTIONS, AND POINT BISERIALS, GRADE 8.....</b>	<b>41</b>
<b>TABLE 9. NYSTP MATHEMATICS 2008 TEST FORM STATISTICS AND RELIABILITY .....</b>	<b>44</b>
<b>TABLE 10. NYSTP MATHEMATICS 2008 CLASSICAL DIF SAMPLE N-COUNTS.....</b>	<b>45</b>
<b>TABLE 11. NUMBER OF ITEMS FLAGGED BY SMD AND MANTEL-HAENZEL DIF METHODS .....</b>	<b>45</b>
<b>TABLE 12. NYSTP MATHEMATICS 2008 CALIBRATION RESULTS.....</b>	<b>48</b>
<b>TABLE 13. NYSTP MATHEMATICS 2008 FINAL TRANSFORMATION CONSTANTS .....</b>	<b>52</b>
<b>TABLE 14. MATHEMATICS ANCHOR EVALUATION SUMMARY.....</b>	<b>53</b>
<b>TABLE 15A. 2008 OPERATIONAL ITEM PARAMETER ESTIMATES, GRADE 3.....</b>	<b>58</b>
<b>TABLE 15B. 2008 OPERATIONAL ITEM PARAMETER ESTIMATES, GRADE 4.....</b>	<b>59</b>
<b>TABLE 15C. 2008 OPERATIONAL ITEM PARAMETER ESTIMATES, GRADE 5.....</b>	<b>60</b>
<b>TABLE 15D. 2008 OPERATIONAL ITEM PARAMETER ESTIMATES, GRADE 6.....</b>	<b>61</b>
<b>TABLE 15E. 2008 OPERATIONAL ITEM PARAMETER ESTIMATES, GRADE 7.....</b>	<b>62</b>
<b>TABLE 15F. 2008 OPERATIONAL ITEM PARAMETER ESTIMATES, GRADE 8.....</b>	<b>63</b>
<b>TABLE 16A. GRADE 3 RAW SCORE TO SCALE SCORE (WITH STANDARD ERROR).....</b>	<b>69</b>
<b>TABLE 16B. GRADE 4 RAW SCORE TO SCALE SCORE (WITH STANDARD ERROR).....</b>	<b>70</b>
<b>TABLE 16C. GRADE 5 RAW SCORE TO SCALE SCORE (WITH STANDARD ERROR).....</b>	<b>72</b>
<b>TABLE 16D. GRADE 6 RAW SCORE TO SCALE SCORE (WITH STANDARD ERROR).....</b>	<b>73</b>

<b>TABLE 16E. GRADE 7 RAW SCORE TO SCALE SCORE (WITH STANDARD ERROR).....</b>	<b>75</b>
<b>TABLE 16F. GRADE 8 RAW SCORE TO SCALE SCORE (WITH STANDARD ERROR).....</b>	<b>76</b>
<b>TABLE 17. SPI TARGET RANGES .....</b>	<b>79</b>
<b>TABLE 18. NUMBER OF ITEMS FLAGGED FOR DIF BY THE LINN-HARNISCH METHOD.....</b>	<b>82</b>
<b>TABLE 19. RELIABILITY AND STANDARD ERROR OF MEASUREMENT ...</b>	<b>83</b>
<b>TABLE 20. RELIABILITY AND STANDARD ERROR OF MEASUREMENT—MC ITEMS ONLY .....</b>	<b>84</b>
<b>TABLE 21. RELIABILITY AND STANDARD ERROR OF MEASUREMENT—CR ITEMS ONLY .....</b>	<b>84</b>
<b>TABLE 22A. GRADE 3 TEST RELIABILITY BY SUBGROUP .....</b>	<b>85</b>
<b>TABLE 22B. GRADE 4 TEST RELIABILITY BY SUBGROUP .....</b>	<b>86</b>
<b>TABLE 22C. GRADE 5 TEST RELIABILITY BY SUBGROUP.....</b>	<b>86</b>
<b>TABLE 22D. GRADE 6 TEST RELIABILITY BY SUBGROUP .....</b>	<b>87</b>
<b>TABLE 22E. GRADE 7 TEST RELIABILITY BY SUBGROUP .....</b>	<b>88</b>
<b>TABLE 22F. GRADE 8 TEST RELIABILITY BY SUBGROUP .....</b>	<b>89</b>
<b>TABLE 23. DECISION CONSISTENCY (ALL CUTS).....</b>	<b>92</b>
<b>TABLE 24. DECISION CONSISTENCY (LEVEL III CUT).....</b>	<b>92</b>
<b>TABLE 25. DECISION AGREEMENT (ACCURACY) .....</b>	<b>93</b>
<b>TABLE 26. MATHEMATICS SCALE SCORE DISTRIBUTION SUMMARY GRADES 3–8.....</b>	<b>94</b>
<b>TABLE 27. SCALE SCORE DISTRIBUTION SUMMARY, BY SUBGROUP, GRADE 3.....</b>	<b>95</b>
<b>TABLE 28. SCALE SCORE DISTRIBUTION SUMMARY, BY SUBGROUP, GRADE 4.....</b>	<b>97</b>
<b>TABLE 29. SCALE SCORE DISTRIBUTION SUMMARY, BY SUBGROUP, GRADE 5.....</b>	<b>98</b>
<b>TABLE 30. SCALE SCORE DISTRIBUTION SUMMARY, BY SUBGROUP, GRADE 6.....</b>	<b>100</b>
<b>TABLE 31. SCALE SCORE DISTRIBUTION SUMMARY, BY SUBGROUP, GRADE 7.....</b>	<b>101</b>
<b>TABLE 32. SCALE SCORE DISTRIBUTION SUMMARY, BY SUBGROUP, GRADE 8.....</b>	<b>102</b>

<b>TABLE 33. MATHEMATICS GRADES 3–8 PERFORMANCE LEVEL CUT SCORES.....</b>	<b>103</b>
<b>TABLE 34. MATHEMATICS TEST PERFORMANCE LEVEL DISTRIBUTIONS GRADES 3–8.....</b>	<b>104</b>
<b>TABLE 35. PERFORMANCE LEVEL DISTRIBUTIONS, BY SUBGROUP, GRADE 3.....</b>	<b>104</b>
<b>TABLE 36. PERFORMANCE LEVEL DISTRIBUTION SUMMARY, BY SUBGROUP, GRADE 4.....</b>	<b>106</b>
<b>TABLE 37. PERFORMANCE LEVEL DISTRIBUTION SUMMARY, BY SUBGROUP, GRADE 5.....</b>	<b>107</b>
<b>TABLE 38. PERFORMANCE LEVEL DISTRIBUTION SUMMARY, BY SUBGROUP, GRADE 6.....</b>	<b>108</b>
<b>TABLE 39. PERFORMANCE LEVEL DISTRIBUTION SUMMARY, BY SUBGROUP, GRADE 7.....</b>	<b>110</b>
<b>TABLE 40. PERFORMANCE LEVEL DISTRIBUTION SUMMARY, BY SUBGROUP, GRADE 8.....</b>	<b>111</b>
<b>TABLE 41. MATHEMATICS GRADES 3–8 TEST LONGITUDINAL RESULTS .....</b>	<b>112</b>
<b>TABLE C1. FACTOR ANALYSIS RESULTS FOR MATHEMATICS TESTS (SELECTED SUBPOPULATIONS).....</b>	<b>118</b>
<b>TABLE D1. NYSTP MATHEMATICS 2008 CLASSICAL DIF ITEM FLAGS... ..</b>	<b>121</b>
<b>TABLE D2. ITEMS FLAGGED FOR DIF BY THE LINN-HARNISCH METHOD .....</b>	<b>124</b>
<b>TABLE E1. MATHEMATICS GRADE 3 ITEM FIT STATISTICS.....</b>	<b>125</b>
<b>TABLE E2. MATHEMATICS GRADE 4 ITEM FIT STATISTICS.....</b>	<b>126</b>
<b>TABLE E3. MATHEMATICS GRADE 5 ITEM FIT STATISTICS.....</b>	<b>127</b>
<b>TABLE E4. MATHEMATICS GRADE 6 ITEM FIT STATISTICS.....</b>	<b>128</b>
<b>TABLE E5. MATHEMATICS GRADE 7 ITEM FIT STATISTICS.....</b>	<b>129</b>
<b>TABLE E6. MATHEMATICS GRADE 8 ITEM FIT STATISTICS.....</b>	<b>130</b>
<b>TABLE H1. GRADE 3 MATHEMATICS 2008 SS FREQUENCY DISTRIBUTION, STATE .....</b>	<b>139</b>
<b>TABLE H2. GRADE 4 MATHEMATICS 2008 SS FREQUENCY DISTRIBUTION, STATE .....</b>	<b>140</b>
<b>TABLE H3. GRADE 5 MATHEMATICS 2008 SS FREQUENCY DISTRIBUTION, STATE .....</b>	<b>142</b>
<b>TABLE H4. GRADE 6 MATHEMATICS 2008 SS FREQUENCY DISTRIBUTION, STATE .....</b>	<b>144</b>

**TABLE H5. GRADE 7 MATHEMATICS 2008 SS FREQUENCY DISTRIBUTION,  
STATE ..... 145**

**TABLE H6. GRADE 8 MATHEMATICS 2008 SS FREQUENCY DISTRIBUTION,  
STATE ..... 147**

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

---

### ***Introduction***

An overview of the New York State Testing Program (NYSTP), Grades 3–8, Mathematics 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 State. The Mathematics Tests target student progress toward five content standards in Grades 3–7 and four content standards in Grade 8 as described in Section II, “Test Design and Development,” subsection “Content Rationale.” The Grades 3–8 Mathematics Tests are written for all students to have the opportunity to demonstrate their knowledge and skills in these standards. The established cut scores classify student proficiency into one of four levels based on their test performance.

### ***Target Population***

Students in New York State public schools, Grades 3, 4, 5, 6, 7, and 8 (and ungraded students of equivalent age) are the target population for the Grades 3–8 Mathematics Tests. Nonpublic schools may participate in the testing program but the participation is not mandatory for them. In 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/elintmath.html>.

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

The Grades 3–8 Mathematics Tests are used to measure the extent to which individual students achieve the New York State Learning Standards in mathematics and to determine whether schools, districts, and the State meet the required progress targets specified in the New York State accountability system. There are several types of scores available from the Grades 3–8 Mathematics Tests and these are discussed in this section.

#### **Scale Scores**

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

Equating.” Uses of Grades 3–8 Mathematics Test 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 Score 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 mathematics proficiency cut scores reflect the abilities intended by the New York State Education Department. Performance of students on the Grades 3–8 Mathematics 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. More 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 Mathematics* and the *NYS Measurement Review Technical Report 2006 for Mathematics*.

### **Standard Performance Index Scores**

Standard performance index (SPI) scores are obtained from the Grades 3–8 Mathematics Tests. The SPI score is an indicator of student ability, knowledge, and skills in specific learning standards and is used primarily for diagnostic purposes to help teachers evaluate academic strengths and weaknesses of their students. These scores can be effectively used by teachers at the classroom level to modify their instructional content and format to best serve their students’ specific needs. Detailed information on the properties and 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 files for the large type editions are created by CTB/McGraw-Hill and printed by NYSED, and the braille editions are produced by Braille Publishers, Inc. The lead transcribers are members of 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 testing program.

Camera-copy versions of the regular tests are provided to the braille vendor, who then proceeds to create the braille editions. Proofs of the braille editions are submitted to NYSED for review and approval prior to reproduction of the braille editions.

### ***Test Translations***

Since these are tests of mathematical ability, the NYSTP 3–8 Mathematics tests are translated into five other languages: Chinese, Haitian-Creole, Korean, Russian, and Spanish. These tests are translated to provide students the opportunity to demonstrate mathematical ability independent of their command of the English language. Sample tests are available in each translated language at the following locations:

<http://www.emsc.nysed.gov/3-8/math-sample/chinese/home.htm> (Chinese),  
<http://www.emsc.nysed.gov/3-8/math-sample/haitian/home.htm> (Haitian-Creole),  
<http://www.emsc.nysed.gov/3-8/math-sample/korean/home.htm> (Korean),  
<http://www.emsc.nysed.gov/3-8/math-sample/russian/home.htm> (Russian),  
<http://www.emsc.nysed.gov/3-8/math-sample/spanish/home.htm> (Spanish).

In addition, each year’s operational test translations are released and posted to NYSED’s web site after the testing administration window is over.

Limited English proficient (LEP) students may be provided with an oral translation of the mathematics tests when a written translation is not available in the student’s first language. The following testing accommodations were made available to LEP students: time extension, separate testing location, bilingual glossaries, simultaneous use of English and alternative language editions, oral translation for lower-incidence languages, and writing responses in the native language.

## Section II: Test Design and Development

### *Test Description*

The Grades 3–8 Mathematics Tests are New York State Learning Standards-based criterion-referenced tests composed of multiple-choice (MC) and constructed-response (CR) items differentiated by maximum score point. MC items have a maximum score of 1, short-response (SR) items have a maximum score of 2, and extended response (ER) items have a maximum score of 3. The tests were administered in New York State classrooms during March 2008 over a two-day period for Grades 3, 5, 6, and 7 and over a three-day period for Grades 4 and 8. The tests were printed in black and white and incorporated the concepts of universal design. Copies of the operational tests are available online at <http://www.nysedregents.org/testing/mathei/08exams/home.htm>.

Details on the administration and scoring of these tests can be found in Section IV, “Test Administration and Scoring.”

### *Test Configuration*

The OP tests books were administered, in order, on two 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. Book 1 contained only MC items. Book 2 and Book 3 contained only CR items. The 2008 *Teacher’s Directions* (available online at <http://www.nysedregents.org/testing/mathei/08exams/home.htm>) and the 2008 *School Administrator’s Manual* provide more detail on security, scheduling, classroom organization and preparation, test materials, and administration.

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

Grade	Day	Book	Number of Items				Allotted Time ( minutes)	
			MC	SR	ER	Total	Testing	Prep
3	1	1	25	0	0	25	45	10
	2	2	0	4	2	6	40	10
	Totals		25	4	2	31	85	20
4	1	1	30	0	0	30	50	10
	2	2	0	7	2	9	50	10
	3	3	0	7	2	9	50	10
	Totals		30	14	4	48	150	30
5	1	1	26	0	0	26	45	10
	2	2	0	4	4	8	50	10
	Totals		26	4	4	34	95	20
6	1	1	25	0	0	25	45	10
	2	2	0	6	4	10	60	10
	Totals		25	6	4	35	105	20

(Continued on next page)

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

Grade	Day	Book	Number of Items				Allotted Time ( minutes)	
			MC	SR	ER	Total	Testing	Prep
7	1	1	30	0	0	30	55	10
	2	2	0	4	4	8	55	10
	Totals		30	4	4	38	110	20
8	1	1	27	0	0	27	50	10
	1	2	0	4	2	6	40	10
	2	3	0	8	4	12	70	10
	Totals		27	12	6	45	160	30

***Test Blueprint***

The NYSTP Mathematics Tests assess students on the content and process strands of New York State Mathematics Learning Standard 3. The test items are indicators used to assess a variety mathematics skills and abilities. Each item is aligned with one content-performance indicator for reporting purposes but is also aligned to one or more process-performance indicators, as appropriate for the concepts embodied in the task. As a result of the alignment to both process and content strands, the tests assess students' conceptual understanding, procedural fluency, and problem-solving abilities, rather than solely assessing their knowledge of isolated skills and facts. The five content strands, to which the items are aligned for reporting purposes, are Number Sense and Operations, Algebra, Geometry, Measurement, and Statistics and Probability. The distribution of score points across the strands was determined during blueprint specifications meetings held with panels of New York State educators at the start of the testing program, prior to item development. The distribution in each grade reflects the number of assessable performance indicators in each strand at that grade and the emphasis placed on those performance indicators by the blueprint-specifications panel members. Table 2 shows the Grades 3–8 Mathematics Test blueprint and actual number of score points in 2008 OP tests.

**Table 2. NYSTP Mathematics 2008 Test Blueprint**

Grade	Total Points	Content Strand	Target # Points	Selected # Points	Target % of Test	Selected % of Test
3	39	Number Sense and Operations	19	16	48.0	41.0
		Algebra	5	7	13.0	18.0
		Geometry	5	5	13.0	13.0
		Measurement	5	5	13.0	13.0
		Statistics and Probability	5	6	13.0	15.0

*(Continued on next page)*

**Table 2. NYSTP Mathematics 2008 Test Blueprint (cont.)**

Grade	Total Points	Content Strand	Target # Points	Selected # Points	Target % of Test	Selected % of Test
4	70	Number Sense and Operations	32	33	45.0	47.0
		Algebra	10	11	14.0	16.0
		Geometry	8	9	12.0	13.0
		Measurement	12	10	17.0	14.0
		Statistics and Probability	8	7	12.0	10.0
5	46	Number Sense and Operations	18	15	39.0	33.0
		Algebra	5	6	11.0	13.0
		Geometry	12	12	25.0	26.0
		Measurement	6	6	14.0	13.0
		Statistics and Probability	5	7	11.0	15.0
6	49	Number Sense and Operations	18	15	37.0	31.0
		Algebra	9	9	19.0	18.0
		Geometry	8	7	16.5	14.0
		Measurement	6	6	11.0	12.0
		Statistics and Probability	8	12	16.5	25.0
7	50	Number Sense and Operations	15	13	30.0	26.0
		Algebra	6	8	12.0	16.0
		Geometry	7	7	14.0	14.0
		Measurement	7	7	14.0	14.0
		Statistics and Probability	15	15	30.0	30.0
8	69	Number Sense and Operations	8	7	11.0	10.0
		Algebra	30	28	44.0	41.0
		Geometry	24	27	35.0	39.0
		Measurement	7	7	10.0	10.0

Tables 3a–3f present Grades 3–8 Mathematics Test item maps with the item type indicator, the answer key, the maximum number of points obtainable from each item, the current strand, and the performance indicator.

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

Item #	Item Type	Answer Key	Max Points	Content Strand	Performance Indicator
1	MC	B	1	1	3.N.2
2	MC	C	1	4	3.M.2
3	MC	A	1	4	3.M.7
4	MC	B	1	1	3.N.6
5	MC	C	1	1	3.N.18
6	MC	B	1	1	3.N.16
7	MC	B	1	3	3.G.5
8	MC	C	1	4	3.M.7
9	MC	A	1	1	3.N.6
10	MC	A	1	2	3.A.2
11	MC	C	1	1	3.N.19
12	MC	C	1	1	3.N.3
13	MC	B	1	1	3.N.21
14	MC	D	1	1	3.N.13
15	MC	D	1	1	3.N.10
16	MC	A	1	4	3.M.1
17	MC	D	1	3	3.G.1
18	MC	B	1	1	3.N.7
19	MC	C	1	2	3.A.1
20	MC	B	1	1	3.N.10
21	MC	C	1	3	3.G.4
22	MC	B	1	4	3.M.9
23	MC	C	1	2	3.A.2
24	MC	D	1	2	3.A.1
25	MC	D	1	5	3.S.7
26	SR	n/a	2	1	3.N.18
27	SR	n/a	2	3	3.G.1
28	SR	n/a	2	1	3.N.18
29	SR	n/a	2	5	3.S.7
30	ER	n/a	3	2	3.A.2
31	ER	n/a	3	5	3.S.5

**Table 3b. NYSTP Mathematics 2008 Operational Test Map, Grade 4**

Item #	Item Type	Answer Key	Max Points	Content Strand	Performance Indicator
1	MC	C	1	1	4.N.2
2	MC	B	1	4	4.M.2
3	MC	A	1	1	4.N.26
4	MC	A	1	1	4.N.20
5	MC	C	1	3	4.G.4
6	MC	B	1	1	4.N.2
7	MC	C	1	1	4.N.13
8	MC	C	1	1	3.N.19
9	MC	B	1	1	3.N.20
10	MC	C	1	3	4.G.1
11	MC	A	1	1	4.N.17
12	MC	A	1	5	4.S.6
13	MC	C	1	1	4.N.4
14	MC	D	1	2	3.A.1
15	MC	D	1	1	4.N.6
16	MC	A	1	4	4.M.4
17	MC	A	1	2	4.A.5
18	MC	C	1	1	4.N.27
19	MC	D	1	1	3.N.15
20	MC	B	1	4	4.M.1
21	MC	B	1	2	4.A.2
22	MC	B	1	3	3.G.2
23	MC	C	1	1	3.N.14
24	MC	B	1	4	4.M.9
25	MC	B	1	1	4.N.22
26	MC	C	1	5	4.S.5
27	MC	A	1	1	4.N.15
28	MC	B	1	1	3.N.14
29	MC	A	1	2	4.A.1
30	MC	D	1	3	4.G.2
31	SR	n/a	2	1	4.N.14
32	SR	n/a	2	1	4.N.18
33	SR	n/a	2	4	4.M.8
34	SR	n/a	2	4	4.M.3
35	ER	n/a	3	3	4.G.3
36	SR	n/a	2	1	4.N.22
37	SR	n/a	2	1	4.N.17
38	SR	n/a	2	2	4.A.4
39	ER	n/a	3	5	4.S.3
40	SR	n/a	2	1	4.N.18

*(Continued on next page)*

**Table 3b. NYSTP Mathematics 2008 Operational Test Map, Grade 4 (cont.)**

Item #	Item Type	Answer Key	Max Points	Content Strand	Performance Indicator
41	SR	n/a	2	3	4.G.3
42	SR	n/a	2	4	4.M.8
43	ER	n/a	3	2	4.A.5
44	ER	n/a	3	1	3.N.20
45	SR	n/a	2	2	4.A.4
46	SR	n/a	2	1	4.N.21
47	SR	n/a	2	5	4.S.3
48	SR	n/a	2	1	4.N.6

**Table 3c. NYSTP Mathematics 2008 Operational Test Map, Grade 5**

Item #	Item Type	Answer Key	Max Points	Content Strand	Performance Indicator
1	MC	C	1	1	5.N.3
2	MC	D	1	3	4.G.7
3	MC	C	1	1	5.N.11
4	MC	D	1	1	4.N.19
5	MC	D	1	2	4.A.2
6	MC	B	1	5	5.S.3
7	MC	C	1	1	4.N.8
8	MC	B	1	1	4.N.25
9	MC	D	1	3	5.G.11
10	MC	D	1	4	5.M.7
11	MC	C	1	3	5.G.9
12	MC	C	1	2	5.A.7
13	MC	B	1	1	5.N.22
14	MC	A	1	3	5.G.1
15	MC	A	1	1	4.N.11
16	MC	D	1	2	5.A.6
17	MC	B	1	1	5.N.18
18	MC	B	1	4	5.M.5
19	MC	D	1	5	5.S.2
20	MC	A	1	1	5.N.9
21	MC	C	1	1	5.N.24
22	MC	A	1	3	5.G.9
23	MC	D	1	1	4.N.23
24	MC	C	1	3	5.G.1
25	MC	B	1	1	5.N.20
26	MC	A	1	3	5.G.2
27	SR	n/a	2	5	5.S.3

*(Continued on next page)*

**Table 3c. NYSTP Mathematics 2008 Operational Test Map, Grade 5 (cont.)**

Item #	Item Type	Answer Key	Max Points	Content Strand	Performance Indicator
28	ER	n/a	3	2	5.A.7
29	SR	n/a	2	4	5.M.8
30	ER	n/a	3	1	5.N.11
31	SR	n/a	2	3	5.G.11
32	ER	n/a	3	5	4.S.4
33	SR	n/a	2	4	5.M.3
34	ER	n/a	3	3	5.G.1

**Table 3d. NYSTP Mathematics 2008 Operational Test Map, Grade 6**

Item #	Item Type	Answer Key	Max Points	Content Strand	Performance Indicator
1	MC	A	1	2	5.A.4
2	MC	B	1	1	6.N.21
3	MC	D	1	1	6.N.13
4	MC	C	1	3	6.G.4
5	MC	C	1	1	6.N.7
6	MC	B	1	5	5.S.6
7	MC	C	1	2	5.A.3
8	MC	C	1	5	5.S.7
9	MC	C	1	4	6.G.7
10	MC	D	1	3	6.M.5
11	MC	C	1	1	6.N.18
12	MC	D	1	3	5.G.13
13	MC	D	1	2	6.A.2
14	MC	B	1	1	6.N.19
15	MC	C	1	5	6.S.5
16	MC	D	1	1	6.N.26
17	MC	C	1	2	6.A.6
18	MC	C	1	1	6.N.23
19	MC	D	1	4	6.M.3
20	MC	A	1	5	6.S.7
21	MC	D	1	1	6.N.25
22	MC	B	1	3	6.G.5
23	MC	D	1	1	6.N.16
24	MC	D	1	1	6.N.9
25	MC	A	1	5	5.S.5
26	SR	n/a	2	4	6.M.1
27	ER	n/a	3	2	5.A.4
28	ER	n/a	3	3	5.G.13

*(Continued on next page)*

**Table 3d. NYSTP Mathematics 2008 Operational Test Map, Grade 6 (cont.)**

Item #	Item Type	Answer Key	Max Points	Content Strand	Performance Indicator
29	SR	n/a	2	5	5.S.6
30	ER	n/a	3	1	6.N.12
31	SR	n/a	2	5	6.S.8
32	SR	n/a	2	4	6.M.7
33	ER	n/a	3	5	6.S.8
34	SR	n/a	2	1	6.N.10
35	SR	n/a	2	2	6.A.2

**Table 3e. NYSTP Mathematics 2008 Operational Test Map, Grade 7**

Item #	Item Type	Answer Key	Max Points	Content Strand	Performance Indicator
1	MC	D	1	3	6.G.10
2	MC	B	1	4	7.M.2
3	MC	B	1	2	6.A.4
4	MC	C	1	3	6.G.11
5	MC	B	1	1	7.A.6
6	MC	B	1	1	7.N.2
7	MC	D	1	1	7.N.9
8	MC	D	1	5	7.S.6
9	MC	C	1	2	6.A.2
10	MC	B	1	5	7.S.10
11	MC	D	1	1	7.N.11
12	MC	B	1	3	7.G.1
13	MC	A	1	1	7.N.6
14	MC	B	1	5	7.S.4
15	MC	A	1	2	6.A.3
16	MC	C	1	5	7.S.12
17	MC	B	1	1	7.N.18
18	MC	D	1	5	7.S.10
19	MC	B	1	4	7.M.4
20	MC	C	1	1	7.N.8
21	MC	B	1	5	7.S.6
22	MC	D	1	5	7.S.8
23	MC	D	1	4	7.M.11
24	MC	C	1	5	6.S.10
25	MC	B	1	1	7.N.12
26	MC	B	1	1	7.N.7
27	MC	D	1	4	7.M.9
28	MC	A	1	5	7.S.9

*(Continued on next page)*

**Table 3e. NYSTP Mathematics 2008 Operational Test Map, Grade 7 (cont.)**

Item #	Item Type	Answer Key	Max Points	Content Strand	Performance Indicator
29	MC	D	1	4	7.M.2
30	MC	A	1	3	7.G.3
31	ER	n/a	3	1	7.N.13
32	SR	n/a	2	2	6.A.5
33	SR	n/a	2	2	7.A.1
34	ER	n/a	3	5	6.S.3
35	ER	n/a	3	3	7.G.2
36	SR	n/a	2	1	7.N.10
37	SR	n/a	2	4	7.M.8
38	ER	n/a	3	5	6.S.2

**Table 3f. NYSTP Mathematics 2008 Operational Test Map, Grade 8**

Item #	Item Type	Answer Key	Max Points	Content Strand	Performance Indicator
1	MC	D	1	3	8.G.1
2	MC	D	1	2	8.A.2
3	MC	B	1	3	8.G.5
4	MC	A	1	3	8.G.6
5	MC	D	1	2	8.A.8
6	MC	B	1	3	7.G.5
7	MC	D	1	2	8.A.9
8	MC	A	1	2	8.A.1
9	MC	A	1	3	8.G.2
10	MC	B	1	1	8.N.4
11	MC	D	1	2	8.A.8
12	MC	B	1	4	7.M.1
13	MC	D	1	3	8.G.3
14	MC	B	1	1	8.N.5
15	MC	B	1	3	8.G.4
16	MC	A	1	2	7.A.4
17	MC	A	1	3	8.G.8
18	MC	D	1	2	8.A.2
19	MC	A	1	2	8.A.6
20	MC	C	1	2	7.A.2
21	MC	C	1	4	7.M.1
22	MC	A	1	3	8.G.2
23	MC	A	1	2	8.A.10
24	MC	B	1	2	7.A.3
25	MC	C	1	3	7.G.8

*(Continued on next page)*

**Table 3f. NYSTP Mathematics 2008 Operational Test Map, Grade 8 (cont.)**

Item #	Item Type	Answer Key	Max Points	Content Strand	Performance Indicator
26	MC	A	1	3	8.G.12
27	MC	B	1	2	8.A.7
28	ER	n/a	3	2	8.A.16
29	SR	n/a	2	1	8.N.4
30	SR	n/a	2	2	8.A.6
31	ER	n/a	3	2	8.A.12
32	SR	n/a	2	3	8.G.7
33	SR	n/a	2	2	7.A.7
34	SR	n/a	2	3	8.G.6
35	SR	n/a	2	2	8.A.7
36	SR	n/a	2	3	7.G.8
37	SR	n/a	2	3	8.G.4
38	SR	n/a	2	4	7.M.7
39	SR	n/a	2	2	8.A.3
40	ER	n/a	3	1	8.N.4
41	ER	n/a	3	3	8.G.10
42	ER	n/a	3	4	7.M.5
43	ER	n/a	3	3	8.G.11
44	SR	n/a	2	2	8.A.12
45	SR	n/a	2	3	8.G.3

***2008 Item Mapping by New York State Standards and Strands*****Table 4. NYSTP Mathematics 2008 Strand Coverage**

Grade	Strand	MC Item #	SR Item #	ER Item #	Total Items
3	Number Sense and Operations	1, 4, 5, 6, 9, 11, 12, 13, 14, 15, 18, 20	26, 28	n/a	14
	Algebra	10, 19, 23, 24	n/a	30	5
	Geometry	7, 17, 21	27	n/a	4
	Measurement	2, 3, 8, 16, 22	n/a	n/a	5
	Statistics and Probability	25	29	31	3

*(Continued on next page)*

**Table 4. NYSTP Mathematics 2008 Strand Coverage (cont.)**

Grade	Strand	MC Item #s	SR Item #s	ER Item #s	Total Items
4	Number Sense and Operations	1, 3, 4, 6, 7, 8, 9, 11, 13, 15, 18, 19, 23, 25, 27, 28	31, 32, 36, 37, 40, 46, 48	44	24
	Algebra	14, 17, 21, 29	38, 45	43	7
	Geometry	5, 10, 22, 30	41	35	6
	Measurement	2, 16, 20, 24	33, 34, 42	n/a	7
	Statistics and Probability	12, 26	47	39	4
5	Number Sense and Operations	1, 3, 4, 7, 8, 13, 15, 17, 20, 21, 23, 25	n/a	30	13
	Algebra	5, 12, 16	n/a	28	4
	Geometry	2, 9, 11, 14, 22, 24, 26	31	34	9
	Measurement	10, 18	29, 33	n/a	4
	Statistics and Probability	6, 19	27	32	4
6	Number Sense and Operations	2, 3, 5, 11, 14, 16, 18, 21, 23, 24	34	30	12
	Algebra	1, 7, 13, 17	35	27	6
	Geometry	4, 9, 12, 22	n/a	28	5
	Measurement	10, 19	26, 32	n/a	4
	Statistics and Probability	6, 8, 15, 20, 25	29, 31	33	8
7	Number Sense and Operations	6, 7, 11, 13, 17, 20, 25, 26	36	31	10
	Algebra	3, 5, 9, 15	32, 33	n/a	6
	Geometry	1, 4, 12, 30	n/a	35	5
	Measurement	2, 19, 23, 27, 29	37	n/a	6
	Statistics and Probability	8, 10, 14, 16, 18, 21, 22, 24, 28	n/a	34, 38	11
8	Number Sense and Operations	10, 14	29	40	4
	Algebra	2, 5, 7, 8, 11, 16, 18, 19, 20, 23, 24, 27	30, 33, 35, 39, 44	28, 31	19
	Geometry	1, 3, 4, 6, 9, 13, 15, 17, 22, 25, 26	32, 34, 36, 37, 45	41, 43	18
	Measurement	12, 21	38	42	4

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

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

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

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

## ***Content Rationale***

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

- which performance indicators were to be assessed
- which item types were to be used for the assessable performance indicators (For example, some performance indicators lend themselves more easily to assessment by 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 selected for their grade-level expertise, were grouped by grade band (i.e., 3/4, 5/6, 7/8), and met for four days. The committees were composed of approximately 10 participants per grade band. Upon completion of the committee meetings, NYSED reviewed the committees' determinations and approved them, with minor adjustments when necessary, to maintain consistency across the grades. In January 2006, a second specifications meeting was held again with New York State educators from around the state in order to review changes made to the New York State Mathematics Learning Standards and all the items were revisited before field testing to certify alignment.

### ***Item Development***

Based on the decisions made during the item specifications meetings, the content-lead editors at CTB/McGraw-Hill distributed writing assignments to experienced item writers. The writers' assignments outlined the number and type of items (including depth-of-knowledge or thinking skill level) to write for each assignment. Writers were familiarized with the New York State Testing Program and the test specifications. They were also provided with sample test items, a style guide, and a document outlining the criteria for acceptable items (see Appendix A) to help them in their writing process.

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

### ***Item Review***

As was done for the specifications and passage review meetings, committees composed of New York State educators were selected for their content and grade-level expertise for item review. Each committee was composed of approximately 10 participants per grade band. The committee members were provided with the items, the New York State Learning Standards, and the test specifications, and 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 distracters (multiple-choice items)
- the conciseness, preciseness, clarity, and readability of the items
- the existence of any ethnic, gender, regional, or other possible bias evident in the items

Upon completion of the committee work, NYSED reviewed the decisions of the committee members; NYSED either approved the changes to the items or suggested additional revisions so that the nature and format of the items were consistent across grades and with the format

and style of the testing program. All approved changes were then incorporated into the items prior to field testing.

### ***Materials Development***

Following item review, CTB/McGraw-Hill staff assembled the approved items into field test forms and submitted the field test forms to NYSED for their review and approval. The field tests were administered to students across New York State during the week of March 19, 2007. In addition, CTB/McGraw-Hill, in conjunction with NYSED's input and approval, developed a combined *Teacher's Directions and School Administrator's Manual* so that the field tests were administered in a uniform manner to all participating students.

After administration of the field tests, rangefinding meetings were conducted in April 2007 in New York State to examine a sampling of the short- and extended-student responses to the field tests. Committees of New York State educators with content and grade-level expertise were again assembled. Each committee was composed of approximately 8 to 10 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 scoring 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 Grades 3–8 Mathematics operational Tests were administered in March 2008. The test items were selected from the pool of field-tested items, 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 B). Item selection for the NYSTP Grades 3–8 Mathematics Tests was based on the classical and IRT statistics of the test items. Selection was conducted by content experts from CTB/McGraw-Hill and NYSED and reviewed by psychometricians at CTB/McGraw-Hill and at NYSED. Two criteria governed the item selection process. The first of these was to meet the content specifications provided by NYSED. Second, within the limits set by these requirements, developers selected items with the best psychometric characteristics from the field-test item pool. 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 March 2008.

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 B).

NYSED staff (including their content and research representative experts) traveled to CTB/McGraw-Hill in Monterey in August 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 approval by NYSED, the tests were produced and administered in March 2008.

In addition to the test books, CTB/McGraw-Hill produced two *School Administrator's Manuals*, one for public schools and the other for nonpublic schools, as well as *Teacher's Directions* for each grade, 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/elintmath.html>
- <http://www.nysedregents.org/testing/mathei/08exams/home.htm>

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

Proficiency cut score recommendations and the drafting of performance standards occurred at the NYSTP mathematics standard setting in Albany, July 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 proficiency cut score setting, please refer to the *Bookmark Standard Setting Technical Report 2006 for Grades 3, 4, 5, 6, 7, and 8 Mathematics* and *NYS Measurement Review Technical Report 2006 for Mathematics*.

## Section III: Validity

---

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

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

### ***Content Validity***

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

Logical analyses of test content indicate the degree to which the content of a test covers the domain of content the test is intended to measure. In the case of 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 2–4 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 the 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 Mathematics Tests was conducted using Norman Webb's method. The

results of the study provided additional evidence of test content validity (refer to *An External Alignment Study for New York State's Assessment Program*, April 2006, Educational Testing Services).

### ***Construct (Internal Structure) Validity***

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

#### **Internal Consistency**

Empirical studies of the internal structure of the test provide one type of evidence of construct validity. For example, high internal consistency constitutes evidence of validity. This is because high coefficients imply that the test questions are measuring the same domain of skill and are reliable and consistent. Reliability coefficients of the tests for total populations and subgroups of students are presented in Section VII, “Reliability and Standard Error of Measurement.” For the total populations the reliability coefficients (Cronbach’s alpha) ranged from 0.88–0.94, and for all subgroups, the reliability coefficients are greater than 0.83. Overall, high internal consistency of the New York State Mathematics Tests provides sound evidence of construct validity.

#### **Unidimensionality**

Other evidence comes from analyses of the degree to which the test questions conform to the requirements of the statistical models used to scale and equate the tests, as well as to generate student scores. Among other things, the models require that the items fit the model well and that the questions in a test measure a single domain of skill: that they are unidimensional. The item-model fit was assessed using Q1 statistics (Yen, 1981) and the results are described in detail in Section VI, “IRT Scaling and Equating.” It was found that all items in Grades 4 and 7 Mathematics Tests displayed good item-model fit. Two items in Grade 3, two items in Grade 5, one item in Grade 6, and one item in Grade 8 were flagged for poor fit. The fact that only a few items were deemed to have unacceptable fit across grades of the mathematics tests provided solid evidence for the appropriateness of IRT models used to calibrate and scale the test data. Another evidence for the efficacy of modeling ability was provided by demonstrating that the questions on New York State Mathematics Tests were related. What relates the questions is most parsimoniously claimed to be the common ability acquired by students studying the content area. Factor analysis of the test data is one way of modeling the common ability. This analysis may show that there is a single or main factor that can account for much of the variability among responses to test questions. A large first component would provide evidence of the latent ability students have in common with respect to the particular questions asked. A large main factor found from a factor analysis of an achievement test would suggest a primary ability construct that may be related to what the questions were designed to have in common, i.e., mathematics ability.

To demonstrate the common factor (ability) underlying student responses to mathematics test items, a principal component factor analysis was conducted on a correlation matrix of individual items for each test. Factoring a correlation matrix rather than actual item response

data is preferable when dichotomous variables are in the analyzed data set. Because the New York State Mathematics Tests contain both MC and CR items, the matrix of polychoric correlations was used as input for factor analysis (polychoric correlation is an extension of tetrachoric correlations that are appropriate only for MC items). The study was conducted on the total population of New York State public and charter school students in each grade. A large first principal component was evident in each analysis, demonstrating essential unidimensionality of the trait measured by each test.

More than one factor with an eigenvalue greater than 1.0 present in each data set would suggest the presence of small additional factors. However, the ratio of the variance accounted for by the first factor to the remaining factors was sufficiently large to support the claim that these tests were essentially unidimensional. These ratios showed that the first eigenvalues were at least five times as large as the second eigenvalues for all 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 of the New York State Grades 3–8 Mathematics Tests exhibited first principle components accounting for more than 20 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 5.

**Table 5. Factor Analysis Results for Mathematics Tests (Total Population)**

Grade	Initial Eigenvalues			
	Component	Total	% of Variance	Cumulative %
3	<b>1</b>	<b>7.53</b>	<b>24.28</b>	<b>24.28</b>
	2	1.32	4.26	28.54
	3	1.09	3.51	32.04
4	<b>1</b>	<b>12.87</b>	<b>26.80</b>	<b>26.80</b>
	2	1.56	3.26	30.06
	3	1.10	2.30	32.35
	4	1.05	2.20	34.55
	5	1.02	2.13	36.68
5	<b>1</b>	<b>8.55</b>	<b>25.16</b>	<b>25.16</b>
	2	1.23	3.61	28.76
	3	1.09	3.20	31.96
	4	1.05	3.07	35.03

(Continued on next page)

**Table 5. Factor Analysis Results for Mathematics Tests (Total Population) (cont.)**

Grade	Initial Eigenvalues			
	Component	Total	% of Variance	Cumulative %
6	<b>1</b>	<b>10.25</b>	<b>29.27</b>	<b>29.27</b>
	2	1.40	4.00	33.27
	3	1.04	2.97	36.24
7	<b>1</b>	<b>8.74</b>	<b>23.00</b>	<b>23.00</b>
	2	1.34	3.52	26.52
	3	1.16	3.06	29.58
8	<b>1</b>	<b>13.32</b>	<b>29.59</b>	<b>29.59</b>
	2	1.46	3.24	32.83
	3	1.25	2.79	35.62
	4	1.12	2.50	38.11

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

As an additional evidence for construct validity, the same factor analysis procedure was employed to assess dimensionality of mathematics construct for selected subgroups of students in each grade: 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 mathematics tests for the analyzed subgroups. Factor analysis results for LEP, SWD and SUA classifications are provided in Table C1 of Appendix C.

### **Minimization of Bias**

Minimizing item bias contributes to minimization of construct-irrelevant variance and contributes to improved test validity. The developers of the NYSTP tests gave careful attention to questions of possible ethnic, gender, translation, and socioeconomic status (SES) bias. All materials were written and reviewed to conform to the 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 Mathematics Tests.

The first procedure was based on the premise that careful editorial attention to validity is an essential step in keeping bias to a minimum. Bias occurs if the test is differentially valid for a

given group of test takers. If the test entails irrelevant skills or knowledge, the possibility of DIF is increased. Thus, preserving content validity is essential.

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

In the third procedure, New York State educators reviewed all field test materials. These professionals 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”). Although several items in each grade were flagged for DIF, typically the amount of DIF present was not large and very few items were flagged by multiple methods. Items that were flagged for statistically significant DIF were carefully reviewed by multiple reviewers during the 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 Scoring Site Operations Manual (2008) located at <http://www.emsc.nysed.gov/3-8/archived.htm#scoring>.

### ***Test Administration***

NYSTP Grades 3–8 Mathematics Tests were administered at the classroom level, during March 2008. The testing window for Grades 3, 4, and 5 was March 3–7, 2008. The testing window for Grades 6, 7, and 8 was March 6–12. The makeup test administration window was March 10–14 for Grades 3–5 and from March 13–19 for Grades 6–8. The makeup test administration window allowed students who were ill or otherwise unable to test during the assigned window to take the test.

### ***Scoring Procedures of Operational Tests***

The scoring of the 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 oversight of the scoring process. At each site, designated trainers taught scoring committee members the basic criteria for scoring each question and monitored the scoring sessions in the room. The trainers were assisted by facilitators or leaders who also helped in monitoring the sessions and enforcing the accuracy of scoring. The titles for administrators, trainers, and facilitators varied per scoring model chosen. At the regional level, oversight was conducted by a site coordinator. A scoring leader trained the scoring committee members and monitored sessions, and a table facilitator assisted in monitoring sessions. At the districtwide level, a school district administrator oversaw operational scoring. A district mathematics leader trained and monitored sessions, and a school mathematics leader assisted in monitoring sessions. For schoolwide scoring, oversight was provided by the principal. Otherwise, titles for the schoolwide model were the same as those for the districtwide model. The general title “scoring committee members” included scorers at every site.

### ***Scoring Models***

For the 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 Mathematics Tests. Schools were able to score these tests regionally, districtwide, or individually. Schools were required to enter one of the following scoring model codes on student answer sheets:

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

2. Schools from two districts—The first readers for the school’s test papers included staff from two school districts, nonpublic schools, charter school districts, or a combination thereof;
3. Three or more schools within a district—The first readers for the school’s test papers included staff from all schools administering this test in a district, provided at least three schools are represented;
4. Two schools within a district—The first readers for the school’s test papers included staff from all schools administering this test in a district, provided that two schools are represented; or
5. One school only (local scoring)—The first readers for the school’s test papers included staff from the only school in the district administering this test, staff from one charter school, or staff from one nonpublic school.

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

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

### ***Scoring of Constructed-Response Items***

The scoring of 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. The CTB/McGraw-Hill mathematics 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 April 2007, CTB/McGraw-Hill staff met with groups of teachers from across the state in rangefinding sessions. Sets of actual field-test student responses 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. Handscoring and content-development specialists created scoring guides based on rangefinding decisions and conferences with NYSED. 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 Mathematics Frequently Asked Questions (FAQs) document, and a DVD, which highlighted important elements of the scoring guides. Each test book was scored by three separate scoring committee members, who scored three distinct sections of the test book. After test books were completed, the table facilitator or mathematics leader conducted a “read-behind” of approximately 12 sets of test books per hour to verify the accuracy of scoring. If a question arose that was not covered in the training materials, facilitators or trainers were to call the New York State Helpline (see the subsection “Quality Control Process”).

### ***Scorer Qualifications and Training***

The scoring of the operational test was conducted by qualified administrators and teachers. Trainers used the scoring guides 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 allows trainers to retrain their scorers, if necessary. The CAS also acknowledged those scorers who had grasped all aspects of the content area being scored and were well prepared to score test responses. After training, each scoring committee member was deemed prepared and verified as ready to score the 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, mathematics FAQs, and DVD, they called the New York State Helpline. This call center was established to aid teachers and administrators during operational scoring. The helpline staff consisted of trained CTB/McGraw-Hill handscoring personnel who answered questions by phone, fax, or e-mail. 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’ operational test 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 at the end of April 2008 for Grades 3, 4, and 5, and in the beginning of May 2008 for Grades 6, 7, and 8. These data were used for all data analysis. Phase 2 involved submitting of “straggler files” to CTB/McGraw-Hill in mid-May 2008. The straggler files contained less than 2% of the total population cases and due to late submission were excluded from research data analyses. Nonpublic school data were excluded from all data analyses.

### *Data Processing*

Data processing refers to the cleaning and screening procedures used to identify errors (such as out-of-range data) and the decisions made to exclude student cases or to suppress particular items in analyses. CTB/McGraw-Hill established a scoring program, EDITCHECKER, to do initial quality assurance on data and identify errors. This program verifies that the data fields are in-range (as defined), that students’ identifying information is present, and that the data are acceptable for delivery to CTB/McGraw-Hill research. NYSED and the data repository were provided with the results of the checking. CTB/McGraw-Hill research performed data cleaning to the delivered data and excluded some student cases in order to obtain a sample of the utmost integrity. It should be noted that the two major groups of cases excluded from the data set were out-of-grade students (students whose grade level did not match the test level) and students from nonpublic schools. Other deleted cases included students with no grade-level data and duplicate record cases. In addition, Grade 8 students who were administered an incorrect version of the Grade 8 test were excluded from Grade 8 data files (refer to “Item Rescoring and Replacing” sub-section for details). A list of the data cleaning procedures conducted by research and accompanying case counts is presented in Tables 6a–6f.

**Table 6a. NYSTP Mathematics Data Cleaning, Grade 3**

Exclusion Rule	# Deleted	# Cases Remain
Initial N		196641
Out of grade	122	196519
No grade	143	196376
Duplicate record	0	196376
Non-public and out-of-district schools	2808	193568
Missing values for ALL items on OP form	2	193566
Out-of-range CR scores	0	193566

**Table 6b. NYSTP Mathematics Data Cleaning, Grade 4**

Exclusion Rule	# Deleted	# Cases Remain
Initial N		212656
Out of grade	155	212501
No grade	148	212353
Duplicate record	0	212353
Non-public and out-of-district schools	17000	195353
Missing values for ALL items on OP form	3	195350
Out-of-range CR scores	0	195350

**Table 6c. NYSTP Mathematics Data Cleaning, Grade 5**

Exclusion Rule	# Deleted	# Cases Remain
Initial N		199521
Out of grade	161	199360
No grade	170	199190
Duplicate record	0	199190
Non-public and out-of-district schools	2936	196254
Missing values for ALL items on OP form	3	196251
Out-of-range CR scores	0	196251

**Table 6d. NYSTP Mathematics Data Cleaning, Grade 6**

Exclusion Rule	# Deleted	# Cases Remain
Initial N		207439
Out of grade	231	207208
No grade	35	207173
Duplicate record	0	207173
Non-public and out-of-district schools	8735	198438
Missing values for ALL items on OP form	2	198436
Out-of-range CR scores	0	198436

**Table 6e. NYSTP Mathematics Data Cleaning, Grade 7**

Exclusion Rule	# Deleted	# Cases Remain
Initial N		208744
Out of grade	287	208457
No grade	40	208417
Duplicate record	0	208417
Non-public and out-of-district schools	3183	205234
Missing values for ALL items on OP form	2	205232
Out-of-range CR scores	0	205232

**Table 6f. NYSTP Mathematics Data Cleaning, Grade 8**

Exclusion Rule	# Deleted	# Cases Remain
Initial N		225436
Out of grade	393	225043
No grade	44	224999
Duplicate record	1	224998
Non-public and out-of-district schools	18181	206817
Missing values for ALL items on OP form	4	206813
Out-of-range CR scores	0	206813
Incorrect test version	369	206444

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

The demographic characteristics of students in the classical analysis and calibration sample datasets are presented in the proceeding tables. 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 variables as it was found that the New York State population is fairly evenly split by gender categories.

**Table 7a. Grade 3 Sample Characteristics (N = 193566)**

Demographic Category		N-count	% of Total N-count
NRC	NYC	69755	36.04
	Big cities	8166	4.22
	Urban-suburban	15969	8.25
	Rural	10920	5.64
	Average needs	56713	29.30
	Low needs	29084	15.03
	Charter	2959	1.53
Ethnicity	Asian	14612	7.55
	Black	36727	18.97
	Hispanic	41491	21.44
	American Indian	963	0.50
	Multi-Racial	237	0.12
	White	99478	51.39
	Unknown	58	0.03
LEP	No	178268	92.10
	Yes	15298	7.90
SWD	No	167687	86.63
	Yes	25879	13.37
SUA	No	153771	79.44
	Yes	39795	20.56

**Table 7b. Grade 4 Sample Characteristics (N = 195350)**

Demographic Category		N-count	% of Total N-count
NRC	NYC	70271	35.97
	Big cities	7771	3.98
	Urban-suburban	15777	8.08
	Rural	10960	5.61
	Average needs	58266	29.83
	Low needs	29924	15.32
	Charter	2381	1.22
Ethnicity	Asian	14473	7.41
	Black	37212	19.05
	Hispanic	41588	21.29
	American Indian	939	0.48
	Multi-Racial	199	0.10
	White	100858	51.63
	Unknown	81	0.04
LEP	No	182180	93.26
	Yes	13170	6.74

*(Continued on next page)*

**Table 7b. Grade 4 Sample Characteristics (N = 195350) (cont.)**

Demographic Category		N-count	% of Total N-count
SWD	No	166789	85.38
	Yes	28561	14.62
SUA	No	153557	78.61
	Yes	41793	21.39

**Table 7c. Grade 5 Sample Characteristics (N = 196251)**

Demographic Category		N-count	% of Total N-count
NRC	NYC	69688	35.51
	Big cities	7628	3.89
	Urban-suburban	15365	7.83
	Rural	10802	5.50
	Average needs	59103	30.12
	Low needs	30380	15.48
	Charter	3285	1.67
Ethnicity	Asian	14843	7.56
	Black	37575	19.15
	Hispanic	40960	20.87
	American Indian	902	0.46
	Multi-Racial	170	0.09
	White	101732	51.84
	Unknown	69	0.04
LEP	No	185722	94.63
	Yes	10529	5.37
SWD	No	167163	85.18
	Yes	29088	14.82
SUA	No	155816	79.40
	Yes	40435	20.60

**Table 7d. Grade 6 Sample Characteristics (N = 198436)**

Demographic Category		N-count	% of Total N-count
NRC	NYC	69379	34.96
	Big cities	7626	3.84
	Urban-suburban	15039	7.58
	Rural	11332	5.71
	Average needs	60908	30.69
	Low needs	31453	15.85
	Charter	2699	1.36

*(Continued on next page)*

**Table 7d. Grade 6 Sample Characteristics (N = 198436) (cont.)**

Demographic Category		N-count	% of Total N-count
Ethnicity	Asian	14855	7.49
	Black	37048	18.67
	Hispanic	40820	20.57
	American Indian	900	0.45
	Multi-Racial	170	0.09
	White	104566	52.70
	Unknown	77	0.04
LEP	No	189906	95.70
	Yes	8530	4.30
SWD	No	168996	85.16
	Yes	29440	14.84
SUA	No	160176	80.72
	Yes	38260	19.28

**Table 7e. Grade 7 Sample Characteristics (N = 205232)**

Demographic Category		N-count	% of Total N-count
NRC	NYC	71595	34.88
	Big cities	8210	4.00
	Urban-suburban	15836	7.72
	Rural	12082	5.89
	Average needs	64217	31.29
	Low needs	31014	15.11
	Charter	2278	1.11
Ethnicity	Asian	14817	7.22
	Black	39629	19.31
	Hispanic	42171	20.55
	American Indian	1003	0.49
	Multi-Racial	138	0.07
	White	107414	52.34
	Unknown	60	0.03
LEP	No	197100	96.04
	Yes	8132	3.96
SWD	No	176059	85.79
	Yes	29173	14.21
SUA	No	166435	81.10
	Yes	38797	18.90

**Table 7f. Grade 8 Sample Characteristics (N = 206444)**

Demographic Category		N-count	% of Total N-count
NRC	NYC	71833	34.80
	Big cities	8374	4.06
	Urban-suburban	15245	7.38
	Rural	12647	6.13
	Average needs	65828	31.89
	Low needs	31108	15.07
	Charter	1409	0.68
Ethnicity	Asian	14743	7.14
	Black	39540	19.15
	Hispanic	41242	19.98
	American Indian	1025	0.50
	Multi-Racial	119	0.06
	White	109723	53.15
	Unknown	52	0.03
LEP	No	199344	96.56
	Yes	7100	3.44
SWD	No	177762	86.11
	Yes	28682	13.89
SUA	No	168449	81.60
	Yes	37995	18.40

### ***Classical Data Analysis***

Classical data analysis of the Grades 3–8 Mathematics Tests consists of four primary elements. One element is the analysis of item level statistical information about student performance. It is important to verify that the items and test forms function as intended. Information on item response patterns, item difficulty (p-value) and item-test correlation (point biserial) is examined thoroughly. If any serious error were to occur with an item (i.e., a printing error or potentially correct distractor), item analysis is the stage in which errors should be flagged and evaluated for rectification (suppression, credit, or other acceptable solution). Analyses of test level data comprise the second element of classical data analysis. These include examination of the raw score statistics (mean and standard deviation) and test reliability measures (Cronbach’s alpha and Feldt-Raju coefficient). Assessment of test speededness is another important element of classical analysis. Additionally, classical 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 Rescoring and Replacing**

One item in Grade 7 Spanish language version was rescored during the data analysis, and one item in Grade 8 book was replaced approximately 1 week before administration.

#### ***Rescored Item***

In item 33 of the Spanish edition only of the Grade 7 Mathematics Test, the phrase *less than* was translated as *less*. As a result, students who used this edition might have written the wrong algebraic expression or might have been unable to answer this question. To adjust for this, any student who used the Spanish edition, either exclusively or in conjunction with the English edition, was given credit for this item.

#### ***Replaced Item***

It was discovered that item 20 on the Grade 8 Mathematics Test was a repeated item from the previous (2007) administration. Because all NYS tests are released after operational administration each year, no items can be repeated in different operational administrations. The repeated item was replaced with a new item and Grade 8 books were reprinted in time before operational administration. However, because this error was discovered one week before the test administration, schools received both versions of the Grade 8 Test (one with the repeated item 20 and one with all unique items). Despite NYSED's best efforts to notify schools about the test version change, five schools administered the old version of the test. Based on the number of books ordered by these five schools (300, 55, 15, 9, and 9), we believe that no more (and likely fewer) than 390 students took the old version of the test. Also, approximately 20 braille students took the old version of the test. The NYSED made a policy decision to exclude students who took the old version from the calibration sample but to score them as if they took a new test version. Subsequently, schools that administered the incorrect version of the test were excluded from the calibration sample (not used in scaling and equating and scoring table development), and the regular Grade 8 scoring table was applied to raw score of students who took either version of the Grade 8 Test.

### **Item Difficulty and Response Distribution**

Item difficulty and response distribution tables (Table 8a–8f) 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 selecting 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” and “% Sel” 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 for each MC item or the average percent of the maximum score that students earned on each CR item. It is important to have a good range of p-values, to increase test information and to avoid floor or ceiling effects. Generally, p-values should range between 0.30 and 0.90. P-values represent the overall degree of difficulty, but do not account for demonstrated student performance on other test items. Usually, p-value

information is coupled with point biserial (pbis) statistics to verify that items are functioning as intended. (Point biserials are discussed in the next subsection.) Item difficulties (p-values) on the tests ranged from 0.36 to 0.98. For Grade 3, the item p-values were between 0.57 and 0.95 with a mean of 0.84. For Grade 4, the item p-values were between 0.48 and 0.98 with a mean of 0.75. For Grade 5, the item p-values were between 0.55 and 0.96 with a mean of 0.74. For Grade 6, the item p-values were between 0.36 and 0.95 with a mean of 0.72. For Grade 7, the item p-values were between 0.44 and 0.89 with a mean of 0.71. For Grade 8, the item p-values were between 0.47 and 0.94 with a mean of 0.68. These statistics are also provided in Table 9, along with other classical test summary statistics.

**Table 8a. P-values, Scored Response Distributions, and Point Biserials, 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	193566	0.94	0.02	0.03	1.37	*94.39	2.89	1.29	-0.19	*0.36	-0.20	-0.24	0.36
2	193566	0.92	0.04	0.06	1.53	2.96	*91.85	3.56	-0.17	-0.2	*0.37	-0.24	0.37
3	193566	0.88	0.07	0.07	*88.12	2.92	3.46	5.36	*0.44	-0.26	-0.23	-0.23	0.44
4	193566	0.90	0.05	0.04	4.75	*90.30	2.24	2.62	-0.18	*0.36	-0.21	-0.23	0.36
5	193566	0.88	0.06	0.04	5.61	3.88	*88.40	2.01	-0.26	-0.26	*0.45	-0.21	0.45
6	193566	0.82	0.07	0.04	8.61	*81.76	4.90	4.62	-0.34	*0.51	-0.20	-0.26	0.51
7	193566	0.91	0.04	0.04	7.03	*91.16	1.22	0.51	-0.21	*0.29	-0.14	-0.16	0.29
8	193566	0.82	0.07	0.04	4.89	6.39	*82.47	6.14	-0.26	-0.21	*0.43	-0.23	0.43
9	193566	0.93	0.05	0.02	*93.32	2.07	1.54	3.00	*0.39	-0.19	-0.24	-0.23	0.39
10	193566	0.94	0.04	0.06	*94.02	2.77	2.67	0.43	*0.31	-0.23	-0.15	-0.15	0.31
11	193566	0.78	0.05	0.04	15.99	2.73	*78.42	2.76	-0.44	-0.14	*0.52	-0.16	0.52
12	193566	0.84	0.06	0.04	2.57	3.41	*84.42	9.50	-0.22	-0.19	*0.49	-0.37	0.49
13	193566	0.88	0.08	0.07	2.12	*87.65	5.66	4.42	-0.26	*0.45	-0.22	-0.29	0.45
14	193566	0.93	0.07	0.04	2.74	1.78	1.90	*93.47	-0.25	-0.22	-0.22	*0.42	0.42
15	193566	0.88	0.08	0.05	1.96	0.63	9.35	*87.93	-0.25	-0.16	-0.32	*0.44	0.44
16	193566	0.74	0.09	0.07	*73.52	11.76	7.78	6.77	*0.38	-0.17	-0.28	-0.13	0.38
17	193566	0.89	0.09	0.06	5.16	3.50	2.64	*88.55	-0.19	-0.20	-0.16	*0.33	0.33
18	193566	0.87	0.07	0.03	8.26	*87.12	3.15	1.37	-0.21	*0.41	-0.26	-0.27	0.41
19	193566	0.87	0.09	0.03	4.94	5.51	*86.83	2.61	-0.21	-0.28	*0.43	-0.23	0.43
20	193566	0.93	0.10	0.03	0.91	*92.67	4.55	1.73	-0.19	*0.43	-0.29	-0.23	0.43
21	193566	0.95	0.11	0.04	1.79	2.29	*95.16	0.63	-0.21	-0.22	*0.35	-0.13	0.35
22	193566	0.85	0.12	0.03	9.73	*84.86	4.24	1.03	-0.31	*0.43	-0.24	-0.12	0.43

(Continued on next page)

**Table 8a. P-values, Scored Response Distributions, and Point Biseriars, Grade 3 (cont.)**

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
23	193566	0.76	0.16	0.04	3.76	7.32	*76.12	12.59	-0.23	-0.20	*0.45	-0.28	0.45
24	193566	0.78	0.24	0.03	7.29	8.12	6.74	*77.58	-0.24	-0.26	-0.17	*0.43	0.43
25	193566	0.75	0.51	0.03	3.89	5.60	14.52	*75.44	-0.19	-0.16	-0.45	*0.55	0.55
26	193566	0.72	0.09	13.20	28.49	58.22							
27	193566	0.66	0.10	10.69	45.90	43.30							
28	193566	0.57	0.14	21.40	42.12	36.34							
29	193566	0.74	0.13	16.85	17.07	65.95							
30	193566	0.76	0.09	1.28	10.71	46.47	41.46						
31	193566	0.84	0.14	6.41	7.10	14.67	71.68						

**Table 8b. P-values, Scored Response Distributions, and Point Biseriars, 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	195350	0.94	0.02	0.02	1.19	3.63	*93.53	1.60	-0.22	-0.25	*0.36	-0.15	0.36
2	195350	0.98	0.02	0.02	0.86	*97.84	0.99	0.27	-0.10	*0.20	-0.14	-0.11	0.20
3	195350	0.86	0.04	0.03	*86.40	8.40	4.16	0.97	*0.37	-0.24	-0.19	-0.21	0.37
4	195350	0.77	0.07	0.06	*76.74	8.53	5.69	8.91	*0.61	-0.23	-0.28	-0.44	0.61
5	195350	0.89	0.06	0.03	7.61	2.55	*88.75	1.01	-0.18	-0.16	*0.28	-0.15	0.28
6	195350	0.87	0.04	0.03	3.09	*87.18	6.07	3.60	-0.22	*0.35	-0.11	-0.28	0.35
7	195350	0.70	0.10	0.03	10.84	11.29	*70.46	7.28	-0.21	-0.19	*0.38	-0.17	0.38
8	195350	0.84	0.04	0.02	10.36	2.81	*84.47	2.29	-0.48	-0.17	*0.54	-0.14	0.54
9	195350	0.89	0.04	0.02	4.62	*89.34	3.14	2.85	-0.28	*0.41	-0.19	-0.21	0.41
10	195350	0.80	0.05	0.04	10.08	6.09	*79.79	3.94	-0.23	-0.24	*0.41	-0.18	0.41
11	195350	0.88	0.05	0.03	*87.70	6.03	2.71	3.47	*0.5	-0.33	-0.19	-0.29	0.50
12	195350	0.53	0.08	0.05	*52.75	12.30	29.69	5.12	*0.13	-0.16	0.03	-0.10	0.13
13	195350	0.75	0.07	0.04	10.59	3.21	*74.54	11.55	-0.13	-0.24	*0.45	-0.36	0.45
14	195350	0.91	0.06	0.03	4.91	3.49	0.84	*90.67	-0.21	-0.26	-0.16	*0.38	0.38
15	195350	0.83	0.10	0.04	3.39	4.93	8.14	*83.40	-0.24	-0.23	-0.20	*0.40	0.40
16	195350	0.68	0.10	0.04	*67.61	12.34	10.43	9.49	*0.29	-0.13	-0.19	-0.12	0.29
17	195350	0.76	0.09	0.05	*75.78	6.69	12.22	5.17	*0.56	-0.17	-0.45	-0.22	0.56
18	195350	0.71	0.14	0.04	11.32	7.30	*70.69	10.52	-0.18	-0.25	*0.45	-0.26	0.45
19	195350	0.67	0.08	0.03	27.41	2.15	3.67	*66.66	-0.40	-0.16	-0.10	*0.47	0.47
20	195350	0.58	0.07	0.03	9.30	*58.32	27.21	5.08	-0.16	*0.35	-0.14	-0.29	0.35
21	195350	0.75	0.09	0.03	11.86	*74.89	7.99	5.14	-0.22	*0.45	-0.29	-0.19	0.45
22	195350	0.79	0.08	0.04	5.54	*78.70	12.98	2.67	-0.19	*0.34	-0.20	-0.17	0.34
23	195350	0.65	0.13	0.06	9.09	6.22	*64.65	19.85	-0.24	-0.08	*0.48	-0.35	0.48
24	195350	0.63	0.12	0.05	2.84	*63.30	9.12	24.56	-0.19	*0.46	-0.24	-0.27	0.46
25	195350	0.63	0.17	0.06	4.80	*63.42	16.03	15.51	-0.17	*0.55	-0.30	-0.32	0.55
26	195350	0.48	0.20	0.08	12.34	2.18	*47.89	37.31	-0.25	-0.14	*0.46	-0.26	0.46
27	195350	0.76	0.25	0.05	*75.83	6.86	12.47	4.55	*0.42	-0.24	-0.22	-0.22	0.42

(Continued on next page)

**Table 8b. P-values, Scored Response Distributions, and Point Biserials, Grade 4 (cont.)**

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
28	195350	0.48	0.23	0.08	10.80	*47.69	11.20	29.99	-0.29	*0.53	-0.12	-0.29	0.53
29	195350	0.62	0.31	0.05	*61.93	8.85	13.53	15.32	*0.44	-0.26	-0.12	-0.27	0.44
30	195350	0.75	0.64	0.05	2.70	5.08	16.83	*74.71	-0.21	-0.19	-0.27	*0.41	0.41
31	195350	0.83	0.06	9.36	14.25	76.33							
32	195350	0.76	0.12	13.07	21.63	65.18							
33	195350	0.71	0.10	8.08	42.01	49.80							
34	195350	0.59	0.29	34.24	11.97	53.50							
35	195350	0.68	0.17	10.76	26.39	11.62	51.06						
36	195350	0.61	0.26	31.61	14.11	54.02							
37	195350	0.84	0.26	10.32	9.98	79.44							
38	195350	0.62	0.27	20.06	34.42	45.26							
39	195350	0.74	0.18	3.76	18.28	30.28	47.51						
40	195350	0.80	0.06	13.10	13.96	72.88							
41	195350	0.91	0.10	3.28	11.24	85.38							
42	195350	0.72	0.12	15.13	26.20	58.54							
43	195350	0.84	0.11	6.73	4.23	17.70	71.23						
44	195350	0.79	0.10	7.81	6.04	28.03	58.02						
45	195350	0.65	0.15	27.97	13.73	58.14							
46	195350	0.70	0.18	25.14	8.38	66.30							
47	195350	0.88	0.20	4.36	13.96	81.49							
48	195350	0.84	0.32	5.03	21.25	73.39							

**Table 8c. P-values, Scored Response Distributions, and Point Biserials, 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	196251	0.79	0.02	0.02	4.83	2.63	*79.38	13.12	-0.23	-0.17	*0.44	-0.30	0.44
2	196251	0.75	0.04	0.02	5.16	11.11	8.39	*75.29	-0.24	-0.26	-0.27	*0.49	0.49
3	196251	0.96	0.02	0.01	1.60	0.47	*96.47	1.43	-0.09	-0.12	*0.22	-0.19	0.22
4	196251	0.71	0.03	0.04	12.64	10.69	5.45	*71.15	-0.45	-0.19	-0.04	*0.48	0.48
5	196251	0.77	0.06	0.04	3.68	4.10	15.44	*76.68	-0.18	-0.18	-0.15	*0.30	0.30
6	196251	0.67	0.09	0.03	17.88	*66.72	11.78	3.51	-0.32	*0.55	-0.29	-0.22	0.55
7	196251	0.55	0.08	0.03	32.66	7.82	*54.55	4.86	-0.40	-0.23	*0.54	-0.09	0.54
8	196251	0.77	0.04	0.05	4.56	*76.93	10.15	8.28	-0.11	*0.45	-0.39	-0.17	0.45
9	196251	0.75	0.03	0.02	9.91	4.47	10.29	*75.29	-0.39	-0.12	-0.19	*0.46	0.46
10	196251	0.75	0.05	0.03	6.14	10.43	8.56	*74.79	-0.17	-0.18	-0.10	*0.29	0.29
11	196251	0.87	0.04	0.02	1.92	6.08	*87.03	4.91	-0.16	-0.23	*0.33	-0.15	0.33
12	196251	0.87	0.04	0.02	1.39	4.74	*86.90	6.91	-0.17	-0.27	*0.47	-0.31	0.47
13	196251	0.67	0.07	0.02	13.32	*67.03	15.88	3.68	-0.23	*0.61	-0.45	-0.24	0.61
14	196251	0.95	0.04	0.01	*95.50	2.43	0.91	1.12	*0.25	-0.15	-0.12	-0.15	0.25
15	196251	0.92	0.03	0.04	*91.76	5.71	1.04	1.42	*0.24	-0.12	-0.14	-0.20	0.24

(Continued on next page)

**Table 8c. P-values, Scored Response Distributions, and Point Biserials, Grade 5 (cont.)**

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
16	196251	0.61	0.06	0.05	3.36	30.84	5.08	*60.62	-0.20	-0.40	-0.15	*0.52	0.52
17	196251	0.61	0.08	0.03	28.62	*60.83	3.87	6.57	-0.18	*0.41	-0.25	-0.28	0.41
18	196251	0.63	0.17	0.03	10.81	*63.08	17.70	8.21	-0.22	*0.51	-0.31	-0.21	0.51
19	196251	0.80	0.05	0.03	6.95	8.58	3.96	*80.43	-0.12	-0.22	-0.22	*0.34	0.34
20	196251	0.68	0.11	0.03	*67.64	9.42	14.05	8.74	*0.42	-0.19	-0.24	-0.21	0.42
21	196251	0.77	0.09	0.05	6.30	13.40	*76.97	3.19	-0.25	-0.29	*0.47	-0.19	0.47
22	196251	0.95	0.08	0.07	*95.36	1.65	1.29	1.55	*0.20	-0.10	-0.11	-0.13	0.20
23	196251	0.79	0.13	0.06	7.86	8.70	3.98	*79.27	-0.28	-0.35	-0.18	*0.52	0.52
24	196251	0.63	0.19	0.04	10.47	17.17	*63.23	8.90	-0.20	-0.25	*0.49	-0.28	0.49
25	196251	0.68	0.27	0.03	16.92	*67.88	5.58	9.33	-0.31	*0.51	-0.16	-0.29	0.51
26	196251	0.77	0.39	0.02	*77.31	12.68	5.26	4.35	*0.29	-0.12	-0.21	-0.15	0.29
27	196251	0.76	0.08	12.23	22.72	64.97							
28	196251	0.62	0.15	16.13	16.64	30.46	36.63						
29	196251	0.73	0.08	13.85	26.75	59.32							
30	196251	0.55	0.19	15.39	32.90	22.32	29.20						
31	196251	0.62	0.20	19.54	37.28	42.97							
32	196251	0.63	0.09	2.46	42.13	17.66	37.66						
33	196251	0.90	0.09	1.70	16.59	81.62							
34	196251	0.76	0.11	6.60	19.42	13.14	60.73						

**Table 8d. 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	Pbis Option 1	Pbis Option 2	Pbis Option 3	Pbis Option 4	Pbis Key
1	198436	0.86	0.07	0.01	*85.76	9.62	2.28	2.26	*0.50	-0.40	-0.26	-0.12	0.50
2	198436	0.68	0.06	0.03	20.98	*67.55	6.43	4.95	-0.52	*0.62	-0.24	-0.10	0.62
3	198436	0.69	0.06	0.04	6.38	3.29	21.16	*69.08	-0.23	-0.18	-0.16	*0.33	0.33
4	198436	0.74	0.10	0.03	3.19	12.64	*74.47	9.57	-0.18	-0.21	*0.38	-0.21	0.38
5	198436	0.86	0.12	0.03	5.99	4.24	*85.54	4.08	-0.27	-0.24	*0.44	-0.20	0.44
6	198436	0.78	0.05	0.03	8.44	*78.12	3.37	9.98	-0.29	*0.43	-0.26	-0.16	0.43
7	198436	0.66	0.09	0.03	4.98	3.56	*66.35	25.00	-0.30	-0.24	*0.38	-0.16	0.38
8	198436	0.82	0.06	0.03	6.58	6.14	*82.07	5.14	-0.24	-0.31	*0.48	-0.23	0.48
9	198436	0.74	0.10	0.03	5.87	9.71	*74.35	9.95	-0.24	-0.26	*0.34	-0.06	0.34
10	198436	0.85	0.08	0.03	1.38	9.67	3.86	*84.99	-0.15	-0.20	-0.20	*0.32	0.32
11	198436	0.71	0.08	0.04	6.80	6.09	*71.03	15.96	-0.25	-0.29	*0.58	-0.35	0.58
12	198436	0.86	0.06	0.03	10.15	1.64	2.02	*86.10	-0.34	-0.17	-0.19	*0.44	0.44
13	198436	0.78	0.10	0.04	3.51	4.98	13.38	*77.99	-0.26	-0.27	-0.38	*0.57	0.57
14	198436	0.56	0.13	0.04	13.20	*55.88	26.06	4.69	-0.10	*0.27	-0.13	-0.18	0.27
15	198436	0.69	0.11	0.04	13.74	7.12	*69.18	9.82	-0.26	-0.18	*0.51	-0.33	0.51
16	198436	0.49	0.20	0.04	5.60	21.44	23.69	*49.03	-0.26	-0.30	-0.08	*0.44	0.44
17	198436	0.88	0.17	0.03	1.70	8.43	*87.77	1.90	-0.15	-0.34	*0.43	-0.17	0.43

(Continued on next page)

**Table 8d. P-values, Scored Response Distributions, and Point Biserials, Grade 6 (cont.)**

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
18	198436	0.95	0.08	0.03	0.89	2.66	*94.50	1.83	-0.11	-0.22	*0.32	-0.19	0.32
19	198436	0.64	0.14	0.03	6.49	17.76	11.79	*63.79	-0.28	-0.29	-0.23	*0.53	0.53
20	198436	0.80	0.17	0.04	*80.01	3.77	10.49	5.50	*0.41	-0.14	-0.21	-0.32	0.41
21	198436	0.84	0.18	0.07	2.57	10.96	2.40	*83.82	-0.28	-0.42	-0.22	*0.57	0.57
22	198436	0.76	0.19	0.06	8.79	*75.67	7.63	7.67	-0.19	*0.45	-0.26	-0.26	0.45
23	198436	0.59	0.23	0.03	2.89	27.00	10.44	*59.41	-0.18	-0.43	-0.19	*0.58	0.58
24	198436	0.79	0.29	0.03	9.97	3.14	8.04	*78.52	-0.32	-0.22	-0.31	*0.54	0.54
25	198436	0.84	0.35	0.03	*83.52	5.39	6.54	4.17	*0.41	-0.21	-0.25	-0.20	0.41
26	198436	0.84	0.11	12.59	5.96	81.34							
27	198436	0.78	0.20	10.32	5.67	22.01	61.79						
28	198436	0.59	0.15	8.00	21.68	54.27	15.91						
29	198436	0.36	0.66	53.14	19.87	26.33							
30	198436	0.46	0.26	38.34	15.01	15.04	31.34						
31	198436	0.73	0.26	17.47	18.93	63.34							
32	198436	0.44	0.50	42.88	26.18	30.45							
33	198436	0.66	0.26	13.25	20.06	22.75	43.68						
34	198436	0.68	0.90	22.21	18.38	58.50							
35	198436	0.75	0.81	14.73	18.63	65.83							

**Table 8e. P-values, Scored Response Distributions, and Point Biserials, 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	205232	0.83	0.02	0.04	0.81	1.70	14.71	*82.73	-0.16	-0.13	-0.32	*0.38	0.38
2	205232	0.79	0.07	0.03	10.30	*79.49	6.12	3.99	-0.15	*0.36	-0.25	-0.20	0.36
3	205232	0.83	0.13	0.04	8.68	*82.89	4.52	3.75	-0.17	*0.40	-0.26	-0.26	0.40
4	205232	0.88	0.06	0.03	2.26	4.62	*87.89	5.14	-0.19	-0.16	*0.22	-0.04	0.22
5	205232	0.84	0.27	0.03	5.54	*84.02	4.89	5.26	-0.17	*0.41	-0.24	-0.26	0.41
6	205232	0.44	0.14	0.05	16.49	*43.75	8.58	30.99	-0.23	*0.41	-0.20	-0.14	0.41
7	205232	0.67	0.14	0.04	9.82	14.31	8.95	*66.73	-0.34	-0.14	-0.15	*0.42	0.42
8	205232	0.84	0.05	0.05	5.54	1.72	8.58	*84.06	-0.21	-0.16	-0.20	*0.34	0.34
9	205232	0.80	0.07	0.03	6.15	11.33	*80.47	1.95	-0.36	-0.24	*0.47	-0.16	0.47
10	205232	0.82	0.05	0.03	6.21	*81.53	3.73	8.45	-0.23	*0.43	-0.13	-0.30	0.43
11	205232	0.59	0.19	0.05	11.31	7.48	21.87	*59.10	-0.04	-0.21	-0.35	*0.44	0.44
12	205232	0.68	0.17	0.04	16.12	*67.72	5.54	10.41	-0.05	*0.33	-0.19	-0.30	0.33
13	205232	0.75	0.07	0.03	*74.61	15.58	3.89	5.81	*0.39	-0.19	-0.26	-0.20	0.39
14	205232	0.74	0.14	0.03	7.72	*74.14	7.69	10.28	-0.11	*0.50	-0.26	-0.38	0.50
15	205232	0.78	0.09	0.03	*77.90	9.19	5.48	7.30	*0.40	-0.19	-0.23	-0.21	0.40
16	205232	0.82	0.14	0.03	9.93	5.00	*81.90	3.00	-0.28	-0.17	*0.38	-0.14	0.38
17	205232	0.75	0.12	0.04	6.51	*75.11	6.77	11.45	-0.18	*0.50	-0.27	-0.32	0.50
18	205232	0.86	0.08	0.05	3.52	4.64	5.68	*86.02	-0.22	-0.27	-0.25	*0.46	0.46

(Continued on next page)

**Table 8e. P-values, Scored Response Distributions, and Point Biserials, Grade 7 (cont.)**

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
19	205232	0.64	0.17	0.04	23.35	*64.04	6.02	6.39	-0.15	*0.37	-0.28	-0.18	0.37
20	205232	0.55	0.21	0.05	19.82	8.83	*55.43	15.66	0.06	-0.15	*0.24	-0.27	0.24
21	205232	0.53	0.10	0.05	16.47	*53.20	20.98	9.20	-0.13	*0.41	-0.24	-0.19	0.41
22	205232	0.84	0.10	0.05	1.75	10.90	2.85	*84.36	-0.18	-0.35	-0.17	*0.45	0.45
23	205232	0.64	0.24	0.05	5.03	10.84	19.86	*63.98	-0.22	-0.19	-0.28	*0.46	0.46
24	205232	0.51	0.25	0.04	26.89	8.90	*50.53	13.38	-0.22	-0.07	*0.27	-0.05	0.27
25	205232	0.52	0.34	0.06	21.05	*52.31	9.04	17.21	-0.15	*0.36	-0.18	-0.16	0.36
26	205232	0.62	0.22	0.05	27.26	*61.87	6.61	3.99	-0.32	*0.46	-0.21	-0.12	0.46
27	205232	0.81	0.20	0.06	5.61	1.47	11.97	*80.69	-0.22	-0.16	-0.10	*0.26	0.26
28	205232	0.78	0.26	0.04	*77.63	6.43	5.03	10.62	*0.32	-0.20	-0.23	-0.10	0.32
29	205232	0.89	0.34	0.03	2.48	3.12	5.03	*88.99	-0.16	-0.22	-0.26	*0.39	0.39
30	205232	0.83	0.35	0.03	*82.86	2.85	12.09	1.82	*0.38	-0.13	-0.29	-0.16	0.38
31	205232	0.66	0.11	12.34	18.10	27.46	41.99						
32	205232	0.68	0.53	25.42	12.02	62.03							
33	205232	0.61	0.77	30.42	15.83	52.98							
34	205232	0.57	0.18	9.05	42.89	16.22	31.66						
35	205232	0.58	0.34	19.32	19.98	28.14	32.22						
36	205232	0.64	2.74	25.68	15.60	55.99							
37	205232	0.66	0.56	20.10	26.40	52.95							
38	205232	0.73	0.71	10.65	16.69	14.77	57.17						

**Table 8f. 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	Pbis Option 1	Pbis Option 2	Pbis Option 3	Pbis Option 4	Pbis Key
1	206444	0.94	0.04	0.01	2.42	1.80	1.95	*93.76	-0.21	-0.20	-0.17	*0.35	0.35
2	206444	0.71	0.03	0.03	13.13	13.92	2.05	*70.85	-0.26	-0.23	-0.17	*0.43	0.43
3	206444	0.77	0.03	0.02	4.30	*76.98	17.98	0.69	-0.22	*0.32	-0.20	-0.13	0.32
4	206444	0.87	0.05	0.03	*87.45	10.28	1.65	0.54	*0.35	-0.26	-0.21	-0.11	0.35
5	206444	0.67	0.03	0.03	8.90	13.32	10.82	*66.90	-0.36	-0.35	-0.14	*0.57	0.57
6	206444	0.86	0.06	0.02	4.81	*86.12	6.41	2.57	-0.21	*0.41	-0.27	-0.19	0.41
7	206444	0.70	0.09	0.03	4.05	4.53	20.96	*70.33	-0.23	-0.19	-0.36	*0.51	0.51
8	206444	0.61	0.05	0.03	*60.96	26.42	7.96	4.58	*0.34	-0.13	-0.24	-0.20	0.34
9	206444	0.64	0.05	0.03	*64.20	6.45	21.20	8.08	*0.51	-0.21	-0.34	-0.20	0.51
10	206444	0.58	0.05	0.02	35.16	*58.33	3.80	2.63	-0.48	*0.53	-0.09	-0.09	0.53
11	206444	0.61	0.09	0.03	19.39	7.63	11.46	*61.40	-0.34	-0.26	-0.21	*0.56	0.56
12	206444	0.90	0.03	0.03	2.19	*89.72	3.72	4.30	-0.16	*0.29	-0.12	-0.20	0.29
13	206444	0.83	0.04	0.03	3.39	7.35	5.86	*83.33	-0.29	-0.27	-0.33	*0.54	0.54
14	206444	0.70	0.08	0.03	8.86	*69.82	12.00	9.20	-0.21	*0.50	-0.23	-0.33	0.50
15	206444	0.86	0.05	0.02	5.45	*85.76	5.80	2.92	-0.24	*0.42	-0.23	-0.22	0.42
16	206444	0.50	0.16	0.03	*49.53	18.26	16.30	15.72	*0.46	-0.20	-0.21	-0.20	0.46

(Continued on next page)

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

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
17	206444	0.71	0.05	0.04	*70.70	7.32	9.88	12.01	*0.44	-0.17	-0.26	-0.24	0.44
18	206444	0.85	0.04	0.03	11.29	1.52	2.31	*84.81	-0.08	-0.17	-0.23	*0.23	0.23
19	206444	0.77	0.08	0.03	*76.67	12.50	6.24	4.49	*0.54	-0.39	-0.27	-0.15	0.54
20	206444	0.53	0.06	0.03	25.30	7.84	*52.78	13.99	0.11	-0.20	*0.26	-0.35	0.26
21	206444	0.84	0.05	0.04	3.60	4.97	*83.88	7.47	-0.19	-0.27	*0.43	-0.24	0.43
22	206444	0.59	0.06	0.04	*59.44	12.63	14.77	13.06	*0.58	-0.24	-0.28	-0.31	0.58
23	206444	0.59	0.16	0.04	*59.34	25.84	8.33	6.29	*0.54	-0.29	-0.30	-0.22	0.54
24	206444	0.67	0.14	0.04	4.93	*66.57	20.73	7.59	-0.22	*0.34	-0.14	-0.20	0.34
25	206444	0.71	0.11	0.04	3.23	13.04	*71.00	12.58	-0.24	-0.26	*0.50	-0.29	0.50
26	206444	0.78	0.11	0.03	*77.75	6.14	5.83	10.15	*0.54	-0.30	-0.31	-0.26	0.54
27	206444	0.79	0.16	0.03	5.27	*79.50	8.39	6.66	-0.27	*0.55	-0.35	-0.25	0.55
28	206444	0.68	0.39	11.50	12.68	34.93	40.50						
29	206444	0.66	0.52	22.43	22.10	54.95							
30	206444	0.61	1.26	24.44	25.72	48.59							
31	206444	0.56	1.18	28.61	16.34	9.11	44.76						
32	206444	0.66	0.59	13.01	40.52	45.87							
33	206444	0.67	0.55	11.45	41.23	46.77							
34	206444	0.80	0.68	17.10	4.38	77.84							
35	206444	0.47	0.94	41.84	20.99	36.23							
36	206444	0.58	0.47	26.07	31.85	41.61							
37	206444	0.51	0.73	23.12	50.96	25.19							
38	206444	0.58	0.78	17.51	46.95	34.77							
39	206444	0.54	0.59	35.33	19.68	44.40							
40	206444	0.64	1.03	20.10	12.64	17.99	48.25						
41	206444	0.76	0.83	13.16	8.10	13.35	64.55						
42	206444	0.69	0.90	10.45	10.25	39.12	39.28						
43	206444	0.58	0.60	27.69	12.79	16.66	42.26						
44	206444	0.51	1.42	38.00	18.82	41.76							
45	206444	0.72	1.04	23.91	5.14	69.92							

**Point-Biserial Correlation Coefficients**

Point biserial statistics are used to examine item-test correlations or item discrimination. In the Tables 8a–8f, 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. Point biserials for the correct answer option should be equal to or greater than 0.15, which would indicate that students who responded correctly also tended to do well on the overall test. For incorrect answer options (distractors), the point biserial should be negative, which indicates that students who scored lower on the overall test had a tendency to pick a distractor. Grade 4 item 12 was the only item flagged for having a point biserial for the correct answer below 0.15. Point biserials for correct answer options (pbis\*) on the tests ranged from 0.13–0.62. For Grade 3, the pbis\*

were between 0.29 and 0.55. For Grade 4, the pbis\* were between 0.13 and 0.61. For Grade 5, the pbis\* were between 0.20 and 0.61. For Grade 6, pbis\* were between 0.27 and 0.62. For Grade 7, the pbis\* were between 0.22 and 0.50. For Grade 8, the pbis\* were between 0.23 and 0.58.

### **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 choice). The proportions of students selecting incorrect responses while responding to MC items are provided in Tables 8a–8f. Distribution of student responses across answer choices was evaluated. It is expected that the proportion of students selecting the correct answer will be higher than proportions of students selecting any other answer choice. This was true for all New York State mathematics items.

As mentioned in the “Point Biserial Correlations Coefficients” subsection, items were flagged if the point biserial of any distractor was positive. One Grade 4 item was flagged for positive point biserial values on a distractor (incorrect) answer option (item 12, 0.03). One Grade 7 item was flagged for positive point biserial values on distractor (incorrect) answer options (item 20, 0.06).

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

Test statistics including raw-score mean and standard deviation are presented in Table 9. Reliability coefficients provide measures of internal consistency that range from zero to one. Two reliability coefficients: Cronbach’s alpha and Feldt-Raju were computed for the Grades 3–8 Mathematics Tests. Both types of reliability estimates are appropriate to use when a test contains both MC and CR items. Calculated Cronbach’s alpha reliabilities ranged 0.89–0.94. Feldt-Raju reliability coefficients ranged from 0.90–0.95. The lowest reliability was observed for the Grade 3 test, but as that test has the lowest number of score points it is reasonable that its reliability would not be as high as the other grades’ tests. The highest reliability was observed for the Grade 8 test. All reliabilities exceeded 0.85, across statistics, which is a good indication that the NYSTP 3–8 Mathematics Tests are acceptably reliable. High reliability indicates that scores are consistent and not unduly influenced by random error. (For more information on test reliability and standard error of measurement, see Section VIII, “Reliability and Standard Error of Measurement.”)

**Table 9. NYSTP Mathematics 2008 Test Form Statistics and Reliability**

Grade	Max RS	RS Mean	RS SD	P-value Mean	Minimum P-value	Maximum P-value	Cronbach's Alpha	Feldt-Raju Alpha
3	39	31.86	6.62	0.82	0.57	0.95	0.89	0.90
4	70	52.46	14.05	0.75	0.48	0.98	0.94	0.94
5	46	33.40	9.10	0.73	0.55	0.96	0.90	0.91
6	49	33.93	10.86	0.69	0.34	0.95	0.92	0.93
7	50	34.64	10.37	0.69	0.44	0.89	0.90	0.92
8	69	45.96	16.22	0.67	0.47	0.94	0.94	0.95

### Speededness

Speededness is the term used to refer to interference in test score observation due to insufficient testing time. Test developers considered speededness in the development of the NYSTP tests. NYSED believes that achievement tests should not be speeded; little or no useful instructional information can be obtained from the fact that a student did not finish a test, while a great deal can be learned from student responses to questions. Further, NYSED prefers all scores to be based on actual student performance, because all students should have ample opportunity to demonstrate that performance to enhance the validity of their scores. Test reliability is directly impacted by the number of test questions, so excluding questions that were impacted by a lack of timing would negatively impact reliability. For these reasons, sufficient administration time limits were set for the NYSTP tests. The research department at 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 8a–8f show the omit rates for items on the Grades 3–8 Mathematics Tests. These results provide no evidence of speededness on these tests.

### Differential Item Functioning

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

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), ethnicity (focal groups: Black, Hispanic, and Asian; reference group: White) and test language (focal group: Spanish; reference group: English). All cases in clean data sets were used to compute DIF statistics. Table 10 shows the number of students in each focal and reference group.

**Table 10. NYSTP Mathematics 2008 Classical DIF Sample N-Counts**

Grade	Ethnicity				Gender		Needs Resource Category		Test Language	
	Black	Hispanic	Asian	White	Female	Male	High	Low	Spanish	English
3	36727	41491	14612	100736	94215	99351	104810	85797	3216	189927
4	37212	41588	14473	102077	95901	99449	104779	88190	2803	193509
5	37575	40960	14843	102873	96097	100154	103483	89483	2543	193190
6	37048	40820	14855	105713	96869	101567	103376	92361	2681	195019
7	39629	42171	14817	108615	100791	104441	107723	95231	2892	201619
8	39540	41242	14743	110919	101105	105339	108099	96936	3106	202555

Table 11 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 11. Number of Items Flagged by SMD and Mantel-Haenszel DIF Methods**

Grade	Number of Flagged Items
3	4
4	11
5	5
6	7
7	11
8	15

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

## Section VI: IRT Scaling and Equating

---

### *IRT Models and Rationale for Use*

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

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

IRT models typically vary according to the number of parameters estimated. For the 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 sample data were used for calibration and scaling of New York State Mathematics Tests. It should be noted that the scaling was done on approximately 98% of the New York State school student population. Exclusion of some cases during the data cleaning process had a very small effect on parameter estimation. The 2008 samples were comparable to 2007 populations in terms of needs resource category, student ethnicity, proportions of students with limited English proficiency, proportions of students with disabilities, and proportions of students using testing accommodations.

### ***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 & Aitkin, 1981; Thissen, 1982). Simulation studies have compared PARDUX with MULTILOG (Thissen, 1991), PARSCALE (Muraki & Bock, 1991), and BIGSTEPS (Wright & Linacre, 1992). PARSCALE, MULTILOG, and BIGSTEPS are among the most widely known and used IRT programs. PARDUX was found to perform at least as well as these other programs (Fitzpatrick, 1990; Fitzpatrick, 1994; Fitzpatrick and Julian, 1996).

The NYSTP Mathematics tests item calibrations did not incur any problems. The number of estimation cycles was set to 50 with convergence criterion of 0.001 for all grades. The maximum value of  $a$ -parameter was set to 3.4, and range for  $b$ -parameter was set to be between -7.5 and 7.5. The maximum  $c$ -parameter value was set to 0.50. These are default parameters that have always been used for calibration of NYS test data. The estimated  $a$ - and  $b$ -parameters were in the original theta metric and all of the items were well within the prescribed parameter ranges. The  $c$ -parameter was not estimated in the 2008 calibration process but was fixed and remained unchanged for anchor items between field test and operational administration. It should be noted that there were a number of items with the default value for the  $c$ -parameter on the operational test. When the PARDUX program encounters difficulty estimating the  $c$ -parameter, it assigns a default  $c$ -parameter value of 0.2000. These default values of  $c$ -parameter were obtained during the field test calibration and remained unchanged between field test and operational administrations. Table 4 presents a summary of calibration results. For the Grades 3–8 Mathematics Tests, all of the calibration estimation results are reasonable.

**Table 12. NYSTP Mathematics 2008 Calibration Results**

Grade	Largest $a$ -parameter	$b$ -parameter Range		# Items with Default $c$ -parameter	Theta Mean	Theta Standard Deviation	# Students
3	2.202	-3.754	-0.911	14	0.17	1.409	193566
4	2.616	-4.348	2.145	14	0.05	1.173	195350
5	2.787	-3.760	0.202	10	0.07	1.224	196251
6	3.318	-3.530	2.390	8	0.04	1.190	198436
7	2.550	-2.614	1.656	14	0.04	1.200	205232
8	2.704	-3.570	0.848	8	0.03	1.173	206444

### ***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

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

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,  $Q_j$  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,  $Q_j$  was transformed to  $Z_{Q_j}$

where

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

The value of  $Z$  still will increase with sample size, all else being equal. To use this standardized statistic to flag items for potential misfit, 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  $Z_{Q_j}Crit$  used to flag items was calculated using the expression

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

where

$N$  is the calibration sample size.

Items were considered to have poor fit if the value of the obtained  $Z_{Q_j}$  was greater than the value of  $Z_{Q_j}$  critical. If the obtained  $Z_{Q_j}$  was less than  $Z_{Q_j}$  critical the items were rated as having acceptable fit. It should be noted that most items in the NYSTP 2008 Mathematics tests demonstrated a good model fit, further supporting use of the chosen models. No items in Grades 4 and 7 exhibited poor item-model fit statistics. The following items exhibited misfit: Grade 3 items 28 ( $Z_{Q_j}$ = 1019.66,  $Z_{Q_j}$  critical= 484.07) and 30 ( $Z_{Q_j}$ = 511.32,  $Z_{Q_j}$  critical= 484.07), Grade 5 items 27 ( $Z_{Q_j}$ = 1015.56,  $Z_{Q_j}$  critical= 511.06) and 31 ( $Z_{Q_j}$ = 522.73,  $Z_{Q_j}$  critical= 511.06), Grade 6 item 27 ( $Z_{Q_j}$ = 634.68,  $Z_{Q_j}$  critical= 521.25) and Grade 8 item 43 ( $Z_{Q_j}$ = 662.79,  $Z_{Q_j}$  critical= 544.89). Fit statistics and status for all items in the Grades 3–8 Mathematics Tests are presented in Appendix E.

### ***Local Independence***

In using IRT models, one of the assumptions made is that the items are locally independent; that is, student response on one item are not dependent upon their response to another item. Statistically speaking, when a student's ability is accounted for, their responses to each item are statistically independent.

One way to assess the validity of this assumption, and to measure the statistical independence of items within a test is via the  $Q_3$  statistic (Yen, 1984). This statistic was obtained by correlating differences between students' observed and expected responses for pairs of items after taking into account their overall test performance. The  $Q_3$  for binary items was computed as follows:

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

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

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

If a substantial number of items in the test demonstrate local dependence, these items may need to be calibrated separately. All pairs of items with  $Q_3$  values greater than 0.20 were classified as locally dependent. The maximum value for this index is 1.00. The content of the flagged items was examined in order to identify possible sources of the local dependence.

The  $Q_3$  statistics were examined on all of the Grade 3–8 Mathematics Tests and no items were found to be locally dependent in any grade. Anchor items were excluded from  $Q_3$  computation.

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

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: Grades 3 through 8 Mathematics Field Test Technical Report*, Page 44.

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 NYS OP scale).

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

In the second equating procedure, the 2008 Mathematics OP data were calibrated after the 2008 OP administration. Equated to OP scale FT parameters for all MC items in OP tests were used as anchors to transform the 2008 OP item parameters to NYS 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 method (Stocking & Lord, 1983). 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.

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 11 presents the 2008 OP transformation parameters for New York State Grades 3–8 Mathematics.

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:

$$M1 = A * MI_{Ft}$$

$$M2 = A * M2_{Ft} + B$$

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  $MI_{Ft}$  and  $M2_{Ft}$  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:

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

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

where

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

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

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

$Mean_{Ft}$  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 13 presents the 2008 OP transformation parameters for New York State Grades 3–8 Mathematics.

**Table 13. NYSTP Mathematics 2008 Final Transformation Constants**

Grade	$M1$	$M2$
3	27.0861	685.7917
4	32.6764	681.9752
5	30.3176	678.6393
6	31.9774	674.6304
7	30.7436	673.7211
8	32.7613	666.1757

### Anchor Set 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. Outlined below, procedures 1 and 2 refer to evaluation of the overall anchor set, and procedures 3, 4 and 5 were applied to evaluate individual anchor items.

1. Anchor set input and estimate TCC alignment. The overall alignment of TCCs for anchor set input and estimate was evaluated to determine the overall stability of anchor item parameters between FT and the 2008 OP administration.
2. Correlations of anchor input and estimate of  $a$ - and  $b$ -parameters and p-values. Correlations of anchor input and estimate of  $a$ - and  $b$ -parameters and p-values was evaluated for magnitude. Ideally, the correlations between anchor input and estimate for  $a$ -parameter should be at least 0.80 and for  $b$ -parameter and p-value should be at least 0.90.

3. Iterative linking using Stocking and Lord's TCC method. This procedure, also called 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. The differential item performance was evaluated by examining previous (input/field test) and transformed (estimated /operational) item parameters. The items with an absolute difference of parameters greater than two times the root mean square deviation are 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 will be calculated. The p-values will then be 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 perpendicular distance to the line-of-best-fit. The fitted line is chosen so as to minimize the sum of squared perpendicular distances of the points to the line. Items lying more than two standard deviations of the distance away 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 then 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 then the item is flagged for performance differences between the two item administrations.

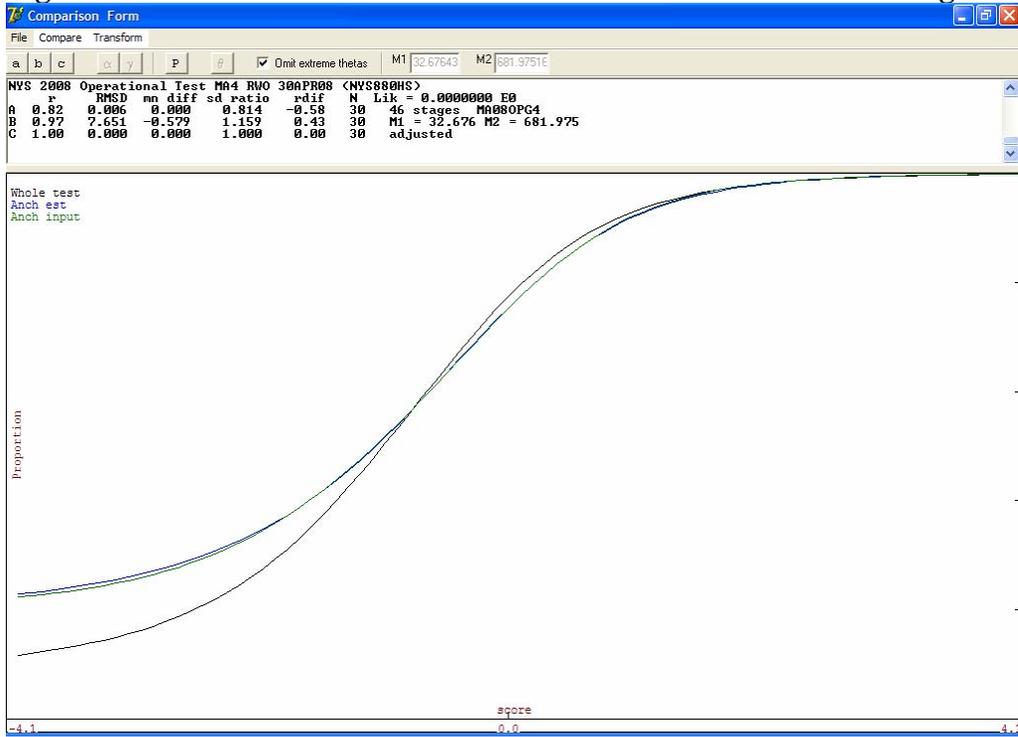
Table 14 provides a summary of anchor item evaluation and item flags.

**Table 14. Mathematics Anchor Evaluation Summary**

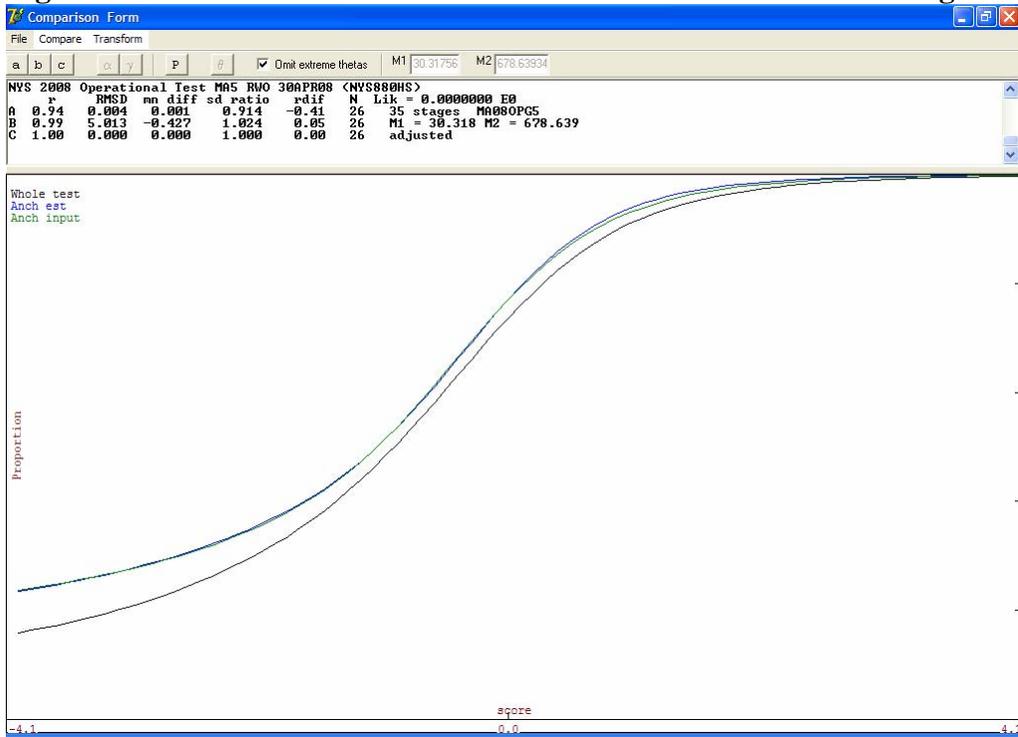
Grade	Number of Anchors	Anchor Input/ Estimate Correlation			Flagged Anchors			
		<i>a</i> -par	<i>b</i> -par	p-value	RMSD <i>a</i> -par	RMSD <i>b</i> -par	Delta	Lord's Chi-Square
3	25	0.793	0.852	0.933		7,18	2,18	7,18
4	30	0.822	0.973	0.973	12	2	2,11	2
5	26	0.942	0.985	0.964	18	17,19	17,19	
6	25	0.867	0.884	0.911	4	3	3	3
7	30	0.906	0.932	0.952	11	27	7,27	27
8	27	0.926	0.923	0.915	23	12,14,20	12,20	12,14



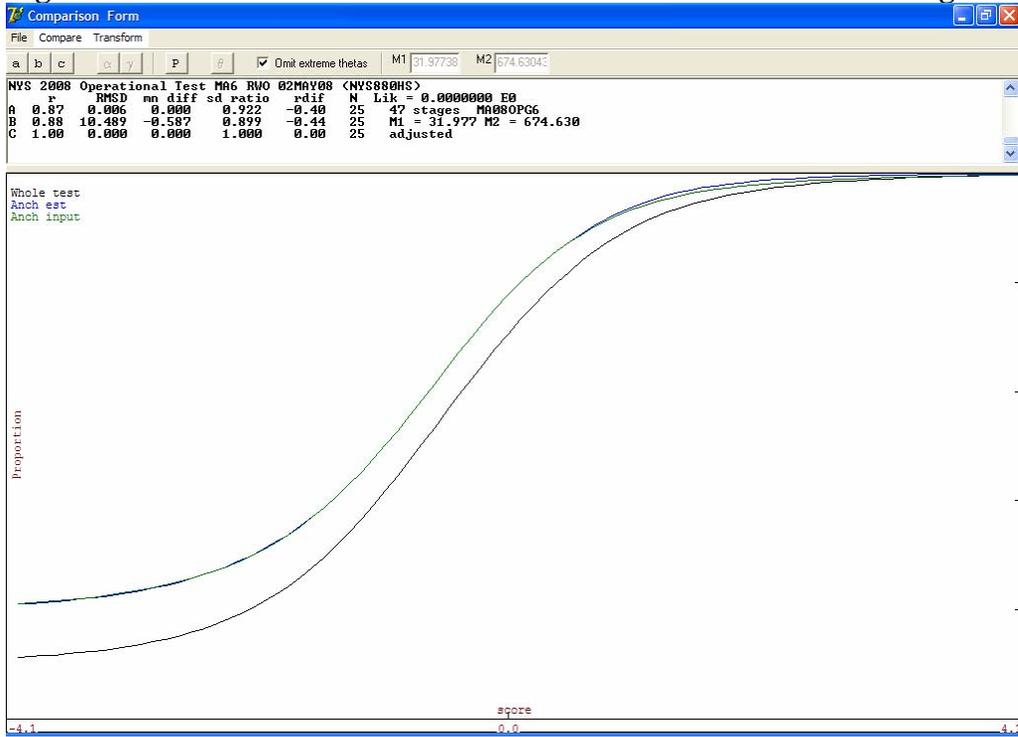
**Figure 2. Mathematics Grade 4 Anchor Set and Whole Test TCC Alignment**



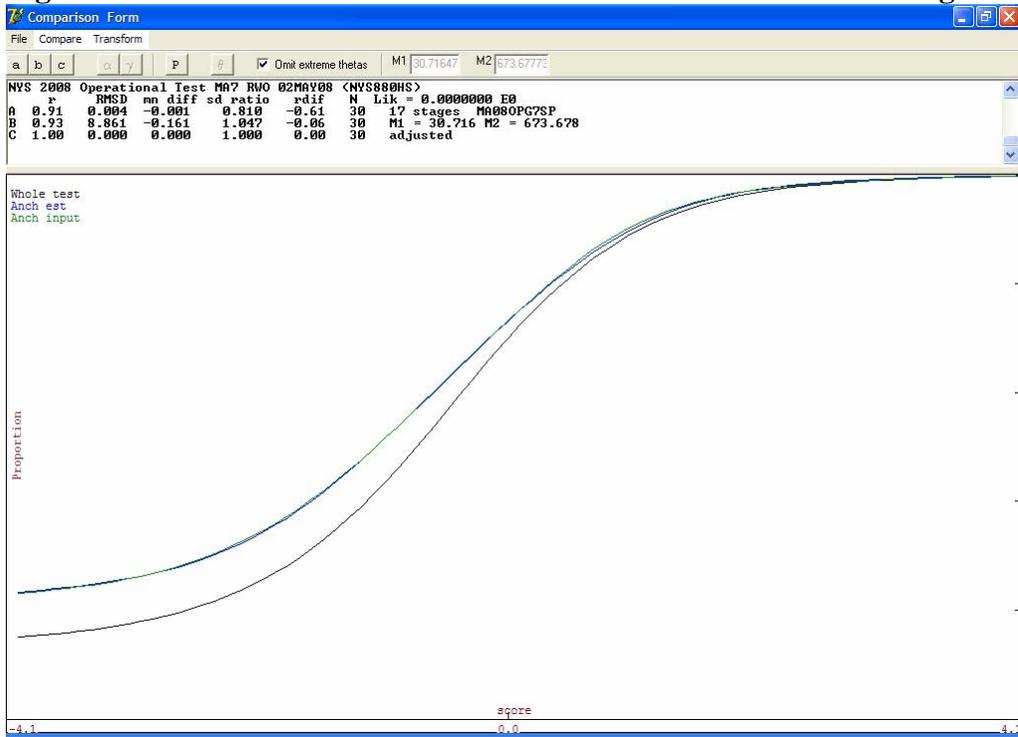
**Figure 3. Mathematics Grade 5 Anchor Set and Whole Test TCC Alignment**



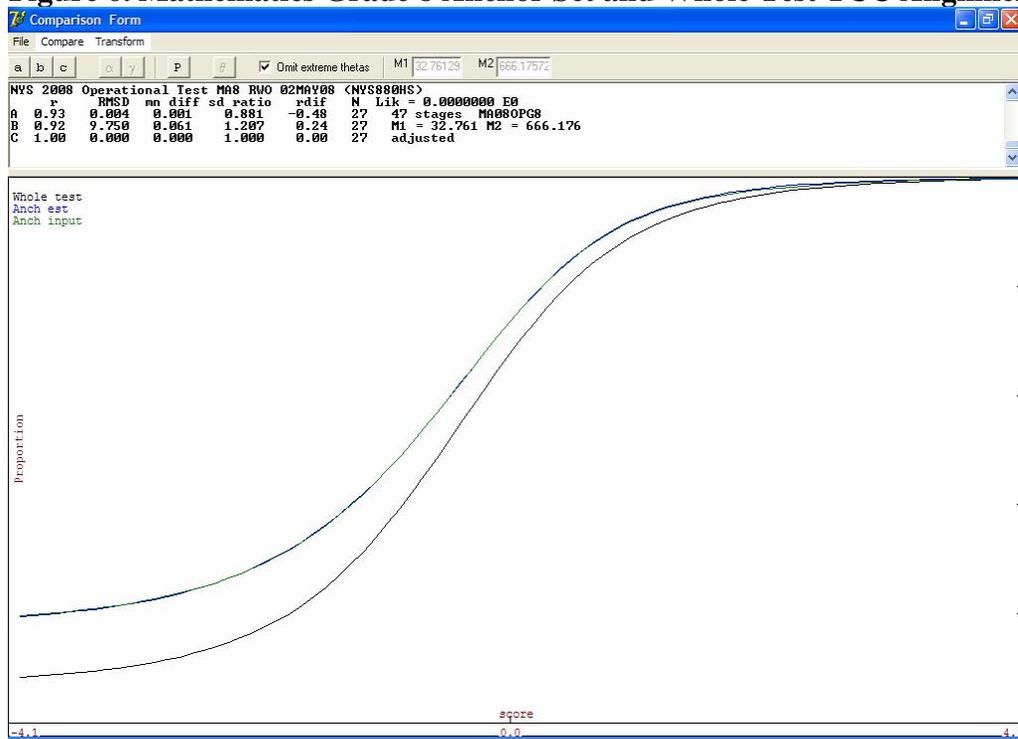
**Figure 4. Mathematics Grade 6 Anchor Set and Whole Test TCC Alignment**



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



**Figure 6. Mathematics 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 estimate parameters 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 of the figures, the alignment of anchor input and estimate parameters is very good, indicating overall good stability of anchor parameters between FT and OP test administrations.

It should be noted that in most 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 (i.e., Grades 3, 6, 8). 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 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 item parameters were estimated by the software PARDUX (Burket, 2002) and are presented in Tables 15a–15f. The parameter estimates are expressed in scale score metric. Descriptions of what each of the parameter variables mean is presented in the subsection depicting the IRT models and rationale.

**Table 15a. 2008 Operational Item Parameter Estimates, Grade 3**

Item	Max Pts	<i>a</i> -par/ alpha	<i>b</i> -par/ gamma1	<i>c</i> -par/ gamma2	gamma3
1	1	0.03433	625.2502	0.2000	
2	1	0.03049	630.5451	0.2000	
3	1	0.03661	645.4166	0.2000	
4	1	0.02956	634.4478	0.2000	
5	1	0.03559	641.7766	0.1223	
6	1	0.04340	656.8563	0.1114	
7	1	0.02251	621.1761	0.2000	
8	1	0.03202	652.1308	0.1529	
9	1	0.03492	629.8397	0.2000	
10	1	0.02790	619.0710	0.2000	
11	1	0.04341	661.5136	0.1171	
12	1	0.04003	651.8163	0.1231	
13	1	0.03934	649.3296	0.2527	
14	1	0.04035	633.3328	0.2000	
15	1	0.04039	651.8445	0.3298	
16	1	0.02814	666.7496	0.2000	
17	1	0.02406	632.4680	0.2000	
18	1	0.03007	642.8763	0.2000	
19	1	0.03788	650.4362	0.2478	
20	1	0.04165	636.6883	0.2000	
21	1	0.03485	622.4229	0.2000	
22	1	0.03343	650.2195	0.2000	
23	1	0.03267	661.1576	0.1180	
24	1	0.03123	658.3315	0.1156	
25	1	0.04783	664.8890	0.0786	
26	2	0.05161	33.5289	34.7309	
27	2	0.03277	20.5152	22.6501	
28	2	0.04748	31.2382	33.0765	
29	2	0.04970	33.0405	32.7245	
30	3	0.03983	23.7129	25.2145	27.6153
31	3	0.03267	21.2591	21.1367	20.8027

**Table 15b. 2008 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
1	1	0.02784	611.4863	0.2000	
2	1	0.02223	566.9116	0.2000	
3	1	0.02543	641.7311	0.3702	
4	1	0.04709	657.7242	0.1099	
5	1	0.01699	609.0261	0.2000	
6	1	0.01988	621.8652	0.2000	
7	1	0.02669	673.3604	0.3029	
8	1	0.03978	644.5557	0.1541	
9	1	0.02787	627.0570	0.2000	
10	1	0.02413	647.2898	0.2000	
11	1	0.03781	638.5018	0.2000	
12	1	0.03427	718.7979	0.3833	
13	1	0.02625	658.5644	0.2000	
14	1	0.02626	621.0676	0.2000	
15	1	0.0239	638.8638	0.2000	
16	1	0.01464	664.6635	0.2000	
17	1	0.03912	658.3041	0.1301	
18	1	0.02751	664.8837	0.1805	
19	1	0.03035	672.6435	0.2000	
20	1	0.01896	683.2184	0.1769	
21	1	0.02617	658.0112	0.2000	
22	1	0.01838	643.1376	0.2000	
23	1	0.03322	675.6236	0.1905	
24	1	0.03798	681.3790	0.2487	
25	1	0.04412	675.9279	0.1470	
26	1	0.02964	691.6428	0.0998	
27	1	0.04198	671.1219	0.3527	
28	1	0.04618	691.1801	0.1112	
29	1	0.02810	678.0126	0.1760	
30	1	0.02705	661.8978	0.2418	
31	2	0.03980	25.3946	25.2652	
32	2	0.04947	31.7685	32.5846	
33	2	0.04358	26.8471	29.6776	
34	2	0.04416	30.4862	28.9548	
35	3	0.02057	12.6413	14.6928	12.713
36	2	0.03854	26.4761	25.2005	
37	2	0.04030	26.1831	25.1782	
38	2	0.03268	21.1844	22.2251	
39	3	0.03074	18.2334	19.9967	20.715

*(Continued on next page)*

**Table 15b. 2008 Operational Item Parameter Estimates, Grade 4 (cont.)**

Item	Max Pts	<i>a</i> -par/ alpha	<i>b</i> -par/ gamma1	<i>c</i> -par/ gamma2	gamma3
40	2	0.03942	25.6572	25.1183	
41	2	0.02419	14.5007	14.2935	
42	2	0.04228	27.2232	28.1110	
43	3	0.03031	19.8307	18.5515	19.2437
44	3	0.03196	20.7824	19.6627	21.1577
45	2	0.03889	26.5257	25.2736	
46	2	0.03414	23.6957	21.2927	
47	2	0.03344	20.4330	20.7852	
48	2	0.03442	20.9207	22.0690	

**Table 15c. 2008 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	0.03013	649.6743	0.2000	
2	1	0.03745	660.9011	0.2472	
3	1	0.02388	586.0185	0.2000	
4	1	0.03127	662.0327	0.1841	
5	1	0.01787	644.7401	0.2000	
6	1	0.04031	667.6013	0.1371	
7	1	0.04692	681.1663	0.1326	
8	1	0.02922	653.2141	0.2000	
9	1	0.02967	655.3510	0.1841	
10	1	0.02625	677.3439	0.4814	
11	1	0.02205	624.7828	0.2000	
12	1	0.04631	646.7650	0.3328	
13	1	0.05407	667.8020	0.1242	
14	1	0.02420	593.8528	0.2000	
15	1	0.01759	595.6953	0.2000	
16	1	0.03454	671.9972	0.0882	
17	1	0.02400	672.3331	0.1260	
18	1	0.04774	676.4774	0.2157	
19	1	0.02142	640.7186	0.2000	
20	1	0.03880	674.8788	0.2827	
21	1	0.03301	654.9667	0.2000	
22	1	0.01880	578.5602	0.2000	
23	1	0.04275	653.7816	0.1860	
24	1	0.03330	669.1621	0.0915	

*(Continued on next page)*

**Table 15c. 2008 Operational Item Parameter Estimates, Grade 5 (cont.)**

Item	Max Pts	<i>a</i> -par/ alpha	<i>b</i> -par/ gamma1	<i>c</i> -par/ gamma2	gamma3
25	1	0.04117	668.6409	0.1818	
26	1	0.01599	635.9699	0.1430	
27	2	0.03823	24.3939	24.8693	
28	3	0.03159	20.6664	20.6419	21.6032
29	2	0.03175	20.2460	20.7904	
30	3	0.04050	25.8403	27.7922	27.8825
31	2	0.02409	15.3976	16.3265	
32	3	0.03187	17.9221	22.2754	21.2358
33	2	0.03944	22.9937	24.8253	
34	3	0.02950	18.0250	20.0095	18.6035

**Table 15d. 2008 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
1	1	0.03881	636.1817	0.1938	
2	1	0.04975	661.8613	0.1004	
3	1	0.01700	653.0756	0.1638	
4	1	0.03098	665.7202	0.3849	
5	1	0.03023	632.3477	0.2000	
6	1	0.02443	641.5952	0.1628	
7	1	0.02072	660.0728	0.1507	
8	1	0.03356	641.4959	0.2000	
9	1	0.01921	646.9359	0.2000	
10	1	0.01898	620.2113	0.2000	
11	1	0.04891	660.5938	0.1635	
12	1	0.03048	631.1019	0.2000	
13	1	0.04784	654.5779	0.2834	
14	1	0.06103	697.6693	0.4032	
15	1	0.03424	662.2089	0.2000	
16	1	0.03028	684.3991	0.1186	
17	1	0.03351	632.9966	0.2963	
18	1	0.02804	600.5942	0.2000	
19	1	0.03650	666.2261	0.1225	
20	1	0.02472	640.0280	0.2000	
21	1	0.04837	640.8737	0.1434	
22	1	0.03403	658.7795	0.3134	
23	1	0.04170	669.4921	0.0820	
24	1	0.03899	647.5701	0.1451	

*(Continued on next page)*

**Table 15d. 2008 Operational Item Parameter Estimates, Grade 6 (cont.)**

Item	Max Pts	<i>a</i> -par/ alpha	<i>b</i> -par/ gamma1	<i>c</i> -par/ gamma2	gamma3
25	1	0.02509	631.7213	0.1788	
26	2	0.03500	23.3205	20.8251	
27	3	0.03057	20.1398	18.7486	19.6694
28	3	0.03320	20.4220	21.2812	24.1753
29	2	0.03116	21.9032	21.2656	
30	3	0.04108	28.0626	27.8505	27.7886
31	2	0.03234	21.0444	20.6468	
32	2	0.05070	34.2984	34.8861	
33	3	0.03001	19.0353	19.8782	19.8931
34	2	0.04304	28.3690	28.0326	
35	2	0.05135	33.0280	33.3482	

**Table 15e. 2008 Operational Item Parameter Estimates, Grade 7**

Item	Max Pts	<i>a</i> -par/ alpha	<i>b</i> -par/ gamma1	<i>c</i> -par/ gamma2	gamma3
1	1	0.025250	634.3683	0.2000	
2	1	0.023530	639.9555	0.2000	
3	1	0.027280	635.6143	0.2000	
4	1	0.014420	595.7092	0.2000	
5	1	0.029130	633.5173	0.1673	
6	1	0.048790	693.3912	0.1998	
7	1	0.023560	658.2588	0.1307	
8	1	0.021700	626.9392	0.2000	
9	1	0.032120	640.9448	0.1430	
10	1	0.029190	639.8499	0.2000	
11	1	0.032810	676.2680	0.2211	
12	1	0.020080	660.0719	0.2000	
13	1	0.024160	649.6418	0.2000	
14	1	0.034140	651.5708	0.1386	
15	1	0.027640	647.8138	0.2360	
16	1	0.024610	635.6469	0.2000	
17	1	0.036990	652.4307	0.1668	
18	1	0.036840	634.8156	0.2000	
19	1	0.034320	680.0516	0.3554	
20	1	0.015110	684.3490	0.2000	
21	1	0.037610	685.8198	0.2383	
22	1	0.033870	636.2902	0.1914	
23	1	0.043800	674.2627	0.2878	

*(Continued on next page)*

**Table 15e. 2008 Operational Item Parameter Estimates, Grade 7 (cont.)**

Item	Max Pts	<i>a</i> -par/ alpha	<i>b</i> -par/ gamma1	<i>c</i> -par/ gamma2	gamma3
24	1	0.036570	700.3654	0.3269	
25	1	0.027870	686.7902	0.2121	
26	1	0.034770	670.7913	0.1813	
27	1	0.014480	621.5944	0.2000	
28	1	0.024860	662.2660	0.4567	
29	1	0.031240	624.4990	0.2000	
30	1	0.025090	633.9534	0.2000	
31	3	0.024270	15.4011	15.7491	16.1289
32	2	0.030980	21.1165	19.3127	
33	2	0.039260	26.5261	25.4714	
34	3	0.025890	15.3696	18.3338	17.1224
35	3	0.041260	26.8289	27.2740	28.2622
36	2	0.038870	26.1386	25.0886	
37	2	0.054440	35.2807	36.1977	
38	3	0.028790	18.2010	19.1844	18.1953

**Table 15f. 2008 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
1	1	0.03113	598.7266	0.2000	
2	1	0.03089	657.5236	0.3061	
3	1	0.01750	630.3878	0.2000	
4	1	0.02239	610.0970	0.2000	
5	1	0.03710	652.3215	0.1035	
6	1	0.02855	621.0581	0.2000	
7	1	0.03282	649.2592	0.1475	
8	1	0.02151	669.7732	0.2492	
9	1	0.03716	659.7260	0.1699	
10	1	0.03506	663.7114	0.1137	
11	1	0.03924	660.0845	0.1119	
12	1	0.01811	592.3226	0.2000	
13	1	0.04079	630.4462	0.1448	
14	1	0.03627	657.2193	0.2814	
15	1	0.02804	621.0265	0.2000	
16	1	0.03260	675.9458	0.1240	
17	1	0.02389	646.1852	0.1667	
18	1	0.01304	593.7604	0.2000	

*(Continued on next page)*

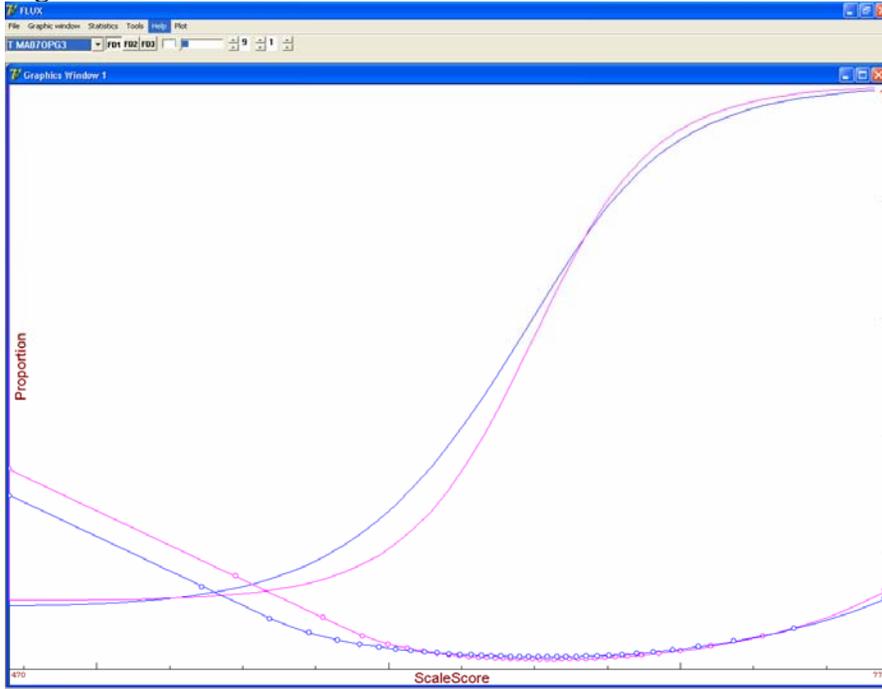
**Table 15f. 2008 Operational Item Parameter Estimates, Grade 8 (cont.)**

Item	Max Pts	<i>a</i> -par/ alpha	<i>b</i> -par/ gamma1	<i>c</i> -par/ gamma2	gamma3
19	1	0.03545	638.9227	0.1111	
20	1	0.02193	688.9215	0.2691	
21	1	0.02727	623.3851	0.1667	
22	1	0.04855	664.0538	0.1385	
23	1	0.04044	663.7283	0.1304	
24	1	0.01744	653.0678	0.2000	
25	1	0.03922	654.7242	0.2465	
26	1	0.03492	636.4811	0.0981	
27	1	0.04157	638.0134	0.1540	
28	3	0.04293	27.0491	27.0359	28.8897
29	2	0.04724	30.4802	30.7686	
30	2	0.02157	14.0796	13.8528	
31	3	0.03772	24.9480	25.5970	24.0776
32	2	0.03312	20.2710	22.1015	
33	2	0.02697	16.2008	17.9465	
34	2	0.03712	25.0966	21.7295	
35	2	0.04075	27.4639	27.1821	
36	2	0.05988	38.6133	40.2235	
37	2	0.04292	27.2094	29.9257	
38	2	0.04115	25.7045	28.1220	
39	2	0.02988	20.1469	19.3736	
40	3	0.02671	17.6538	17.2435	17.0684
41	3	0.03178	20.6457	20.1705	19.7053
42	3	0.03202	20.3795	19.6036	21.6225
43	3	0.02092	14.3469	13.6096	13.2573
44	2	0.05090	34.0157	33.7444	
45	2	0.03525	24.2853	20.9009	

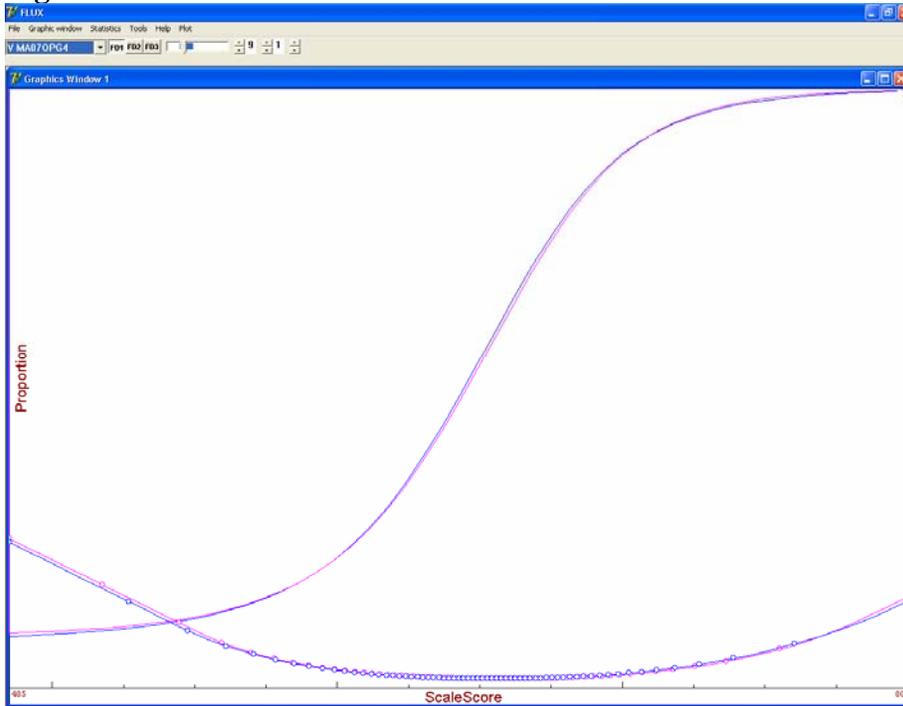
***Test Characteristic Curves***

Test Characteristic Curves (TCCs) provide an overview of the test in IRT scale score metric. The 2007 and 2008 TCCs were generated using final OP item parameters. TCCs are the summation of all the item characteristic curves (ICCs) for items that contribute to the OP scale score. Standard error (SE) curves graphically show the amount of measurement error at different ability levels. The 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 blue TCCs and SE curves represent the 2007 OP test and pink TCCs and SE curves represent the 2008 OP test.

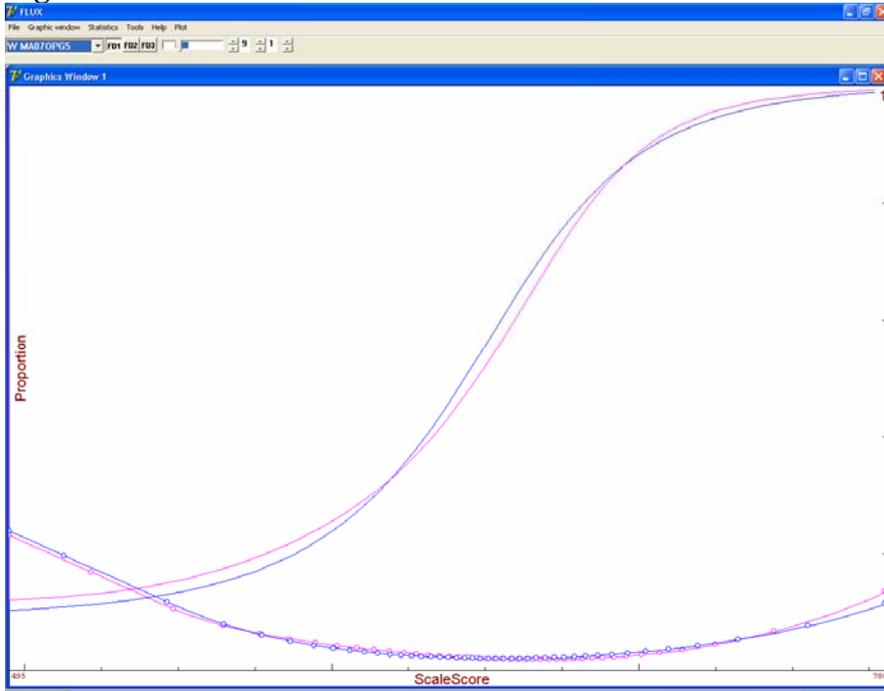
**Figure 7. Grade 3 Mathematics 2007 and 2008 OP TCCs and SE**



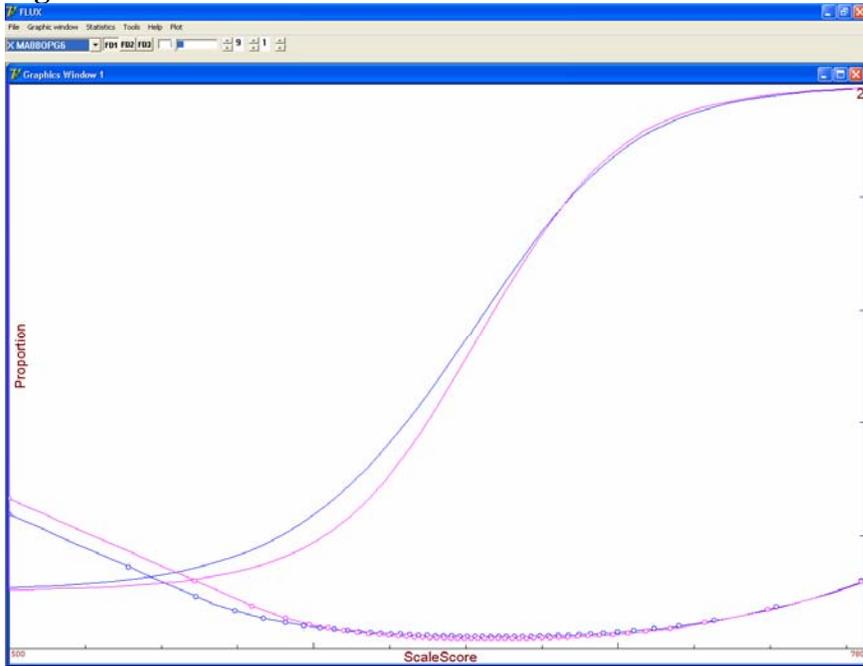
**Figure 8. Grade 4 Mathematics 2007 and 2008 OP TCCs and SE**



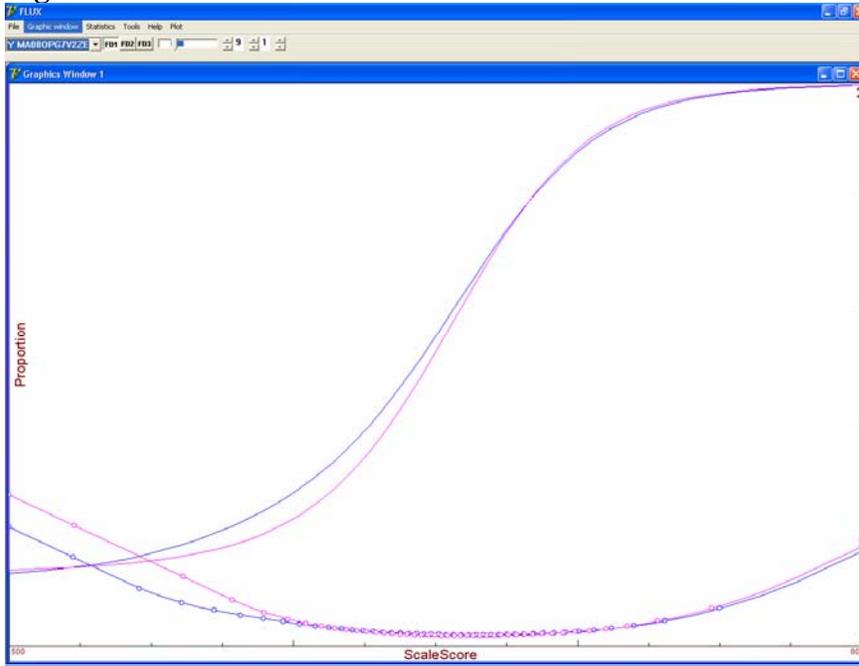
**Figure 9. Grade 5 Mathematics 2007 and 2008 OP TCCs and SE**



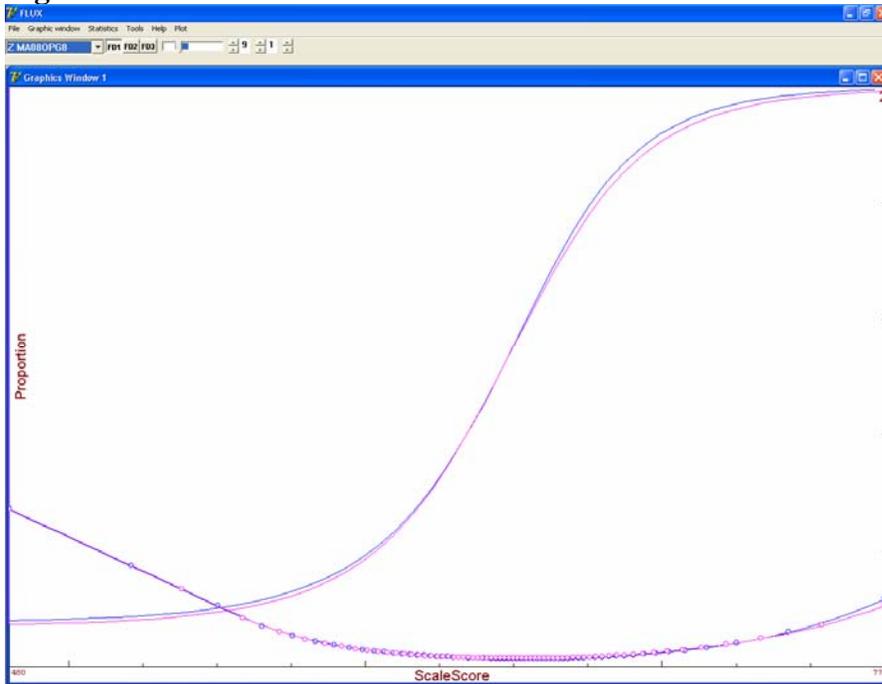
**Figure 10. Grade 6 Mathematics 2007 and 2008 OP TCCs and SE**



**Figure 11. Grade 7 Mathematics 2007 and 2008 OP TCCs and SE**



**Figure 12. Grade 8 Mathematics 2007 and 2008 OP TCCs and SE**



As seen in Figures 8, 9, and 12 very good alignments of the 2007 and 2008 TCCs and SE curves were found for Grades 4, 5 and 8. The TCCs for Grades 3, 6, and 7 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. It should be noted that potential differences in test form difficulty at different ability levels are accounted for in the equating and in the resulting raw score-to-scale score conversion tables, so that students of the same ability are expected to obtain the same scale score regardless of which form they took.

### ***Scoring Procedure***

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

The final item parameters in scale score metric were used to produce raw score-to-scale score conversion tables for the Grades 3–8 Mathematics Tests. An inverse TCC method was employed. The scoring tables were created using CTB/McGraw-Hill’s FLUX program. The inverse of the TCC procedure produces trait values based on unweighted raw scores. These estimates show negligible bias for tests with maximum possible raw scores of at least 30 points (Yen, 1984). The New York State Mathematics Tests have a maximum raw score ranging from 39 points (Grade 3) to 70 points (Grade 4). In the inverse TCC method, a student’s trait estimate is taken to be the trait value that has an expected raw score equal to the student’s observed raw score. It was found that for tests containing all MC items, the inverse of the TCC is an excellent first-order approximation to number correct maximum likelihood estimates (MLE) showing negligible bias for tests of at least 30 items. For tests with a mixture of MC and CR items, the MLE and TCC estimates are even more similar (Yen, 1984).

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

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

where

$x_i$  is a student’s observed raw score on item  $i$ .

$v_i$  is a weight specified in a scoring process (if no weights are specified then  $v_i=1$ ).

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

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

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

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

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

where

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

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

It should be noted that the information is estimated based on thetas in scale score metric; therefore, the SE is also expressed in 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 16a. Grade 3 Raw Score to Scale Score (with Standard Error)**

Raw Score	Scale Score	Standard Error
0	470	145
1	470	145
2	470	145
3	470	145
4	470	145
5	548	67
6	578	37
7	591	24
8	600	18
9	606	15
10	612	13
11	617	12
12	621	11
13	624	10
14	628	9
15	631	9
16	634	8
17	637	8
18	639	8
19	642	8
20	644	7
21	647	7
22	649	7
23	652	7
24	654	7

*(Continued on next page)*

**Table 16a. Grade 3 Raw Score to Scale Score (with Standard Error) (cont.)**

Raw Score	Scale Score	Standard Error
25	657	7
26	659	7
27	662	7
28	665	8
29	668	8
30	671	8
31	674	8
32	678	9
33	682	9
34	687	10
35	693	12
36	700	14
37	710	17
38	728	24
39	770	58

**Table 16b. Grade 4 Raw Score to Scale Score (with Standard Error)**

Raw Score	Scale Score	Standard Error
0	485	106
1	485	106
2	485	106
3	485	106
4	485	106
5	485	106
6	485	106
7	518	73
8	545	46
9	559	32
10	570	25
11	578	21
12	584	18
13	590	16
14	595	14
15	599	13
16	603	12
17	606	12
18	610	11
19	612	10
20	615	10

*(Continued on next page)*

**Table 16b. Grade 4 Raw Score to Scale Score (with Standard Error) (cont.)**

Raw Score	Scale Score	Standard Error
21	618	9
22	620	9
23	623	9
24	625	9
25	627	8
26	629	8
27	631	8
28	633	8
29	635	8
30	637	8
31	638	7
32	640	7
33	642	7
34	644	7
35	645	7
36	647	7
37	649	7
38	650	7
39	652	7
40	654	7
41	655	7
42	657	7
43	659	7
44	660	7
45	662	7
46	664	7
47	665	7
48	667	7
49	669	7
50	671	7
51	673	7
52	674	7
53	676	7
54	678	7
55	681	7
56	683	8
57	685	8
58	687	8
59	690	8
60	693	9

*(Continued on next page)*

**Table 16b. Grade 4 Raw Score to Scale Score (with Standard Error) (cont.)**

Raw Score	Scale Score	Standard Error
61	696	9
62	699	9
63	703	10
64	707	11
65	712	12
66	718	13
67	726	15
68	736	19
69	755	28
70	800	67

**Table 16c. Grade 5 Raw Score to Scale Score (with Standard Error)**

Raw Score	Scale Score	Standard Error
0	495	98
1	495	98
2	495	98
3	495	98
4	495	98
5	495	98
6	522	71
7	548	44
8	565	32
9	577	26
10	586	23
11	595	20
12	602	18
13	608	16
14	614	15
15	619	14
16	623	13
17	628	12
18	631	11
19	635	11
20	638	10
21	642	10
22	645	9
23	647	9
24	650	9
25	653	9
26	655	8

*(Continued on next page)*

**Table 16c. Grade 5 Raw Score to Scale Score (with Standard Error) (cont.)**

Raw Score	Scale Score	Standard Error
27	658	8
28	661	8
29	663	8
30	666	8
31	668	8
32	671	8
33	673	8
34	676	8
35	679	8
36	682	8
37	685	9
38	688	9
39	692	10
40	696	10
41	701	11
42	707	13
43	714	15
44	725	19
45	744	29
46	780	57

**Table 16d. Grade 6 Raw Score to Scale Score (with Standard Error)**

Raw Score	Scale Score	Standard Error
0	500	111
1	500	111
2	500	111
3	500	111
4	500	111
5	500	111
6	561	51
7	580	32
8	591	23
9	599	18
10	605	15
11	610	13
12	614	12
13	618	11
14	622	10
15	625	10

*(Continued on next page)*

**Table 16d. Grade 6 Raw Score to Scale Score (with Standard Error) (cont.)**

Raw Score	Scale Score	Standard Error
16	628	9
17	631	9
18	633	9
19	636	8
20	638	8
21	641	8
22	643	8
23	645	8
24	647	8
25	650	7
26	652	7
27	654	7
28	656	7
29	658	7
30	661	7
31	663	7
32	665	7
33	668	8
34	670	8
35	672	8
36	675	8
37	678	8
38	681	8
39	684	9
40	687	9
41	690	9
42	694	10
43	698	10
44	703	11
45	709	13
46	717	15
47	729	19
48	749	29
49	780	50

**Table 16e. Grade 7 Raw Score to Scale Score (with Standard Error)**

Raw Score	Scale Score	Standard Error
0	500	113
1	500	113
2	500	113
3	500	113
4	500	113
5	500	113
6	500	113
7	523	90
8	561	52
9	579	35
10	590	25
11	598	20
12	604	17
13	610	15
14	615	13
15	619	12
16	622	12
17	626	11
18	629	10
19	632	10
20	635	9
21	638	9
22	640	9
23	643	9
24	645	9
25	647	8
26	650	8
27	652	8
28	654	8
29	657	8
30	659	8
31	661	8
32	663	8
33	666	8
34	668	8
35	671	8
36	673	8
37	676	8
38	678	9
39	681	9

*(Continued on next page)*

**Table 16e. Grade 7 Raw Score to Scale Score (with Standard Error) (cont.)**

Raw Score	Scale Score	Standard Error
40	684	9
41	688	9
42	691	10
43	695	10
44	699	11
45	704	12
46	710	13
47	718	15
48	728	19
49	747	29
50	800	80

**Table 16f. Grade 8 Raw Score to Scale Score (with Standard Error)**

Raw Score	Scale Score	Standard Error
0	480	114
1	480	114
2	480	114
3	480	114
4	480	114
5	480	114
6	538	56
7	559	36
8	571	25
9	580	20
10	587	17
11	592	15
12	597	13
13	601	12
14	604	11
15	608	10
16	610	10
17	613	9
18	616	9
19	618	9
20	620	8
21	622	8
22	624	8
23	626	8
24	628	7

*(Continued on next page)*

**Table 16f. Grade 8 Raw Score to Scale Score (with Standard Error) (cont.)**

Raw Score	Scale Score	Standard Error
25	630	7
26	631	7
27	633	7
28	635	7
29	636	7
30	638	7
31	640	7
32	641	7
33	643	6
34	644	6
35	646	6
36	647	6
37	648	6
38	650	6
39	651	6
40	653	6
41	654	6
42	656	6
43	657	6
44	659	6
45	660	6
46	662	6
47	663	6
48	665	6
49	667	7
50	668	7
51	670	7
52	672	7
53	674	7
54	676	7
55	678	7
56	680	8
57	682	8
58	685	8
59	687	9
60	690	9
61	694	10
62	697	10
63	702	11
64	707	13

*(Continued on next page)*

**Table 16f. Grade 8 Raw Score to Scale Score (with Standard Error) (cont.)**

Raw Score	Scale Score	Standard Error
65	713	14
66	721	17
67	733	21
68	754	31
69	775	45

### ***Standard Performance Index***

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

Because objectives on all achievement tests are measured by relatively small numbers of items, CTB/McGraw-Hill’s scoring system looks not only at how many of those items the student answered correctly, but at additional information as well. In technical terms, the procedure CTB/McGraw-Hill uses to calculate the SPI is based on a combination of 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 New York State Mathematics Tests, the performance on objectives was tied to the Level III cut score by computing the SPI target ranges. The expected SPI cuts were computed for the scale scores that are 1 standard error above and 1 standard error below the Level III cut (scale score of 650 for all grades). Table 17 presents SPI target ranges. The objectives in this table are denoted as follows: 1—Number Sense and Operations, 2—Algebra, 3—Geometry, 4—Measurement, and 5—Statistics and Probability.

**Table 17. SPI Target Ranges**

Grade	Objective	# Items	Total Points	Level III Cut SPI Target Range
3	1	14	16	50–65
	2	5	7	53–64
	3	4	5	61–70
	4	5	5	54–68
	5	3	6	34–54
4	1	24	33	48–62
	2	7	11	45–58
	3	6	9	59–67
	4	7	10	40–50
	5	4	7	51–59
5	1	13	15	40–54
	2	4	6	37–55
	3	9	12	53–67
	4	4	6	55–66
	5	4	7	42–56
6	1	12	15	37–50
	2	6	9	56–72
	3	5	7	52–62
	4	4	6	43–55
	5	8	12	44–56
7	1	10	13	37–50
	2	6	8	46–63
	3	5	7	48–61
	4	6	7	49–63
	5	11	15	50–62
8	1	4	7	43–58
	2	19	28	42–53
	3	18	27	56–68
	4	4	7	60–69

The SPI is most meaningful in terms of its description of the student’s level of skills and knowledge measured by a given objective. The SPI increases the instructional value of test results by breaking down the information provided by the test into smaller, more manageable units. A total test score for a student in Grade 3 who scores below the average on the mathematics test does not provide sufficient information of what specific type of problem the student may be having. On the other hand, this kind of information may be provided by the SPI. For example, evidence that the student has attained an acceptable level of knowledge in the content strand of Number Sense, but has a low level of knowledge in Algebra, provides the teacher with a good indication of what type of educational assistance might be of greatest value to improving student achievement. Instruction focused on the identified needs of students has the best chance of helping those students increase their skills in the areas

measured by the test. SPI reports provide students, parents, and educators the opportunity to identify and target specific areas within the broader content domain 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 (strand) from year to year, differences in item difficulties in a given learning objective from year to year, and the fact that the learning objective sub-scores are not equated. The SPI scores are diagnostic scores and are best used at the classroom level to give teachers some insight into their students' strengths and weaknesses.

### ***IRT DIF Statistics***

In addition to classical DIF analysis, an IRT-based Linn-Harnisch statistical procedure was used to detect DIF on the Grades 3–8 Mathematics Tests (Linn and Harnisch, 1981). In this procedure, item parameters (discrimination, location, and guessing) and the scale score ( $\theta$ ) 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

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

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

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

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), Low Needs districts (by NRC code), and Spanish language test version. Most of the items flagged by IRT DIF were items from the Spanish language versions of the test. Also, as indicated in the classical DIF analysis section, items flagged for DIF do not necessarily display bias. Applying the Linn-Harnisch method revealed that two items were flagged for DIF on the Grade 3 test; four items were flagged on the Grade 4 test; three items were flagged on the Grade 5 test, five items were flagged on the Grade 6 test; eight items were flagged on the Grade 7 test; and three items were flagged on the Grade 8 test, as is shown in Table 18.

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

**Table 18. Number of Items Flagged for DIF by the Linn-Harnisch Method**

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

## **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 Mathematics Tests, we calculated two types of reliability statistics: Cronbach’s alpha (Cronbach, 1951) and Feldt-Raju coefficient (Qualls, 1995). These two measures are appropriate for assessment of a test’s internal consistency when a single test is administered to a group of examinees on one occasion. The reliability of the test is then estimated by considering how well the items that reflect the same construct yield similar results (or how consistent the results are for different items for the same construct measured by the test). Both Cronbach’s alpha and Feldt-Raju coefficient measures are appropriate for tests of multiple-item formats (MC and CR items). Please note that the reliability statistics in Section V, “Operational Test Data Collections and Classical Analysis,” are based upon the classical analysis and calibration sample, whereas the statistics in this section are based on the total student population data.

#### **Reliability for Total Test**

The overall test reliability is a very good indication of each test’s internal consistency. Included in Table 19 are the case counts (N-count), number of test items (# Items), Cronbach’s alpha and associated SEM, and Feldt-Raju coefficient and associated SEM obtained for the total mathematics tests.

**Table 19. Reliability and Standard Error of Measurement**

Grade	N-count	# Items	# RS Points	Cronbach’s Alpha	SEM of Cronbach	Feldt-Raju	SEM of Feldt-Raju
3	197306	31	39	0.89	2.23	0.90	2.10
4	198509	48	70	0.94	3.52	0.95	3.31
5	199474	34	46	0.90	2.88	0.91	2.69
6	201719	35	49	0.92	3.09	0.93	2.86
7	208694	38	50	0.90	3.23	0.92	3.01
8	210265	45	69	0.94	3.91	0.95	3.59

All the coefficients for total test reliability were in the range of 0.90–0.95, which indicated high internal consistency. As expected, the lowest reliabilities were found for shortest tests

(Grades 3, 5, 6, and 7) and the highest reliabilities are associated with the longer tests (Grades 4 and 8).

### Reliability for MC Items

In addition to overall test reliability, Cronbach’s alpha and Feldt-Raju 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 estimated for the overall test form. Table 20 presents reliabilities for the MC subsets.

**Table 20. 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	197306	25	0.86	1.50	0.87	1.48
4	198509	30	0.87	2.06	0.88	2.03
5	199474	26	0.87	1.89	0.87	1.85
6	201719	25	0.88	1.85	0.88	1.83
7	208694	30	0.86	2.14	0.86	2.13
8	210265	27	0.88	1.99	0.88	1.96

### Reliability for CR Items

Reliability coefficients were also computed for the subsets of CR items. It should be noted that the Grades 3–8 Mathematics Tests include 6–18 CR items depending on grade level. The results are presented in Table 21.

**Table 21. 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	197306	6	14	0.77	1.48	0.77	1.47
4	198509	18	40	0.91	2.65	0.91	2.59
5	199474	8	20	0.79	2.01	0.81	1.94
6	201719	10	24	0.86	2.24	0.87	2.19
7	208694	8	20	0.84	2.12	0.84	2.11
8	210265	18	41	0.92	3.08	0.93	3.00

Note: Results should be interpreted with caution for Grades 3, 5, 6, and 7 because the number of items is low.

### Test Reliability for NCLB Reporting Categories

In this section, reliability coefficients that were estimated for the population and NCLB reporting subgroups are presented. The reporting categories include the following: gender, ethnicity, needs resource code (NRC), limited English proficiency (LEP) status, students with disabilities (SWD), and students using test accommodations (SUA). For LEP students, reliability coefficients were computed for the following subgroups: students taking the

English version of the mathematics test and students taking the mathematics tests in each of the five translated languages (Chinese, Haitian-Creole, Korean, Russian, and Spanish). As shown in Tables 22a–22f, the estimated reliabilities for subgroups were close in magnitude to the test reliability estimates of the population. Cronbach’s alpha reliability coefficients across subgroups were equal to or greater than 0.85, with the exception of Grade 3 Low Needs district, Grade 3 Chinese, and Grade 4 Korean subgroups. Feldt-Raju reliability coefficients, which tend to be larger than the Cronbach’s alpha estimates for the same group, were all larger than 0.85, with the exception of Grade 3 Low Needs district and Grade 4 Korean subgroup. Overall, the New York State Mathematics Tests were found to have very good test internal consistency (reliability) for analyzed subgroups of examinees.

**Table 22a. 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	197306	0.89	2.23	0.90	2.10
Gender	Female	95843	0.88	2.23	0.89	2.10
	Male	101463	0.89	2.23	0.91	2.10
Ethnicity	Asian	14788	0.86	1.85	0.88	1.73
	Black	37706	0.89	2.48	0.90	2.34
	Hispanic	42308	0.89	2.40	0.90	2.26
	American Indian	970	0.89	2.42	0.90	2.29
	Multi-Racial	240	0.86	2.23	0.88	2.10
	White	101227	0.87	2.07	0.88	1.97
	Unknown	67	0.77	1.84	0.79	1.77
NRC	New York City	70240	0.90	2.32	0.91	2.16
	Big 4 Cites	8212	0.90	2.63	0.91	2.49
	High Needs Urban/Suburban	15993	0.88	2.35	0.90	2.23
	High Needs Rural	11467	0.87	2.30	0.89	2.19
	Average Needs	58664	0.86	2.12	0.88	2.02
	Low Needs	29518	0.83	1.89	0.85	1.80
	Charter	2999	0.86	2.25	0.87	2.14
SWD	All Codes	26744	0.91	2.66	0.92	2.51
SUA	All Codes	40980	0.90	2.62	0.91	2.46
LEP	English	15535	0.89	2.53	0.90	2.38
	Chinese	249	0.84	2.21	0.86	2.08
	Haitian-Creole	64	0.91	2.73	0.92	2.55
	Korean	68	0.85	1.85	0.87	1.70
	Russian	48	0.88	2.67	0.90	2.41
	Spanish	3258	0.89	2.68	0.90	2.53
	All Translations	3687	0.90	2.66	0.91	2.50

**Table 22b. 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	198509	0.94	3.52	0.95	3.31
Gender	Female	97188	0.93	3.53	0.94	3.32
	Male	101321	0.94	3.50	0.95	3.28
Ethnicity	Asian	14639	0.93	2.89	0.94	2.73
	Black	38146	0.94	3.83	0.94	3.59
	Hispanic	42116	0.93	3.74	0.94	3.52
	American Indian	945	0.93	3.74	0.94	3.52
	Multi-Racial	203	0.94	3.62	0.94	3.39
	White	102369	0.93	3.33	0.94	3.16
	Unknown	91	0.93	3.22	0.94	3.03
	NRC	New York City	70844	0.94	3.62	0.95
Big 4 Cites		7793	0.94	3.95	0.94	3.71
High Needs Urban/Suburban		15803	0.93	3.71	0.94	3.49
High Needs Rural		11488	0.93	3.70	0.93	3.51
Average Needs		59580	0.93	3.41	0.93	3.24
Low Needs		30342	0.92	3.02	0.92	2.89
Charter		2404	0.92	3.61	0.93	3.44
SWD	All Codes	29510	0.94	4.03	0.95	3.75
SUA	All Codes	42946	0.94	4.00	0.95	3.73
LEP	English	13265	0.93	3.94	0.94	3.70
	Chinese	281	0.93	3.32	0.94	3.12
	Haitian-Creole	80	0.95	4.09	0.96	3.68
	Korean	75	0.81	2.85	0.83	2.71
	Russian	54	0.94	3.96	0.95	3.69
	Spanish	2826	0.93	4.04	0.94	3.79
	All Translations	3316	0.94	3.99	0.95	3.72

**Table 22c. 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	199474	0.90	2.88	0.91	2.69
Gender	Female	97446	0.90	2.89	0.91	2.70
	Male	102028	0.90	2.87	0.92	2.67

*(Continued on next page)*

**Table 22c. Grade 5 Test Reliability by Subgroup (cont.)**

Group	Subgroup	N-count	Cronbach's Alpha	SEM of Cronbach	Feldt-Raju	SEM of Feldt-Raju
Ethnicity	Asian	15038	0.89	2.45	0.91	2.27
	Black	38471	0.89	3.06	0.91	2.88
	Hispanic	41465	0.90	3.00	0.91	2.82
	American Indian	908	0.89	3.03	0.90	2.85
	Multi-Racial	173	0.89	2.92	0.91	2.74
	White	103342	0.88	2.77	0.90	2.60
	Unknown	77	0.91	2.66	0.92	2.41
NRC	New York City	70240	0.91	2.92	0.92	2.71
	Big 4 Cites	7647	0.89	3.15	0.90	2.97
	High Needs Urban/Suburban	15386	0.90	3.00	0.91	2.82
	High Needs Rural	11370	0.88	2.97	0.89	2.80
	Average Needs	60466	0.88	2.83	0.89	2.66
	Low Needs	30802	0.87	2.58	0.88	2.43
	Charter	3290	0.87	2.97	0.88	2.83
SWD	All Codes	30058	0.90	3.11	0.91	2.94
SUA	All Codes	41669	0.90	3.10	0.91	2.93
LEP	English	10593	0.90	3.08	0.91	2.91
	Chinese	334	0.90	2.74	0.92	2.53
	Haitian-Creole	78	0.88	3.09	0.89	2.91
	Korean	54	0.91	2.42	0.92	2.21
	Russian	58	0.91	3.10	0.92	2.88
	Spanish	2573	0.89	3.12	0.90	2.95
	All Translations	3097	0.91	3.10	0.92	2.91

**Table 22d. 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	201719	0.92	3.09	0.93	2.86
Gender	Female	98210	0.92	3.07	0.93	2.85
	Male	103509	0.92	3.11	0.94	2.86
Ethnicity	Asian	15027	0.91	2.64	0.92	2.42
	Black	37994	0.91	3.28	0.92	3.07
	Hispanic	41414	0.92	3.25	0.93	3.03
	American Indian	916	0.92	3.21	0.93	2.98
	Multi-Racial	173	0.91	3.20	0.93	2.96

*(Continued on next page)*

**Table 22d. Grade 6 Test Reliability by Subgroup (cont.)**

Group	Subgroup	N-count	Cronbach's Alpha	SEM of Cronbach	Feldt-Raju	SEM of Feldt-Raju
Ethnicity	White	106111	0.91	2.95	0.92	2.74
	Unknown	84	0.93	3.01	0.94	2.72
NRC	New York City	69954	0.93	3.20	0.94	2.94
	Big 4 Cites	7675	0.91	3.34	0.92	3.13
	High Needs Urban/Suburban	15488	0.91	3.21	0.92	3.00
	High Needs Rural	11747	0.90	3.11	0.91	2.92
	Average Needs	62115	0.90	2.99	0.92	2.79
	Low Needs	31657	0.89	2.74	0.91	2.55
	Charter	2761	0.91	3.14	0.92	2.94
SWD	All Codes	30508	0.91	3.32	0.92	3.11
SUA	All Codes	39538	0.91	3.32	0.93	3.11
LEP	English	8679	0.91	3.34	0.92	3.12
	Chinese	442	0.90	3.06	0.91	2.83
	Haitian-Creole	140	0.92	3.32	0.93	3.06
	Korean	96	0.88	2.68	0.90	2.50
	Russian	68	0.91	3.21	0.92	3.00
	Spanish	2708	0.90	3.34	0.91	3.15
	All Translations	3454	0.92	3.35	0.93	3.11

**Table 22e. 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	208694	0.90	3.23	0.92	3.01
Gender	Female	102072	0.90	3.21	0.91	3.00
	Male	106622	0.91	3.24	0.92	3.01
Ethnicity	Asian	14982	0.90	2.81	0.92	2.61
	Black	40114	0.89	3.43	0.90	3.24
	Hispanic	42482	0.89	3.41	0.90	3.22
	American Indian	1016	0.89	3.40	0.90	3.21
	Multi-Racial	139	0.91	3.18	0.92	2.96
	White	109894	0.89	3.02	0.90	2.86
	Unknown	67	0.91	2.97	0.93	2.74

*(Continued on next page)*

**Table 22e. Grade 7 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	72418	0.90	3.34	0.92	3.11
	Big 4 Cites	8263	0.88	3.47	0.89	3.29
	High Needs Urban/Suburban	15857	0.89	3.37	0.90	3.18
	High Needs Rural	12506	0.88	3.27	0.89	3.10
	Average Needs	65039	0.89	3.07	0.90	2.91
	Low Needs	31884	0.88	2.75	0.89	2.63
	Charter	2289	0.88	3.20	0.89	3.04
SWD	All Codes	30398	0.88	3.47	0.90	3.29
SUA	All Codes	39953	0.89	3.49	0.90	3.31
LEP	English	8177	0.88	3.45	0.89	3.29
	Chinese	455	0.88	3.08	0.89	2.88
	Haitian-Creole	119	0.86	3.42	0.87	3.30
	Korean	77	0.87	2.77	0.88	2.60
	Russian	75	0.88	3.49	0.89	3.31
	Spanish	2901	0.87	3.34	0.88	3.19
	All Translations	3627	0.90	3.39	0.91	3.19

**Table 22f. 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	210265	0.94	3.91	0.95	3.59
Gender	Female	102514	0.94	3.89	0.95	3.58
	Male	107751	0.95	3.93	0.95	3.59
Ethnicity	Asian	14836	0.94	3.32	0.95	3.06
	Black	40502	0.94	4.11	0.95	3.82
	Hispanic	41989	0.94	4.08	0.95	3.79
	American Indian	1050	0.94	4.07	0.94	3.79
	Multi-Racial	125	0.93	3.89	0.94	3.64
	White	111707	0.93	3.74	0.94	3.47
	Unknown	56	0.95	3.64	0.96	3.29
NRC	New York City	72975	0.95	4.00	0.95	3.66
	Big 4 Cites	8413	0.93	4.13	0.94	3.85
	High Needs Urban/Suburban	16070	0.94	4.08	0.95	3.78

*(Continued on next page)*

**Table 22f. Grade 8 Test Reliability by Subgroup (cont.)**

Group	Subgroup	N-count	Cronbach's Alpha	SEM of Cronbach	Feldt-Raju	SEM of Feldt-Raju
NRC	Average Needs	66080	0.93	3.78	0.94	3.53
	Low Needs	31562	0.92	3.41	0.93	3.21
	Charter	1418	0.93	3.95	0.94	3.69
SWD	All Codes	30167	0.93	4.10	0.94	3.81
SUA	All Codes	39450	0.94	4.12	0.95	3.81
LEP	English	7208	0.94	4.10	0.95	3.79
	Chinese	507	0.94	3.62	0.95	3.35
	Haitian-Creole	107	0.94	4.05	0.95	3.77
	Korean	108	0.92	3.41	0.93	3.16
	Russian	65	0.94	4.15	0.95	3.80
	Spanish	3148	0.93	4.12	0.94	3.85
	All Translations	3935	0.94	4.10	0.95	3.79

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

The standard errors of measurement (SEMs), as computed from Cronbach's alpha and the Feldt-Raju reliability statistics, are presented in Table 19. SEMs based on Cronbach's alpha ranged from 2.23–3.91, which is reasonably small given the maximum number of score points on mathematics tests. In other words, the error of measurement from the observed test score ranged from approximately  $\pm 2$  to  $\pm 4$  raw score points. SEMs are directly related to reliability: the higher the reliability, the lower the standard error. As discussed, the reliability of these tests is relatively high, so it was expected that the SEMs would be very low.

The SEMs for subpopulations, as computed from Cronbach's alpha and the Feldt-Raju reliability statistics, are presented in Tables 22a–22f. The SEMs associated with all reliability estimates for all subpopulations are in the range of 1.70–4.15, which is acceptably close to those for the entire population. This narrow range indicates that across the Grades 3–8 Mathematics Tests, all students' test scores are reasonably reliable with minimal error.

### ***Performance Level Classification Consistency and Accuracy***

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

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

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

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

### **Consistency**

The results for classifying students into four performance levels are separated from results based solely on the Level III cut. Included in Tables 23 and 24 are case counts (N-count), classification consistency (Agreement), classification inconsistency (Inconsistency), and Cohen's kappa (Kappa). Consistency indicates the rate that a second administration would yield the same performance category designation (or a different designation for the inconsistency rate). The agreement index is a sum of the diagonal element in the contingency table. The inconsistency index is equal to the "1 - agreement index." Kappa is a measure of agreement corrected for chance.

Table 23 depicts the consistency study results based on the range of performance levels for all grades. Overall, between 78% and 81% of students were estimated to be classified consistently to one of the four performance categories. The coefficient kappa, which indicates the consistency of the placement in the absence of chance, ranged from 0.61 –0.71.

**Table 23. Decision Consistency (All Cuts)**

Grade	N-count	Agreement	Inconsistency	Kappa
3	197306	0.7853	0.2147	0.6072
4	198509	0.8139	0.1861	0.6949
5	199474	0.7773	0.2227	0.6315
6	201719	0.7932	0.2068	0.6724
7	208694	0.7778	0.2222	0.6518
8	210265	0.8134	0.1866	0.7092

Table 24 depicts the consistency study results based on two performance levels (passing and not passing), as defined by the Level III cut. Overall, about 92%–95% of the classifications of individual students were estimated to remain stable with a second administration. Kappa coefficients for classification consistency based on one cut ranged from 0.73–0.83.

**Table 24. Decision Consistency (Level III Cut)**

Grade	N-count	Agreement	Inconsistency	Kappa
3	197306	0.9531	0.0469	0.7252
4	198509	0.9474	0.0526	0.8088
5	199474	0.9287	0.0713	0.7491
6	201719	0.9334	0.0666	0.7976
7	208694	0.9175	0.0825	0.7539
8	210265	0.9279	0.0721	0.8278

### Accuracy

The results of classification accuracy are presented in Table 25. 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 his or her true ability approximately 83%–86% of the time across all performance levels and approximately 94%–97% of the time in regards to the Level III cut score.

**Table 25. 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	197306	<b>0.8293</b>	0.1318	0.0389	<b>0.9668</b>	0.0166	0.0166
4	198509	<b>0.8640</b>	0.0863	0.0497	<b>0.9624</b>	0.0212	0.0164
5	199474	<b>0.8358</b>	0.1041	0.0599	<b>0.9495</b>	0.0256	0.0249
6	201719	<b>0.8479</b>	0.1001	0.0521	<b>0.9514</b>	0.0301	0.0186
7	208694	<b>0.8406</b>	0.0932	0.0663	<b>0.9401</b>	0.0356	0.0243
8	210265	<b>0.8644</b>	0.0860	0.0496	<b>0.9461</b>	0.0369	0.0170

## 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 Mathematics Tests. These include the scale score means, standard deviations, and percentiles and performance level distributions for each grade’s population and specific subgroups. Gender, ethnic identification, needs resource category (NRC), limited English proficiency (LEP), students with disability (SWD), students using accommodation (SUA), and test language variables (Test Language) 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 H.

### *Scale Score Distribution Summary*

Scale score distribution summary tables are presented and discussed in Table 26. First, scale score statistics for total populations of students from public and charter schools are presented. Next, scale score statistics are presented for selected sub-groups in each grade level. The statistics for groups with small number counts should be interpreted with caution. Some general observations: Females and Males had very similar achievement patterns; Asian and White students outperformed their peers from other ethnic groups; Low Needs and Average Needs schools (as identified by NRC) outperformed other school types (New York City, Big 4 Cities, Urban/Suburban, Rural, and Charter); students taking the Chinese and Korean translations met or exceeded the population at every reported percentile, whereas the other translation subgroups (Haitian-Creole, Spanish, and Russian) were below the population scale score at each percentile; students with LEP, taking the mathematics test in English, SWD and/or SUA status, achieved below the State aggregate (All Students) in every percentile. This pattern of achievement was consistent across all grades. Note that complete scale score frequency distribution tables for the total population of students are located in Appendix H.

**Table 26. Mathematics Scale Score Distribution Summary Grades 3–8**

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	197306	688.36	34.39	649	668	687	710	728
4	198509	683.13	38.11	638	660	683	707	726
5	199474	679.65	36.38	638	658	679	701	725
6	201719	674.85	38.21	631	654	675	698	717
7	208694	674.60	38.27	632	652	673	695	718
8	210265	666.44	38.19	622	644	667	690	713

### Grade 3

Scale score statistics and N-counts of demographic groups for Grade 3 are presented in Table 27. The population scale score mean was 688.36 with a standard deviation of 34.39. The gender subgroups performed very similarly, with a mean difference of less than one scale score point. Asian, Multi-Racial, and White ethnic subgroups had scale score means that exceeded the State mean scale score on the test, as did students from Low Needs and Average Needs districts. The lowest performing NRC subgroup was the Big 4 Cities, with a mean of 665.68, and the lowest performing ethnic subgroup was Black (mean scale score of 674.82). SWD, SUA, and LEP without testing in an alternate language subgroup scored consistently below the Statewide percentile scale score rankings. At the 50<sup>th</sup> percentile, the scale scores on translated forms range from 659 (Haitian-Creole subgroup) to 710 (Korean subgroup), a difference that exceeds a standard deviation. The subgroup who used the Haitian-Creole translation had a scale score mean of 28 scale score units below the population mean, which was the lowest performing group analyzed. At the 50<sup>th</sup> percentile, the following groups exceeded the population scale score of 687: Asian (700), Multi-Racial (687), White (693), Low Needs (700), and students who used the Korean (710) translations.

**Table 27. 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	197306	688.36	34.39	649	668	687	710	728
Gender	Female	95843	688.86	33.66	652	668	687	710	728
	Male	101463	687.89	35.06	649	668	687	710	728
Ethnicity	Asian	14788	707.18	35.26	668	682	700	728	770
	Black	37706	674.82	32.62	639	657	674	693	710
	Hispanic	42308	679.09	32.53	644	662	678	693	710
	American Indian	970	678.41	33.75	644	659	678	693	710
	Multi-Racial	240	688.86	31.60	653	671	687	700	728
	White	101227	694.62	32.79	659	674	693	710	728
	Unknown	67	701.09	27.52	665	682	693	728	728
NRC	New York City	70240	684.93	35.94	644	665	682	700	728
	Big 4 Cities	8212	665.68	33.24	628	647	665	682	700
	High Needs Urban/Suburban	15993	680.44	31.54	644	662	678	700	710
	High Needs Rural	11467	683.11	31.18	649	665	682	700	728
	Average Needs	58664	691.91	31.73	657	674	687	710	728
	Low Needs	29518	702.65	32.12	668	682	700	728	770
	Charter	2999	686.90	30.52	652	668	682	700	728
SWD	All Codes	26744	661.07	35.70	621	642	665	682	700
SUA	All Codes	40980	665.57	34.32	628	649	668	682	700

*(Continued on next page)*

**Table 27. Scale Score Distribution Summary, by Subgroup, Grade 3 (cont.)**

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
LEP	LEP = Y Test Language = English	15535	671.96	31.74	637	654	671	687	710
Test Language	Chinese	249	693.23	32.33	657	674	687	710	728
	Haitian-Creole	64	654.25	44.91	617	638	659	673	700
	Korean	68	704.40	29.39	665	682	710	728	728
	Russian	48	673.02	31.05	634	649	671	693	710
	Spanish	3258	663.22	32.22	628	647	665	682	700
	All Translations	3687	665.97	33.72	628	649	668	682	700

**Grade 4**

Scale score statistics and N-counts of demographic groups for Grade 4 are presented in Table 28. The population scale score mean was 683.13 with a standard deviation of 38.11. The gender subgroups performed very similarly, with a mean difference of less than two scale score points. Asian and White students' scale score means exceeded the State mean scale score on the test. Asian students (the highest performing ethnic subgroup) exceeded the State mean by more than one-half of a standard deviation. Black, Hispanics, and American Indian subgroups had mean scale scores almost one standard deviation below the Asian subgroup. Students from Low Needs and Average Needs districts outperformed the other NRC subgroups. The lowest performing NRC subgroup was the Big 4 Cities, with a mean of 659.45, well more than one-half of a standard deviation below the State mean. SWD, SUA, and LEP without testing in an alternate language subgroup scored consistently below the Statewide percentile scale score rankings. Haitian-Creole and Spanish translated forms had means over one standard deviation below the population. LEP students who took the mathematics test in English outperformed the total group of students who took translated forms in terms of test mean and reported percentile scores except for Chinese and Korean translation subgroups. The subgroup who used the Haitian-Creole translation had a scale score mean much more than one standard deviation units below the population mean and was the lowest performing group analyzed. At the 50<sup>th</sup> percentile, the following groups exceeded the population scale score of 683: Male (685), Asian (707), White (690), Average Needs (687), Low Needs (699), and students who used the Chinese (690) and Korean (703) translations.

**Table 28. 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	198509	683.13	38.11	638	660	683	707	726
Gender	Female	97188	682.84	36.70	640	660	683	703	726
	Male	101321	683.41	39.42	637	660	685	707	726
Ethnicity	Asian	14639	706.82	38.55	664	683	707	726	755
	Black	38146	667.36	36.42	625	647	669	690	712
	Hispanic	42116	671.98	35.63	631	652	673	693	712
	American Indian	945	673.00	35.73	629	652	674	696	718
	Multi-Racial	203	679.21	36.85	637	660	681	703	718
	White	102369	690.30	36.00	649	669	690	712	736
	Unknown	91	693.90	36.91	657	669	693	718	736
NRC	New York City	70844	678.59	39.90	633	655	678	703	726
	Big 4 Cites	7793	659.45	37.06	615	638	660	683	703
	High Needs Urban/Suburban	15803	673.82	35.87	631	654	674	696	712
	High Needs Rural	11488	675.15	33.43	637	657	676	696	712
	Average Needs	59580	686.98	34.69	647	667	687	707	726
	Low Needs	30342	701.14	34.88	662	681	699	718	736
	Charter	2404	678.43	31.31	640	659	678	699	718
SWD	All Codes	29510	649.52	39.29	603	629	652	674	693
SUA	All Codes	42946	654.19	37.97	610	633	657	678	696
LEP	LEP = Y Test Language = English	13265	659.89	35.04	618	640	662	681	699
Test Language	Chinese	281	689.24	35.65	652	673	690	707	726
	Haitian-Creole	80	635.83	46.11	581	605	637	673	692
	Korean	75	712.69	36.74	676	685	703	726	755
	Russian	54	661.02	38.63	610	633	660	690	703
	Spanish	2826	650.46	35.64	606	631	654	674	690
	All Translations	3316	654.97	38.68	610	633	657	678	699

## Grade 5

Grade 5 demographic groups N-counts and scale score statistics are presented in Table 29. The population scale score mean was 679.65 with a standard deviation of 36.38. The gender subgroups performed very similarly, with a mean difference of less than one scale score point. Asian and White students' scale score means exceeded the State mean scale score on the test. Asian students (the highest performing ethnic subgroup) exceeded the State mean by close to 25 scale score points. American Indian, Black, and Hispanic ethnic subgroups had mean scale scores approximately one standard deviation below the Asian subgroup. Students from Low Needs and Average Needs districts outperformed the other NRC subgroups. The lowest performing NRC subgroup was the Big 4 Cities, with a mean of 652.96, nearly one-half of a standard deviation below the second lowest performing NRC subgroup (High Needs, Urban/Suburban, 669.27) and close to 50 scale score units below the Low Needs subgroup mean. SWD, SUA, and LEP without testing in an alternate language subgroup scored consistently below the Statewide percentile scale score rankings. Haitian-Creole and Spanish translated forms had scale score means close to one standard deviation below the population mean. The Haitian-Creole translation subgroup, which had a scale score mean (625.18) of more than 50 units below the population mean, was the lowest performing group analyzed. The Korean subgroup was the highest performing group analyzed, with a scale score mean of 705.20, more than one-half of a standard deviation above the population mean. At the 50<sup>th</sup> percentile, the following groups exceeded the population scale score of 679: Male (682), Asian (701), White (685), Average Needs (682), Low Needs (696), and students who used the Chinese (688) and Korean (701) translations.

**Table 29. 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	199474	679.65	36.38	638	658	679	701	725
Gender	Female	97446	679.57	35.68	638	658	679	701	725
	Male	102028	679.72	37.04	635	658	682	701	725
Ethnicity	Asian	15038	703.36	37.76	661	682	701	725	744
	Black	38471	664.03	34.76	623	645	666	685	701
	Hispanic	41465	668.91	35.03	628	650	671	688	707
	American Indian	908	667.22	32.95	628	647	668	688	707
	Multi-Racial	173	678.02	38.68	638	658	676	696	714
	White	103342	686.42	33.56	647	668	685	707	725
	Unknown	77	690.05	35.69	638	666	692	714	744
NRC	New York City	70240	676.39	39.27	631	653	676	701	725
	Big 4 Cites	7647	652.96	34.69	614	631	655	673	692
	High Needs Urban/Suburban	15386	669.27	34.81	628	650	671	688	707
	High Needs Rural	11370	673.21	31.28	638	655	673	692	707

*(Continued on next page)*

**Table 29. Scale Score Distribution Summary, by Subgroup, Grade 5 (cont.)**

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
NRC	Average Needs	60466	683.15	32.28	645	663	682	701	725
	Low Needs	30802	695.64	32.28	658	676	696	714	744
	Charter	3290	673.52	29.68	638	655	673	692	707
SWD	All Codes	30058	648.01	37.06	602	628	650	671	688
SUA	All Codes	41669	651.60	36.95	608	631	655	673	692
LEP	LEP = Y Test Language = English	10593	653.85	36.29	614	635	655	676	692
Test Language	Chinese	334	689.62	38.16	645	668	688	714	725
	Haitian-Creole	78	625.18	43.52	565	608	633	653	673
	Korean	54	705.20	40.93	661	682	701	725	780
	Russian	58	660.12	35.81	608	635	662	685	714
	Spanish	2573	643.69	37.40	602	623	647	668	685
	All Translations	3097	649.56	41.12	602	628	653	673	696

**Grade 6**

Grade 6 scale score statistics and N-counts of demographic groups are presented in Table 30. The population scale score mean was 674.85 with a standard deviation of 38.21. The gender subgroups performed very similarly, with a mean difference of less than two scale score points. Asian, and White students' scale score means exceeded the State mean scale score. American Indian, Black, and Hispanic ethnic subgroups had mean scale scores approximately one standard deviation below the Asian subgroup. Students from Low Needs and Average Needs districts outperformed the other NRC subgroups. The lowest performing NRC subgroup was the Big 4 Cities, with a mean of 649.54. New York City, High Needs Urban/Suburban, High Needs Rural, and Charter subgroups had similar scale score means (ranging from approximately 664–672). SWD, SUA, and LEP without testing in an alternate language subgroup scored consistently below the Statewide percentile scale score rankings. Haitian-Creole and Spanish translated forms had scale score means more than one standard deviation below the population. The Haitian-Creole translation subgroup, which had a scale score mean (632.68) more than 40 units below the population mean, was the lowest performing group analyzed. Asian students (the highest performing subgroup with a mean of 699.73) exceeded the State mean by nearly 25 scale score points. At the 50<sup>th</sup> percentile, the following groups exceeded the population scale score of 675: Asian (698), White (684), Average Needs (681), Low Needs (690), and students who used the Chinese (681) and Korean (694) translations.

**Table 30. 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	201719	674.85	38.21	631	654	675	698	717
Gender	Female	98210	675.81	36.72	633	656	675	698	717
	Male	103509	673.93	39.55	628	652	675	698	717
Ethnicity	Asian	15027	699.73	38.03	656	678	698	717	749
	Black	37994	657.62	36.59	614	638	661	681	698
	Hispanic	41414	661.11	36.95	618	641	663	684	703
	American Indian	916	664.12	38.68	618	645	668	687	709
	Multi-Racial	173	670.96	36.51	631	652	672	694	717
	White	106111	682.95	34.76	645	663	684	703	729
	Unknown	84	679.73	38.16	625	658	681	703	729
NRC	New York City	69954	668.24	40.94	622	645	668	694	717
	Big 4 Cites	7675	649.54	37.06	605	631	652	672	690
	High Needs Urban/Suburban	15488	664.84	36.18	622	647	668	687	703
	High Needs Rural	11747	671.73	33.32	636	654	672	690	709
	Average Needs	62115	679.82	33.85	643	661	681	698	717
	Low Needs	31657	692.96	33.34	656	672	690	709	729
	Charter	2761	671.35	32.21	633	652	672	690	709
SWD	All Codes	30508	637.83	40.46	591	618	643	663	681
SUA	All Codes	39538	642.51	40.11	599	622	647	668	687
LEP	LEP = Y Test Language = English	8679	643.41	38.80	599	625	647	668	687
Test Language	Chinese	442	678.54	33.02	641	661	681	698	717
	Haitian-Creole	140	632.68	45.80	591	610	636	660	683
	Korean	96	693.84	29.68	654	680	694	709	729
	Russian	68	653.54	30.27	610	639	661	677	690
	Spanish	2708	635.31	40.44	591	618	641	661	678
	All Translations	3454	642.72	42.83	591	622	647	670	690

**Grade 7**

N-counts and scale score statistics of demographic groups for Grade 7 are presented in Table 31. The population scale score mean was 674.60 with a standard deviation of 38.27. The gender subgroups performed very similarly, with a mean difference of less than four scale score points. Asian, and White subgroups' scale score means exceeded the State mean scale score. American Indian, Black, and Hispanic ethnic subgroups had mean scale scores between one-quarter and one-half of a standard deviation below the population. The lowest performing NRC subgroup, Big 4 Cities, had a scale score mean of 646.54, while the Low Needs subgroup's scale score mean was 696.61. SWD, SUA, and LEP without testing in an alternate language subgroup scored consistently below the Statewide percentile scale score

rankings and had means nearly one standard deviation below the population mean. Haitian-Creole and Spanish translation subgroups had scale score means close to one standard deviation below the population. The Haitian-Creole translation was the lowest performing group analyzed, while the Korean translation subgroup was the highest. At the 50<sup>th</sup> percentile, the following groups exceeded the population scale score of 673: Female (676), Asian (695), White (684), Average Needs (681), Low Needs (695), and students who used the Chinese (678) and Korean (695) translations.

**Table 31. 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	208694	674.60	38.27	632	652	673	695	718
Gender	Female	102072	676.05	36.73	635	654	676	695	718
	Male	106622	673.21	39.64	629	652	673	695	718
Ethnicity	Asian	14982	695.95	40.73	652	673	695	718	747
	Black	40114	654.93	33.85	619	638	657	676	691
	Hispanic	42482	659.57	33.99	622	643	661	678	699
	American Indian	1016	663.27	35.33	626	645	663	681	704
	Multi-Racial	139	676.40	40.19	635	654	673	699	728
	White	109894	684.77	35.79	647	663	684	704	728
	Unknown	67	685.82	35.97	640	663	688	710	728
	NRC	New York City	72418	664.28	38.11	622	643	663	684
	Big 4 Cites	8263	646.54	34.62	610	629	650	666	684
	High Needs Urban/Suburban	15857	664.66	34.86	629	645	663	684	704
	High Needs Rural	12506	670.87	31.85	638	654	671	688	704
	Average Needs	65039	682.50	34.49	645	663	681	699	718
	Low Needs	31884	696.61	35.84	659	676	695	718	747
	Charter	2289	673.06	29.37	638	654	673	691	710
SWD	All Codes	30398	639.23	38.42	598	622	643	663	678
SUA	All Codes	39953	642.95	37.90	604	626	647	666	684
LEP	LEP = Y Test Language = English	8177	640.37	37.66	598	622	645	661	678
Test Language	Chinese	455	680.56	31.67	647	663	678	699	718
	Haitian-Creole	119	629.86	41.69	579	615	640	652	666
	Korean	77	698.04	31.51	663	673	695	718	747
	Russian	75	652.40	33.77	610	640	659	676	684
	Spanish	2901	638.72	31.30	604	622	640	659	673
	All Translations	3627	645.22	35.61	604	626	645	666	688

## Grade 8

Grade 8 scale score statistics and N-counts of demographic groups are presented in Table 32. The population scale score mean was 666.44 with a standard deviation of 38.19. The gender subgroups performed similarly, with a mean difference of less than four scale score points. Asian, Multi-Racial, and White ethnic subgroups' scale score means exceeded the State mean scale score. The Black, Hispanic, and American Indian subgroups' scale score means were all close to or more than 10 scale score points below the population mean. The lowest performing NRC subgroup, Big 4 Cities, had a scale score mean of 636.74, while the Low Needs subgroup's scale score mean was 687.77, which indicated a large performance discrepancy by school district NRC designation. SWD, SUA, and LEP without testing in an alternate language subgroup scored consistently below the Statewide percentile scale score rankings. At the 50<sup>th</sup> percentile, the following groups exceeded the population scale score of 667: Female (668), Asian (690), White (676), Average Needs (674), Low Needs (685), and students who used the Chinese (680) and Korean (687) translations.

**Table 32. 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	210265	666.44	38.19	622	644	667	690	713
Gender	Female	102514	668.36	37.22	624	646	668	690	713
	Male	107751	664.62	39.00	618	643	665	687	713
Ethnicity	Asian	14836	692.23	39.08	646	668	690	713	754
	Black	40502	647.10	35.67	608	628	648	668	687
	Hispanic	41989	652.45	34.84	610	631	653	674	694
	American Indian	1050	654.14	34.74	616	635	656	676	694
	Multi-Racial	125	668.82	35.28	631	651	665	690	713
	White	111707	675.40	35.10	636	656	676	697	721
	Unknown	56	674.98	38.10	630	657	676	702	721
NRC	New York City	72975	658.35	39.55	613	635	657	682	707
	Big 4 Cities	8413	636.74	35.52	597	618	638	657	678
	High Needs Urban/Suburban	16070	655.28	35.63	613	635	656	676	697
	High Needs Rural	13077	661.11	31.96	626	644	662	680	697
	Average Needs	66080	673.45	33.96	636	654	674	694	713
	Low Needs	31562	687.77	33.13	650	667	685	707	733
	Charter	1418	663.37	30.85	626	646	665	682	697
SWD	All Codes	30167	629.98	37.93	587	610	635	654	672
SUA	All Codes	39450	635.78	38.42	592	616	640	660	678
LEP	LEP = Y Test Language = English	7208	639.36	37.50	597	618	641	662	682

(Continued on next page)

**Table 32. Scale Score Distribution Summary, by Subgroup, Grade 8 (cont.)**

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
Test Language	Chinese	507	678.75	34.37	633	660	680	702	721
	Haitian-Creole	107	638.16	35.79	597	620	641	659	687
	Korean	108	685.35	31.53	646	666	687	702	733
	Russian	65	653.82	32.90	613	628	657	674	697
	Spanish	3148	634.49	35.11	592	616	636	656	676
	All Translations	3935	642.01	38.61	597	620	643	665	687

***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. Table 33 shows the Mathematics cut scores used for classification of students to the four performance levels in 2008.

**Table 33. Mathematics Grades 3–8 Performance Level Cut Scores**

Grade	Level II cut	Level III cut	Level IV cut
3	624	650	703
4	622	650	702
5	619	650	699
6	616	650	696
7	611	650	693
8	616	650	701

Tables 34–40 show the performance level distribution for all examinees from public and charter school with valid scores. Table 34 presents performance level data for total populations of students in Grades 3–8. Tables 35–40 contain performance level data for selected subgroups of students. In general, these summaries reflect the same achievement trends in the scale score summary discussion. Male and Female students performed similarly across grades; however, Females consistently outperformed Males in all grade levels. More White and Asian students were classified in Level III and above, as compared to their peers from other ethnic subgroups. Students from Low and Average Needs districts outperformed students from High Needs districts (New York City, Big 4 Cities, High Needs Urban/Suburban, and High Needs Rural) and Charter schools. The subgroups that took Korean or Chinese test translations outperformed other test translation subgroups. The Level III and above rates for SWD and SUA subgroups were low compared to the total population of examinees. Across grades, the following subgroups consistently performed above the population average: Asian, White, Average Needs, Low Needs, Chinese translation, and Korean translation. Please note that the case counts for the Haitian/Creole, Korean, and Russian translations subgroups were very low, and the results might have been heavily influenced by very high and/or very low achieving individual students.

**Table 34. Mathematics Test Performance Level Distributions Grades 3–8**

Grade	N-count	Percent of New York State Population in Performance Level				
		Level I	Level II	Level III	Level IV	Levels III & IV
3	197306	2.26	7.80	63.60	26.34	89.94
4	198509	4.70	11.37	54.49	29.45	83.93
5	199474	3.77	12.93	56.27	27.04	83.31
6	201719	5.45	15.04	53.21	26.31	79.52
7	208694	3.82	17.15	51.25	27.77	79.02
8	210265	7.31	22.69	53.10	16.89	69.99

**Grade 3**

Performance level summaries and N-counts of demographic groups for Grade 3 are presented in Table 35. Statewide, 89.94% of third-graders were in Levels III and IV. American Indian, Black, and Hispanic subgroups had a lower percentage of students in Levels III and IV than the rest of the population, but the percentage of Asian, Multi-Racial, and White subgroups in Levels III and IV exceeded the overall State population. Student achievement varied widely by NRC subgroup, as well. Over 96% of students from Low Needs districts were classified in Levels III and IV; whereas, only about 71% of Big 4 Cities students were in Levels III and IV. Only about two-thirds of SWD, SUA, or those who took translated test forms were classified in Levels III or above; however, the subgroups for Korean and Chinese translations had more than 91% in Levels III and IV with Korean students having the greatest percentage, close to 97%.

**Table 35. Performance Level Distributions, by Subgroup, Grade 3**

Demographic Category (Subgroup)		N-count	Level I %	Level II %	Level III %	Level IV %	Levels III & IV %
State	All Students	197306	2.26	7.80	63.60	26.34	89.94
Gender	Female	95843	1.85	7.59	64.11	26.46	90.56
	Male	101463	2.64	8.01	63.12	26.23	89.35
Ethnicity	Asian	14788	0.86	2.64	48.15	48.34	96.50
	Black	37706	4.38	14.24	67.50	13.88	81.38
	Hispanic	42308	3.32	11.48	68.53	16.67	85.20
	American Indian	970	3.20	11.75	69.48	15.57	85.05
	Multi-Racial	240	1.25	7.08	67.08	24.58	91.67
	White	101227	1.22	4.59	62.28	31.91	94.19
	Unknown	67	0.00	1.49	56.72	41.79	98.51

*(Continued on next page)*

**Table 35. Performance Level Distributions, by Subgroup, Grade 3 (cont.)**

Demographic Category (Subgroup)		N-count	Level I %	Level II %	Level III %	Level IV %	Levels III & IV %
NRC	New York City	70240	3.13	9.82	63.27	23.78	87.04
	Big 4 Cites	8212	7.71	21.14	61.74	9.41	71.15
	High Needs Urban/Suburban	15993	2.68	10.66	69.36	17.30	86.66
	High Needs Rural	11467	2.09	8.62	70.11	19.19	89.29
	Average Needs	58664	1.20	5.14	65.25	28.41	93.66
	Low Needs	29518	0.48	2.53	55.68	41.30	96.98
	Charter	2999	1.03	7.97	68.66	22.34	91.00
SWD	All Codes	26744	10.79	22.36	59.64	7.21	66.84
SUA	All Codes	40980	8.22	19.36	63.86	8.57	72.42
LEP	LEP = Y Test Language = English	15535	4.81	15.31	68.83	11.05	79.87
Test Language	Chinese	249	0.40	7.63	62.25	29.72	91.97
	Haitian-Creole	64	15.63	20.31	57.81	6.25	64.06
	Korean	68	0.00	2.94	42.65	54.41	97.06
	Russian	48	2.08	25.00	62.50	10.42	72.92
	Spanish	3258	7.80	21.95	63.75	6.51	70.26
	All Translations	3687	7.21	20.64	63.14	9.00	72.15

**Grade 4**

Performance level summaries and N-counts of demographic groups for Grade 4 are presented in Table 36. Statewide, 83.93% of the fourth-grade population was placed in Levels III and IV. Around 6%–9% of American Indian, Black, and Hispanic students were Level I, as compared to only about 1.5% of Asian students and 3% of White students. American Indian, Black, and Hispanic ethnic subgroups had percentages of students in Levels III and IV ranging from 72%–78%, but the percentages of the White and Asian subgroups students meeting standards for Levels III and IV (89.82% and 94.69%) exceeded the population. Student achievement also varied widely by NRC subgroup. Almost 95% of students from Low Needs districts were meeting standards for Levels III and IV, but only about 63% Big 4 Cities students were. Only about half of SWD or SUA status students or those who took translated test forms met or exceeded the Level III cut; however, the Chinese translations subgroup had a very high percentage of students in Levels III and IV (90.39%). 100% of students in the Korean translation subgroup were in Levels III and IV. The following subgroups had a higher percentage of students meeting standards for Levels III and IV than the State population: Female, Asian, Multi-Racial, White, Average Needs, Low Needs, Chinese translation, and Korean translation.

**Table 36. 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	198509	4.70	11.37	54.49	29.45	83.93
Gender	Female	97188	4.12	11.58	55.85	28.45	84.30
	Male	101321	5.25	11.16	53.18	30.41	83.58
Ethnicity	Asian	14639	1.52	3.80	39.26	55.43	94.69
	Black	38146	8.84	19.17	56.62	15.38	71.99
	Hispanic	42116	6.81	16.36	58.72	18.11	76.84
	American Indian	945	6.56	15.24	58.41	19.79	78.20
	Multi-Racial	203	4.93	13.79	54.19	27.09	81.28
	White	102369	2.73	7.45	54.09	35.73	89.82
	Unknown	91	3.30	4.40	59.34	32.97	92.31
NRC	New York City	70844	6.37	14.14	53.35	26.15	79.50
	Big 4 Cites	7793	12.95	24.43	51.61	11.01	62.62
	High Needs Urban/Suburban	15803	6.17	15.65	58.24	19.95	78.18
	High Needs Rural	11488	4.62	13.83	62.71	18.84	81.55
	Average Needs	59580	2.90	8.42	57.26	31.42	88.68
	Low Needs	30342	1.21	3.86	47.10	47.82	94.92
	Charter	2404	2.41	13.89	61.69	22.01	83.69
SWD	All Codes	29510	20.15	26.32	46.78	6.75	53.53
SUA	All Codes	42946	16.38	24.60	51.10	7.92	59.02
LEP	LEP = Y Test Language = English	13265	11.28	23.20	56.51	9.01	65.52
Test Language	Chinese	281	3.56	6.05	53.74	36.65	90.39
	Haitian-Creole	80	36.25	22.50	38.75	2.50	41.25
	Korean	75	0.00	0.00	48.00	52.00	100.00
	Russian	54	14.81	18.52	55.56	11.11	66.67
	Spanish	2826	18.01	27.39	49.12	5.48	54.60
	All Translations	3316	16.77	24.70	49.34	9.20	58.53

**Grade 5**

Performance level summaries and N-counts of demographic groups for Grade 5 are presented in Table 37. Statewide, 86.31% of the fifth-grade population was placed in Levels III and IV. There was little performance differentiation by gender subgroup, with less than 1% difference between each level. However, across ethnic and test translation subgroups, there were marked differences. American Indian, Black, and Hispanic ethnic subgroups were well below the State average of students meeting standards for Levels III and IV (ranging from 70%–75%), as compared to the percentage of Asian, Multi-Racial, and White students meeting standards for Levels III and IV (between 85% and 95%). Nearly 95% of students from Low Needs districts were in Levels III or IV, but only slightly more than 55% of the Big 4 Cities students were. Only about 5%–8% of SWD or SUA subgroups were placed in

Level IV, compared to the population's 27.04% in Level IV. Less than 10% of students who took translated test forms or who reported LEP with English language test forms were placed in Level IV, except for Russian (13.79%) and the Chinese and Korean translation subgroups that had very high percentages of students in Level IV (36.83% and 53.70%). The following subgroups had a higher percentage of students meeting standards for Levels III and IV than the State population: Female, Asian, Multi-Racial, White, Average Needs, Low Needs, Chinese translation, and Korean translation.

**Table 37. 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	199474	3.77	12.93	56.27	27.04	83.31
Gender	Female	97446	3.38	13.07	57.20	26.35	83.55
	Male	102028	4.14	12.78	55.39	27.69	83.08
Ethnicity	Asian	15038	1.36	4.41	40.18	54.06	94.23
	Black	38471	7.10	22.29	57.62	13.00	70.61
	Hispanic	41465	5.82	18.84	58.87	16.47	75.34
	American Indian	908	5.18	20.15	59.58	15.09	74.67
	Multi-Racial	173	2.89	12.72	61.27	23.12	84.39
	White	103342	2.05	8.25	57.03	32.67	89.71
	Unknown	77	2.60	7.79	50.65	38.96	89.61
NRC	New York City	70240	5.12	15.85	53.90	25.13	79.03
	Big 4 Cites	7647	12.11	30.22	50.31	7.36	57.67
	High Needs Urban/Suburban	15386	5.71	18.04	59.52	16.73	76.25
	High Needs Rural	11370	3.48	15.22	63.76	17.54	81.29
	Average Needs	60466	2.03	9.68	59.87	28.42	88.29
	Low Needs	30802	0.94	4.49	51.05	43.52	94.57
	Charter	3290	2.19	15.68	65.23	16.90	82.13
SWD	All Codes	30058	16.01	31.00	47.34	5.65	52.99
SUA	All Codes	41669	13.87	28.99	50.09	7.05	57.13
LEP	LEP = Y Test Language = English	10593	12.06	28.61	51.52	7.80	59.32
Test Language	Chinese	334	2.10	11.68	49.40	36.83	86.23
	Haitian-Creole	78	33.33	37.18	29.49	0.00	29.49
	Korean	54	1.85	7.41	37.04	53.70	90.74
	Russian	58	10.34	25.86	50.00	13.79	63.79
	Spanish	2573	19.32	32.02	43.96	4.70	48.66
	All Translations	3097	17.34	29.42	44.17	9.07	53.25

## Grade 6

Performance level summaries and N-counts of demographic groups for Grade 6 are presented in Table 38. Statewide, 79.52% of the sixth-grade population was placed in Levels III and IV. There was a slight performance differentiation by gender subgroup with less than 2% difference between each level. There were marked differences across ethnic and test translation subgroups. About 10% of American Indian, Black, and Hispanic students were in Level I, as compared to less than 2% of Asian students and about 3% of White students. American Indian, Black, and Hispanic ethnic subgroups were below the State average of students meeting standards for Levels III and IV (ranging from 64%–70%), as compared to the percentage of Asian, and White students meeting standards for Levels III and IV (92.75% and 87.78%). About 93% of students from Low Needs districts were in Levels III or IV, but only about 54% of the Big 4 Cities students were. Only about 4%–6% of SWD and SUA subgroups were placed in Level IV, compared to the population’s 26.31% in Level IV. Less than 6% of students who took translated test forms or who reported LEP with English language test forms were placed in Level IV, except for the Chinese and Korean translation subgroups that had very high percentages of students in Level IV (26.24% and 48.96%). The following subgroups had a higher percentage of students meeting standards for Levels III and IV than the State population: Asian, White, High Needs Rural, Average Needs, Low Needs, Chinese translation, and Korean translation.

**Table 38. 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	201719	5.45	15.04	53.21	26.31	79.52
Gender	Female	98210	5.45	15.04	53.21	26.31	79.52
	Male	103509	6.25	15.71	51.53	26.50	78.04
Ethnicity	Asian	15027	1.86	5.40	38.90	53.84	92.75
	Black	37994	10.04	25.90	52.44	11.62	64.06
	Hispanic	41414	8.85	23.22	54.04	13.88	67.93
	American Indian	916	9.06	20.20	53.17	17.58	70.74
	Multi-Racial	173	5.20	19.08	54.34	21.39	75.72
	White	106111	2.95	9.27	55.18	32.60	87.78
	Unknown	84	5.95	9.52	53.57	30.95	84.52
NRC	New York City	69954	7.80	20.61	49.53	22.06	71.59
	Big 4 Cites	7675	14.59	30.97	46.79	7.65	54.44
	High Needs Urban/Suburban	15488	7.49	19.62	57.06	15.83	72.89
	High Needs Rural	11747	4.61	14.24	61.54	19.61	81.15
	Average Needs	62115	3.15	10.57	57.71	28.57	86.29
	Low Needs	31657	1.43	5.24	49.15	44.17	93.33
	Charter	2761	3.48	18.98	57.33	20.21	77.54
SWD	All Codes	30508	23.41	33.85	38.74	4.00	42.74
SUA	All Codes	39538	19.98	32.00	42.60	5.42	48.02

*(Continued on next page)*

**Table 38. Performance Level Distribution Summary, by Subgroup, Grade 6 (cont.)**

Demographic Category (Subgroup)		N-count	Level I %	Level II %	Level III %	Level IV %	Levels III & IV %
LEP	LEP = Y Test Language = English	8679	18.50	33.82	41.94	5.74	47.68
Test Language	Chinese	442	2.94	11.09	59.73	26.24	85.97
	Haitian-Creole	140	30.00	31.43	35.00	3.57	38.57
	Korean	96	0.00	7.29	43.75	48.96	92.71
	Russian	68	16.18	20.59	61.76	1.47	63.24
	Spanish	2708	24.00	36.41	36.74	2.84	39.59
	All Translations	3454	20.73	31.85	40.30	7.12	47.42

### Grade 7

Performance level summaries and N-counts of demographic groups for Grade 7 are presented in Table 39. Statewide, 79.02% of the seventh-grade population was placed in Levels III and IV. Overall there was only slight performance differentiation by gender subgroup with only about 3% difference between each level. However, there were marked differences across ethnic and test translation subgroups. Black, Hispanic, and American Indian ethnic subgroups had around 60%–70% of students meeting standards for Levels III and IV, with less than 16% of those students in Level IV, whereas over 91% of Asian students were meeting standards for Levels III and IV (and almost 52% were in Level IV.) About 50% of Big 4 Cities students were meeting standards for Levels III and IV, with less than 7% in Level IV, yet over 94% of students from Low Needs districts were meeting standards for Levels III and IV (with about 51% in Level IV). Less than 6% of SWD and SUA subgroups were placed in Level IV, and close to 15% were in Level I. Less than 8% of students who took translated test forms or who reported LEP with English language test forms were placed in Level IV, except for the Chinese and Korean translation subgroups that had very high rates (31.65% and 55.84%). Across all subgroups, the Haitian-Creole translation subgroup had the largest percentage of students placed in Level I (23.53%) and the Korean translation subgroup had the largest percentage of students (97.40%) who met the standards for Levels III and IV. The following subgroups had a higher percentage of students meeting Levels III and IV standards than the State population: Female, Asian, Multi-Racial, White, High Needs Rural, Average Needs, Low Needs, Chinese translation, and Korean translation.

**Table 39. 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	208694	3.82	17.15	51.25	27.77	79.02
Gender	Female	102072	2.92	16.26	52.58	28.23	80.81
	Male	106622	4.69	18.01	49.97	27.33	77.30
Ethnicity	Asian	14982	1.80	6.79	40.04	51.36	91.40
	Black	40114	7.41	31.60	51.33	9.66	60.99
	Hispanic	42482	6.18	27.26	53.67	12.89	66.56
	American Indian	1016	4.82	24.70	54.63	15.85	70.47
	Multi-Racial	139	4.32	16.55	46.04	33.09	79.14
	White	109894	1.87	9.32	51.79	37.01	88.81
	Unknown	67	1.49	14.93	38.81	44.78	83.58
NRC	New York City	72418	5.83	25.42	49.98	18.77	68.75
	Big 4 Cites	8263	11.15	38.75	43.88	6.22	50.10
	High Needs Urban/Suburban	15857	4.94	22.93	55.65	16.48	72.13
	High Needs Rural	12506	2.98	16.10	60.50	20.42	80.92
	Average Needs	65039	1.81	9.93	54.76	33.50	88.26
	Low Needs	31884	0.88	4.71	43.13	51.29	94.41
	Charter	2289	1.53	17.52	57.23	23.72	80.95
SWD	All Codes	30398	17.20	39.70	38.88	4.22	43.10
SUA	All Codes	39953	14.98	37.74	41.70	5.58	47.28
LEP	LEP = Y Test Language = English	8177	16.11	40.10	39.34	4.45	43.79
Test Language	Chinese	455	1.98	9.01	57.36	31.65	89.01
	Haitian-Creole	119	23.53	44.54	28.57	3.36	31.93
	Korean	77	1.30	1.30	41.56	55.84	97.40
	Russian	75	12.00	26.67	56.00	5.33	61.33
	Spanish	2901	16.89	44.43	36.19	2.48	38.68
	All Translations	3627	14.81	38.71	39.12	7.36	46.48

**Grade 8**

Performance level summaries and N-counts of demographic groups for Grade 8 are presented in Table 40. Statewide, 69.99% of the eighth-grade population was placed in Levels III and IV. Overall, there was little performance differentiation by gender subgroup, with less than 3% difference between each level percentage. Across ethnic and test translation subgroups, there were marked differences in performance. Around 10%–15% of Black, Hispanic, and American Indian students were in Level I, compared to less than 4% of Asian, and White students. American Indian, Black, Hispanic, and Multi-Racial subgroups had around 50%–60% of students meeting standards for Levels III and IV, respectively, whereas over 88% of Asian students were meeting Level III and IV standards. About 34% of Big 4 Cities students were in Levels III and IV, yet over 80% of students from Low Needs districts were classified

in these proficiency levels. Approximately 20%–30% of SWD, SUA, and LEP students were placed in Level I. Less than 8% of students who took translated test forms or who reported LEP with English language test forms were placed in Level IV, except for the Chinese and Korean translation subgroups that had a very high percentage of students in Level IV (25.84% and 28.70%). Across all subgroups, the Spanish translation subgroup had the largest percentage of students placed in Level I (24.36%), and the Asian subgroup had the largest percentage of students placed in Level IV (41.06%). The following subgroups had a higher percentage of students meeting standards for Levels III and IV than the State population: Female, Asian, Multi-Racial, White, Average Needs, Low Needs, Charter, Chinese translation, and Korean translation.

**Table 40. 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	210265	7.31	22.69	53.10	16.89	69.99
Gender	Female	102514	6.00	22.20	54.14	17.66	71.81
	Male	107751	8.56	23.17	52.11	16.16	68.27
Ethnicity	Asian	14836	2.57	9.09	47.28	41.06	88.34
	Black	40502	14.42	37.09	42.88	5.61	48.49
	Hispanic	41989	11.53	33.27	47.88	7.31	55.20
	American Indian	1050	9.90	29.62	53.43	7.05	60.48
	Multi-Racial	125	6.40	15.20	61.60	16.80	78.40
	White	111707	3.75	15.25	59.53	21.46	80.99
	Unknown	56	7.14	12.50	51.79	28.57	80.36
NRC	New York City	72975	10.60	30.15	45.98	13.26	59.25
	Big 4 Cites	8413	21.67	43.69	31.53	3.10	34.64
	High Needs Urban/Suburban	16070	10.18	31.59	49.41	8.82	58.23
	High Needs Rural	13077	6.18	25.46	59.11	9.25	68.36
	Average Needs	66080	3.65	16.04	61.47	18.84	80.31
	Low Needs	31562	1.53	7.84	57.81	32.82	90.63
	Charter	1418	4.80	24.40	61.07	9.73	70.80
SWD	All Codes	30167	28.94	39.77	29.75	1.54	31.30
SUA	All Codes	39450	24.37	38.03	34.60	3.00	37.60
LEP	LEP = Y Test Language = English	7208	21.75	38.08	35.95	4.22	40.16
Test Language	Chinese	507	3.94	12.03	58.19	25.84	84.02
	Haitian-Creole	107	19.63	41.12	35.51	3.74	39.25
	Korean	108	1.85	12.04	57.41	28.70	86.11
	Russian	65	12.31	27.69	52.31	7.69	60.00
	Spanish	3148	24.36	41.90	31.77	1.97	33.74
	All Translations	3935	20.79	36.98	36.32	5.92	42.24

## 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 Mathematics Tests. These include the scale score means, standard deviations, and performance level distributions for each grade's public and charter school population. The longitudinal results are presented in Table 41.

**Table 41. Mathematics 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	197306	688.36	34.39	2.26	7.80	63.6	26.34	89.94
	2007	200071	684.93	36.64	4.09	10.61	55.97	29.33	85.30
	2006	201908	677.49	37.75	6.35	13.13	55.42	25.11	80.52
4	2008	198509	683.13	38.11	4.70	11.37	54.49	29.45	83.93
	2007	199181	679.91	39.85	6.02	13.97	52.52	27.49	80.01
	2006	202695	676.55	40.81	7.41	14.59	52.12	25.88	78.00
5	2008	199474	679.65	36.38	3.77	12.93	56.27	27.04	83.31
	2007	203670	673.69	37.93	5.78	18.01	54.10	22.11	76.20
	2006	209200	665.59	39.85	10.29	21.24	49.31	19.16	68.47
6	2008	201719	674.85	38.21	5.45	15.04	53.21	26.31	79.52
	2007	205976	667.96	40.34	8.71	19.94	51.33	20.02	71.35
	2006	211376	655.94	40.44	13.32	26.23	47.26	13.19	60.45
7	2008	208694	674.60	38.30	3.82	17.15	51.25	27.77	79.02
	2007	213165	662.84	38.16	7.46	26.06	48.13	18.35	66.48
	2006	217225	651.08	40.55	13.19	31.12	43.52	12.17	55.69
8	2008	210265	666.44	38.19	7.31	22.69	53.10	16.89	69.99
	2007	215108	656.93	38.62	12.21	28.90	46.97	11.92	58.89
	2006	219294	651.55	41.15	14.98	31.09	43.74	10.18	53.93

As seen in Table 41 an increase in scale score means was observed for all grades between 2006 and 2007 and again between 2007 and 2008 test administrations. Least gain was observed for Grades 3 and 4 for which total gain was 11 and 6 scale score points, respectively, between 2006 and 2008 test administrations. A total gain of approximately 15 scale score points in the first three years of Mathematics Tests administration was observed for Grades 5 and 8. The most gain in scale score points between 2006 and 2008 test administrations was noted for Grades 6 and 7 (19 and 24 scale score points, respectively).

The variability of scale score distribution was fairly uniform across all grades and years. The scale score standard deviation was around 40 scale score points for Grades 4-8 in the first test administration year and decreased by approximately 2 to 3 scale score points in subsequent

years. The scale score standard deviation was around 38 scale score points for Grade 3 in 2006 and decreased to approximately 34 in administration year 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 each year for each grade but the magnitude of this increase varied depending on the grade level. An increase of 5% of students classified in Levels III and IV was observed for Grade 3 between administration years 2006 and 2007 and again between years 2007 and 2008 resulting in a total increase of 10% percent of students classified in Levels III and IV between administration years 2006 and 2008. An increase of 2% of students classified in Levels III and IV between administration years 2006 and 2007 and 4% between administration years 2007 and 2008 was noted for Grade 4. The total increase in the percentage of students classified in proficiency Levels III and IV for Grade 4 between administration years 2006 and 2008 was approximately 6% (from 78% to 84%). Larger gains in the percentage of students classified in proficiency Levels III and IV between administration years 2006 and 2008 were observed for Grades 5, 6, 7, and 8. Grade 5 proficiency score trend indicates relatively steady increase in the percentage of students classified in Levels III and IV between years 2006 and 2008 with 8% increase between years 2006 and 2007, and approximately 7% increase between years 2007 and 2008. Overall, the percentage of Grade 5 students classified in Levels III and IV increased from approximately 68% to 83% between years 2006 and 2008. Grade 6 trend shows approximately 11% increase in the percentage of students classified in Levels III and IV between years 2006 and 2007 and about 8% increase between years 2007 and 2008. Overall, the percentage of Grade 6 students classified in Levels III and IV increased by about 19% from 60.5% to 80% between years 2006 and 2008. Grade 7 proficiency score trend shows most gain in the three years of Mathematics Tests administration. It was observed that the percentage of students classified in Levels III and IV increased by approximately 10% between years 2006 and 2007 and by approximately 13% between administration year 2007 and 2008. Overall, the percentage of Grade 7 students classified in Levels III and IV increased by approximately 23% from 56% to 79% between years 2006 and 2008. Grade 8 trend shows approximately 5% increase in the percentage of students classified in Levels III and IV between years 2006 and 2007 and about 11% increase between years 2007 and 2008. Overall, the percentage of Grade 8 students classified in Levels III and IV increased by about 16% from approximately 54% to 70% between years 2006 and 2008.

In summary, an increase in the mean scale score and the percentage of students classified in Levels III and IV was observed in the second and third year of Mathematics Tests administration for all grade levels. These changes were not uniform across grades. Least gain was observed for Grades 3 and 4 while largest increase was noted for Grades 6 and 7. 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—Criteria for Item Acceptability

---

### For Multiple-Choice Items:

#### Check that the content of each item

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

#### Check that the format of each item

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

#### Also check that

- one item does not present clues to the correct answer choice for any other item
- any item based on a passage is answerable from the information given in the passage and is not dependent on skills related to other content areas
- any item based on a passage is truly passage-dependent; that is, not answerable without reference to the passage
- there is a balance of reasonable, non-stereotypic representation of economic classes, races, cultures, ages, genders, and persons with disabilities in context and art

## **For Constructed-Response Items:**

### **Check that the content of each item is**

- designed to assess the targeted performance indicator
- appropriate for the grade level being tested
- presented at a reading level suitable for the grade level being tested
- appropriate in context
- written so that a student possessing knowledge or skill being tested can construct a response that is scorable with the specified rubric or scoring tool; that is, the range of possible correct responses must be wide enough to allow for diversity of responses, but narrow enough so that students who do not clearly show their grasp of the objective or skill being assessed cannot obtain the maximum score
- presented without clue to the correct response
- checked for accuracy and documented against reliable, up-to-date sources (including rubrics)
- free of cultural, racial, ethnic, age, gender, disability, or other apparent bias

### **Check that the format of each item is**

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

### **Also check that**

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

## Appendix B—Psychometric Guidelines for Operational Item Selection

---

It is primarily up to the content development department to select items for the 2008 OP test. Research staff will provide support, as necessary, and will review the final item selection. Research staff will provide data files with parameters for all FT items eligible for 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 percent of MC and CR items on the test. An often used criterion for objective coverage is within 5% difference the of score point percentage per objective.
- Avoid selecting poor-fitting items, items with too high/low p-values, items with flagged point biserials (the research department will provide a list of such items).
- Avoid items flagged for local dependency.
- Minimize the number of items flagged for DIF (gender, ethnicity, and High/Low Needs schools). Flagged items should be reviewed for content again. It needs to be remembered that some items may be flagged for DIF by chance only and their content may not necessarily be biased against any of the analyzed groups. Research will provide DIF information for each item. It is also possible to get “significant” DIF, yet not bias if the content is a necessary part of the construct that is measured. That is, some items may be flagged for DIF not out of chance and still not represent bias.
- Verify that the items will be administered in the same relative positions in both the FT and OP forms (e.g., the first item in a FT form should also be the first item in an OP form). When that is impossible, please ensure that they are in the same one-third section of the forms.
- Evaluate the alignment of TCCs and SE curves of the proposed 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 your MC items cover a wide range of the scale.
- Provide research 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 C—Factor Analysis Results

As described in Section III, “Validity,” a principal component factor analysis was conducted on the Grades 3–8 Mathematics 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 C1 contains a table of eigenvalues and proportion of variance accounted for by extracted factors for these subgroups.

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

Grade	Subgroup	Initial Eigenvalues			
		Component	Total	% of Variance	Cumulative %
3	LEP	<b>1</b>	<b>7.57</b>	<b>24.42</b>	<b>24.42</b>
		2	1.41	4.56	28.98
		3	1.07	3.46	32.44
		4	1.05	3.39	35.83
	SWD	<b>1</b>	<b>8.59</b>	<b>27.72</b>	<b>27.72</b>
		2	1.29	4.15	31.87
		3	1.02	3.30	35.18
	SUA	<b>1</b>	<b>8.29</b>	<b>26.74</b>	<b>26.74</b>
		2	1.33	4.30	31.04
3		1.03	3.31	34.35	
4	LEP	<b>1</b>	<b>12.11</b>	<b>25.22</b>	<b>25.22</b>
		2	1.71	3.55	28.77
		3	1.16	2.42	31.19
		4	1.11	2.31	33.50
		5	1.04	2.17	35.68
		6	1.03	2.14	37.81
	SWD	<b>1</b>	<b>12.89</b>	<b>26.84</b>	<b>26.84</b>
		2	1.66	3.46	30.30
		3	1.19	2.47	32.78
		4	1.11	2.32	35.10
		5	1.05	2.18	37.28
	SUA	<b>1</b>	<b>12.66</b>	<b>26.37</b>	<b>26.37</b>
		2	1.65	3.43	29.81
		3	1.18	2.47	32.27
		4	1.12	2.32	34.59
5		1.05	2.18	36.78	
5	LEP	<b>1</b>	<b>8.12</b>	<b>23.87</b>	<b>23.87</b>
		2	1.33	3.90	27.77
		3	1.10	3.22	30.99
		4	1.05	3.10	34.09
		5	1.02	2.99	37.08

*(Continued on next page)*

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

Grade	Subgroup	Initial Eigenvalues			
		Component	Total	% of Variance	Cumulative %
5	SWD	<b>1</b>	<b>7.97</b>	<b>23.44</b>	<b>23.44</b>
		2	1.42	4.19	27.62
		3	1.11	3.26	30.88
		4	1.05	3.08	33.96
	SUA	<b>1</b>	<b>8.10</b>	<b>23.84</b>	<b>23.84</b>
		2	1.40	4.10	27.94
		3	1.10	3.24	31.18
		4	1.04	3.06	34.24
		5	1.00	2.95	37.19
	6	LEP	<b>1</b>	<b>9.36</b>	<b>26.76</b>
2			1.41	4.01	30.77
3			1.09	3.13	33.89
4			1.01	2.89	36.78
SWD		<b>1</b>	<b>9.29</b>	<b>26.56</b>	<b>26.56</b>
		2	1.34	3.82	30.37
		3	1.12	3.21	33.58
		4	1.01	2.89	36.48
SUA		<b>1</b>	<b>9.54</b>	<b>27.25</b>	<b>27.25</b>
		2	1.34	3.84	31.09
		3	1.10	3.15	34.23
		4	1.01	2.87	37.11
7	LEP	<b>1</b>	<b>7.54</b>	<b>19.85</b>	<b>19.85</b>
		2	1.33	3.49	23.34
		3	1.26	3.31	26.65
		4	1.09	2.87	29.52
		5	1.02	2.69	32.21
		6	1.00	2.64	34.84
	SWD	<b>1</b>	<b>7.57</b>	<b>19.91</b>	<b>19.91</b>
		2	1.37	3.59	23.51
		3	1.20	3.15	26.65
		4	1.05	2.77	29.42
		5	1.02	2.69	32.11
	SUA	<b>1</b>	<b>7.74</b>	<b>20.36</b>	<b>20.36</b>
		2	1.35	3.54	23.90
		3	1.21	3.19	27.09
		4	1.08	2.83	29.92
		5	1.00	2.64	32.55

(Continued on next page)

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

Grade	Subgroup	Initial Eigenvalues			
		Component	Total	% of Variance	Cumulative %
8	LEP	<b>1</b>	<b>12.82</b>	<b>28.48</b>	<b>28.48</b>
		2	1.49	3.32	31.80
		3	1.27	2.83	34.62
		4	1.20	2.67	37.30
		5	1.05	2.34	39.63
		6	1.02	2.27	41.90
	SWD	<b>1</b>	<b>11.61</b>	<b>25.81</b>	<b>25.81</b>
		2	1.62	3.61	29.42
		3	1.30	2.89	32.31
		4	1.19	2.64	34.95
		5	1.03	2.30	37.24
	SUA	<b>1</b>	<b>12.37</b>	<b>27.48</b>	<b>27.48</b>
		2	1.59	3.53	31.01
		3	1.29	2.88	33.89
		4	1.18	2.62	36.50
5		1.01	2.25	38.75	

## Appendix D—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 D1 shows items flagged by the SMD and Mantel-Haenszel methods, and Table D2 presents items flagged by the Linn-Harnisch method. Note that positive values of SMD and Delta in Table D1 indicate DIF in favor of a focal group and negative values of SMD and Delta indicate DIF against a focal group.

**Table D1. NYSTP Mathematics 2008 Classical DIF Item Flags**

Grade	Item #	Subgroup	DIF	SMD	Mantel-Haenszel	Delta
3	11	Spanish	Against	-0.11	No flag	No flag
3	13	Asian	Favor	No Flag	196.45	1.52
3	16	Spanish	Against	-0.11	No flag	No flag
3	25	Black	Against	-0.10	1709.80	-1.66
3	25	Hispanic	Against	-0.13	3232.93	-2.19
3	25	Asian	Against	No Flag	552.75	-1.70
3	25	High needs	Against	No Flag	2129.24	-1.54
4	13	Spanish	Against	-0.10	No flag	No flag
4	16	Hispanic	Against	-0.11	No flag	No flag
4	16	Spanish	Against	-0.10	No flag	No flag
4	24	Female	Against	-0.11	No flag	No flag
4	27	Black	Against	-0.10	No flag	No flag
4	27	Hispanic	Against	-0.10	No flag	No Flag
4	27	Asian	Against	No flag	721.79	-1.80
4	27	Spanish	Against	-0.10	No flag	No flag
4	34	Black	Against	-0.15	No flag	No flag
4	34	Hispanic	Against	-0.13	No flag	No flag
4	34	Asian	Against	-0.10	No flag	No flag
4	34	Female	Against	-0.10	No flag	No flag
4	34	High needs	Against	-0.10	No flag	No flag
4	34	Spanish	Against	-0.11	No flag	No flag
4	38	Black	Favor	0.12	No flag	No flag
4	39	Female	Favor	0.11	No flag	No flag
4	40	Spanish	Favor	0.10	No flag	No flag
4	42	Spanish	Favor	0.10	No flag	No flag
4	44	Spanish	Favor	0.10	No flag	No flag
4	45	Black	Favor	0.11	No flag	No flag
5	28	Black	Favor	0.13	n/a	n/a

*(Continued on next page)*

**Table D1. NYSTP Mathematics 2008 Classical DIF Item Flags (cont.)**

Grade	Item #	Subgroup	DIF	SMD	Mantel-Haenszel	Delta
5	28	Hispanic	Favor	0.10	n/a	n/a
5	29	Black	Against	-0.11	n/a	n/a
5	29	Spanish	Favor	0.11	n/a	n/a
5	32	Black	Against	-0.19	n/a	n/a
5	32	Hispanic	Against	-0.22	n/a	n/a
5	32	Asian	Against	-0.18	n/a	n/a
5	32	High needs	Against	-0.16	n/a	n/a
5	33	Spanish	Against	-0.13	n/a	n/a
5	34	Black	Favor	0.10	n/a	n/a
5	34	Hispanic	Favor	0.13	n/a	n/a
5	34	High needs	Favor	0.10	n/a	n/a
5	34	Spanish	Favor	0.17	n/a	n/a
6	8	Spanish	Against	-0.16	357.37	-1.91
6	25	Spanish	Against	-0.10	No flag	No flag
6	26	Spanish	Favor	0.10	No flag	No flag
6	28	Hispanic	Favor	0.10	No flag	No flag
6	29	Hispanic	Against	-0.10	n/a	n/a
6	29	Asian	Against	-0.13	n/a	n/a
6	30	Spanish	Favor	0.15	n/a	n/a
6	33	Spanish	Against	-0.17	n/a	n/a
7	10	Hispanic	Against	No flag	1540.12	-1.52
7	10	Asian	Against	No flag	839.96	-1.93
7	10	Spanish	Against	-0.18	460.92	-2.09
7	18	Spanish	Against	-0.13	262.90	-1.64
7	22	Spanish	Against	-0.11	No flag	No flag
7	25	Female	Against	-0.11	No flag	No flag
7	29	Asian	Favor	No flag	222.39	1.71
7	31	Black	Against	-0.11	n/a	n/a
7	31	Hispanic	Against	-0.16	n/a	n/a
7	31	Female	Against	-0.12	n/a	n/a
7	31	Spanish	Against	-0.24	n/a	n/a
7	33	Hispanic	Favor	0.14	n/a	n/a
7	33	Female	Favor	0.10	n/a	n/a
7	33	Spanish	Favor	1.47	n/a	n/a
7	34	Asian	Against	-0.18	n/a	n/a
7	34	Female	Favor	0.11	n/a	n/a
7	34	High needs	Against	-0.15	n/a	n/a
7	34	Spanish	Against	-0.10	n/a	n/a
7	35	Black	Against	-0.20	n/a	n/a

*(Continued on next page)*

**Table D1. NYSTP Mathematics 2008 Classical DIF Item Flags (cont.)**

Grade	Item #	Subgroup	DIF	SMD	Mantel-Haenszel	Delta
7	35	Hispanic	Against	-0.19	n/a	n/a
7	35	Asian	Against	-0.19	n/a	n/a
7	35	High needs	Against	-0.22	n/a	n/a
7	36	Black	Favor	0.16	n/a	n/a
7	36	Hispanic	Favor	0.16	n/a	n/a
7	36	Asian	Favor	0.11	n/a	n/a
7	36	Female	Favor	0.14	n/a	n/a
7	36	High needs	Favor	0.10	n/a	n/a
7	38	Black	Against	-0.38	n/a	n/a
7	38	Hispanic	Against	-0.38	n/a	n/a
7	38	Asian	Against	-0.22	n/a	n/a
7	38	High needs	Against	-0.27	n/a	n/a
7	38	Spanish	Against	-0.58	n/a	n/a
8	5	Spanish	Favor	0.10	No flag	No flag
8	8	Spanish	Against	-0.16	328.67	-1.65
8	10	Female	Against	-0.11	3404.68	-1.51
8	21	Spanish	Against	-0.10	No flag	No flag
8	24	Black	Favor	0.10	No flag	No flag
8	30	Female	Favor	0.13	n/a	n/a
8	30	Spanish	Favor	0.10	n/a	n/a
8	31	Spanish	Favor	0.14	n/a	n/a
8	35	Spanish	Favor	0.10	n/a	n/a
8	36	Hispanic	Against	-0.12	n/a	n/a
8	36	High needs	Against	-0.12	n/a	n/a
8	39	Black	Against	-0.10	n/a	n/a
8	39	Female	Against	-0.16	n/a	n/a
8	39	Spanish	Against	-0.11	n/a	n/a
8	40	Black	Against	-0.13	n/a	n/a
8	40	Hispanic	Against	-0.14	n/a	n/a
8	40	Asian	Against	-0.10	n/a	n/a
8	40	High needs	Against	-0.11	n/a	n/a
8	41	Black	Against	-0.14	n/a	n/a
8	42	Female	Favor	0.11	n/a	n/a
8	43	Black	Favor	0.15	n/a	n/a
8	43	Hispanic	Favor	0.12	n/a	n/a
8	43	High needs	Favor	0.14	n/a	n/a
8	44	Spanish	Favor	0.10	n/a	n/a

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

Grade	Item	Focal Group	Direction	Magnitude
3	11	Spanish	Against	-0.103
3	16	Spanish	Against	-0.112
4	13	Spanish	Against	-0.105
4	16	Spanish	Against	-0.100
4	34	Spanish	Against	-0.118
4	40	Spanish	In Favor	0.105
5	29	Spanish	In Favor	0.119
5	32	Spanish	Against	-0.110
5	32	Asian	Against	-0.126
5	32	Black	Against	-0.108
5	32	Hispanic	Against	-0.143
5	34	Spanish	In Favor	0.157
6	8	Spanish	Against	-0.151
6	26	Spanish	In Favor	0.117
6	29	Asian	Against	-0.103
6	30	Spanish	In Favor	0.132
6	33	Spanish	Against	-0.169
7	10	Spanish	Against	-0.181
7	18	Spanish	Against	-0.130
7	22	Spanish	Against	-0.121
7	31	Spanish	Against	-0.270
7	33	Spanish	In Favor	1.404
7	34	Asian	Against	-0.162
7	34	Spanish	Against	-0.148
7	35	Asian	Against	-0.149
7	35	Black	Against	-0.113
7	35	High needs	Against	-0.109
7	35	Hispanic	Against	-0.100
7	35	Low needs	In Favor	0.116
7	35	Spanish	Against	-0.103
7	38	Asian	Against	-0.124
7	38	Black	Against	-0.104
7	38	Hispanic	Against	-0.179
7	38	Low needs	In Favor	0.123
7	38	Spanish	Against	-0.606
7	38	White	In Favor	0.138
8	8	Spanish	Against	-0.152
8	31	Spanish	In Favor	0.104
8	39	Spanish	Against	-0.128

## Appendix E—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, N-count, obtained-Z fit statistic, and critical-Z fit statistic are presented for each item. Fit for most items in the Grades 3–8 Mathematics Tests was acceptable (critical  $Z >$  obtained  $Z$ ).

**Table E1. Mathematics Grade 3 Item Fit Statistics**

Item	Model	Chi-Square	DF	N-count	Obtained $Z$	Critical $Z$	Fit OK?
1	3PL	47.77	7	181526	10.90	484.07	Y
2	3PL	174.75	7	181526	44.83	484.07	Y
3	3PL	119.80	7	181526	30.15	484.07	Y
4	3PL	83.50	7	181526	20.44	484.07	Y
5	3PL	54.32	7	181526	12.65	484.07	Y
6	3PL	364.83	7	181526	95.63	484.07	Y
7	3PL	60.64	7	181526	14.34	484.07	Y
8	3PL	188.61	7	181526	48.54	484.07	Y
9	3PL	63.99	7	181526	15.23	484.07	Y
10	3PL	71.26	7	181526	17.17	484.07	Y
11	3PL	253.73	7	181526	65.94	484.07	Y
12	3PL	175.10	7	181526	44.93	484.07	Y
13	3PL	127.81	7	181526	32.29	484.07	Y
14	3PL	84.73	7	181526	20.77	484.07	Y
15	3PL	269.74	7	181526	70.22	484.07	Y
16	3PL	42.28	7	181526	9.43	484.07	Y
17	3PL	284.14	7	181526	74.07	484.07	Y
18	3PL	497.47	7	181526	131.08	484.07	Y
19	3PL	167.63	7	181526	42.93	484.07	Y
20	3PL	116.47	7	181526	29.26	484.07	Y
21	3PL	83.42	7	181526	20.42	484.07	Y
22	3PL	36.69	7	181526	7.93	484.07	Y
23	3PL	29.37	7	181526	5.98	484.07	Y
24	3PL	46.78	7	181526	10.63	484.07	Y
25	3PL	437.47	7	181526	115.05	484.07	Y
26	2PPC	490.72	17	181526	81.24	484.07	Y
27	2PPC	930.51	17	181526	156.66	484.07	Y
28	2PPC	5962.59	17	181526	1019.66	484.07	N
29	2PPC	400.50	17	181526	65.77	484.07	Y
30	2PPC	3713.15	26	181526	511.32	484.07	N
31	2PPC	1319.39	26	181526	179.36	484.07	Y

**Table E2. Mathematics Grade 4 Item Fit Statistics**

Item	Model	Chi-Square	DF	N-count	Obtained Z	Critical Z	Fit OK?
1	3PL	126.71	7	192025	31.99	512.07	Y
2	3PL	68.63	7	192025	16.47	512.07	Y
3	3PL	23.47	7	192025	4.40	512.07	Y
4	3PL	129.09	7	192025	32.63	512.07	Y
5	3PL	312.57	7	192025	81.67	512.07	Y
6	3PL	551.85	7	192025	145.62	512.07	Y
7	3PL	211.38	7	192025	54.62	512.07	Y
8	3PL	52.87	7	192025	12.26	512.07	Y
9	3PL	28.39	7	192025	5.72	512.07	Y
10	3PL	161.48	7	192025	41.29	512.07	Y
11	3PL	25.07	7	192025	4.83	512.07	Y
12	3PL	1884.35	7	192025	501.74	512.07	Y
13	3PL	401.58	7	192025	105.46	512.07	Y
14	3PL	212.91	7	192025	55.03	512.07	Y
15	3PL	30.65	7	192025	6.32	512.07	Y
16	3PL	41.79	7	192025	9.30	512.07	Y
17	3PL	76.39	7	192025	18.55	512.07	Y
18	3PL	29.94	7	192025	6.13	512.07	Y
19	3PL	260.28	7	192025	67.69	512.07	Y
20	3PL	225.39	7	192025	58.37	512.07	Y
21	3PL	142.47	7	192025	36.21	512.07	Y
22	3PL	47.48	7	192025	10.82	512.07	Y
23	3PL	406.82	7	192025	106.86	512.07	Y
24	3PL	173.26	7	192025	44.44	512.07	Y
25	3PL	171.17	7	192025	43.87	512.07	Y
26	3PL	501.09	7	192025	132.05	512.07	Y
27	3PL	925.29	7	192025	245.42	512.07	Y
28	3PL	546.32	7	192025	144.14	512.07	Y
29	3PL	59.74	7	192025	14.10	512.07	Y
30	3PL	380.63	7	192025	99.86	512.07	Y
31	2PPC	342.62	17	192025	55.84	512.07	Y
32	2PPC	396.10	17	192025	65.02	512.07	Y
33	2PPC	1277.24	17	192025	216.13	512.07	Y
34	2PPC	501.59	17	192025	83.11	512.07	Y
35	2PPC	558.24	26	192025	73.81	512.07	Y
36	2PPC	770.64	17	192025	129.25	512.07	Y
37	2PPC	316.46	17	192025	51.36	512.07	Y
38	2PPC	2763.63	17	192025	471.04	512.07	Y
39	2PPC	539.89	26	192025	71.26	512.07	Y
40	2PPC	1093.85	17	192025	184.68	512.07	Y
41	2PPC	347.81	17	192025	56.73	512.07	Y
42	2PPC	677.20	17	192025	113.22	512.07	Y

*(Continued on next page)*

**Table E2. Mathematics Grade 4 Item Fit Statistics (cont.)**

Item	Model	Chi-Square	DF	N-count	Obtained Z	Critical Z	Fit OK?
43	2PPC	675.21	26	192025	90.03	512.07	Y
44	2PPC	456.87	26	192025	59.75	512.07	Y
45	2PPC	293.46	17	192025	47.41	512.07	Y
46	2PPC	380.57	17	192025	62.35	512.07	Y
47	2PPC	560.51	17	192025	93.21	512.07	Y
48	2PPC	419.32	17	192025	69.00	512.07	Y

**Table E3. Mathematics Grade 5 Item Fit Statistics**

Item	Model	Chi-Square	DF	N-count	Obtained Z	Critical Z	Fit OK?
1	3PL	52.90	7	191647	12.27	511.06	Y
2	3PL	166.20	7	191647	42.55	511.06	Y
3	3PL	177.88	7	191647	45.67	511.06	Y
4	3PL	1213.67	7	191647	322.50	511.06	Y
5	3PL	663.19	7	191647	175.37	511.06	Y
6	3PL	391.09	7	191647	102.65	511.06	Y
7	3PL	436.92	7	191647	114.90	511.06	Y
8	3PL	461.83	7	191647	121.56	511.06	Y
9	3PL	359.33	7	191647	94.16	511.06	Y
10	3PL	48.91	7	191647	11.20	511.06	Y
11	3PL	191.49	7	191647	49.31	511.06	Y
12	3PL	83.94	7	191647	20.56	511.06	Y
13	3PL	648.71	7	191647	171.50	511.06	Y
14	3PL	246.85	7	191647	64.10	511.06	Y
15	3PL	584.87	7	191647	154.44	511.06	Y
16	3PL	151.87	7	191647	38.72	511.06	Y
17	3PL	138.09	7	191647	35.03	511.06	Y
18	3PL	239.56	7	191647	62.16	511.06	Y
19	3PL	222.28	7	191647	57.54	511.06	Y
20	3PL	331.53	7	191647	86.74	511.06	Y
21	3PL	42.20	7	191647	9.41	511.06	Y
22	3PL	242.22	7	191647	62.87	511.06	Y
23	3PL	148.83	7	191647	37.91	511.06	Y
24	3PL	347.73	7	191647	91.06	511.06	Y
25	3PL	102.92	7	191647	25.63	511.06	Y
26	3PL	141.02	7	191647	35.82	511.06	Y
27	2PPC	5938.69	17	191647	1015.56	511.06	N
28	2PPC	1418.52	26	191647	193.11	511.06	Y
29	2PPC	291.17	17	191647	47.02	511.06	Y
30	2PPC	1103.95	26	191647	149.49	511.06	Y
31	2PPC	3065.02	17	191647	522.73	511.06	N
32	2PPC	569.80	26	191647	75.41	511.06	Y
33	2PPC	661.14	17	191647	110.47	511.06	Y
34	2PPC	1901.29	26	191647	260.06	511.06	Y

**Table E4. Mathematics Grade 6 Item Fit Statistics**

Item	Model	Chi-Square	DF	N-count	Obtained Z	Critical Z	Fit OK?
1	3PL	32.94	7	195467	6.93	521.25	Y
2	3PL	262.59	7	195467	68.31	521.25	Y
3	3PL	87.86	7	195467	21.61	521.25	Y
4	3PL	116.87	7	195467	29.36	521.25	Y
5	3PL	64.16	7	195467	15.28	521.25	Y
6	3PL	172.79	7	195467	44.31	521.25	Y
7	3PL	43.97	7	195467	9.88	521.25	Y
8	3PL	16.33	7	195467	2.49	521.25	Y
9	3PL	279.30	7	195467	72.78	521.25	Y
10	3PL	232.22	7	195467	60.19	521.25	Y
11	3PL	202.11	7	195467	52.15	521.25	Y
12	3PL	32.88	7	195467	6.92	521.25	Y
13	3PL	1531.20	7	195467	407.36	521.25	Y
14	3PL	1168.76	7	195467	310.49	521.25	Y
15	3PL	538.02	7	195467	141.92	521.25	Y
16	3PL	124.04	7	195467	31.28	521.25	Y
17	3PL	27.62	7	195467	5.51	521.25	Y
18	3PL	208.02	7	195467	53.73	521.25	Y
19	3PL	127.33	7	195467	32.16	521.25	Y
20	3PL	85.54	7	195467	20.99	521.25	Y
21	3PL	155.75	7	195467	39.76	521.25	Y
22	3PL	73.89	7	195467	17.88	521.25	Y
23	3PL	273.20	7	195467	71.14	521.25	Y
24	3PL	44.01	7	195467	9.89	521.25	Y
25	3PL	227.16	7	195467	58.84	521.25	Y
26	2PPC	656.71	17	195467	109.71	521.25	Y
27	2PPC	4602.75	26	195467	634.68	521.25	N
28	2PPC	982.18	26	195467	132.60	521.25	Y
29	2PPC	600.98	17	195467	100.15	521.25	Y
30	2PPC	1411.61	26	195467	192.15	521.25	Y
31	2PPC	561.91	17	195467	93.45	521.25	Y
32	2PPC	668.60	17	195467	111.75	521.25	Y
33	2PPC	724.24	26	195467	96.83	521.25	Y
34	2PPC	2034.79	17	195467	346.05	521.25	Y
35	2PPC	1823.81	17	195467	309.87	521.25	Y

**Table E5. Mathematics Grade 7 Item Fit Statistics**

Item	Model	Chi Square	DF	Total N	Obtained Z	Critical Z	Fit OK?
1	3PL	20.84	7	201713	3.70	537.90	Y
2	3PL	677.11	7	201713	179.09	537.90	Y
3	3PL	575.19	7	201713	151.86	537.90	Y
4	3PL	673.11	7	201713	178.03	537.90	Y
5	3PL	66.11	7	201713	15.80	537.90	Y
6	3PL	391.48	7	201713	102.76	537.90	Y
7	3PL	450.21	7	201713	118.45	537.90	Y
8	3PL	100.56	7	201713	25.01	537.90	Y
9	3PL	917.66	7	201713	243.38	537.90	Y
10	3PL	15.74	7	201713	2.34	537.90	Y
11	3PL	360.72	7	201713	94.54	537.90	Y
12	3PL	644.65	7	201713	170.42	537.90	Y
13	3PL	55.27	7	201713	12.90	537.90	Y
14	3PL	134.95	7	201713	34.20	537.90	Y
15	3PL	448.48	7	201713	117.99	537.90	Y
16	3PL	66.46	7	201713	15.89	537.90	Y
17	3PL	158.02	7	201713	40.36	537.90	Y
18	3PL	282.32	7	201713	73.58	537.90	Y
19	3PL	47.71	7	201713	10.88	537.90	Y
20	3PL	502.42	7	201713	132.41	537.90	Y
21	3PL	131.96	7	201713	33.40	537.90	Y
22	3PL	832.06	7	201713	220.51	537.90	Y
23	3PL	165.50	7	201713	42.36	537.90	Y
24	3PL	277.50	7	201713	72.29	537.90	Y
25	3PL	135.16	7	201713	34.25	537.90	Y
26	3PL	385.10	7	201713	101.05	537.90	Y
27	3PL	426.88	7	201713	112.22	537.90	Y
28	3PL	125.09	7	201713	31.56	537.90	Y
29	3PL	118.81	7	201713	29.88	537.90	Y
30	3PL	28.25	7	201713	5.68	537.90	Y
31	2PPC	1039.15	26	201713	140.50	537.90	Y
32	2PPC	209.92	17	201713	33.09	537.90	Y
33	2PPC	1853.23	17	201713	314.91	537.90	Y
34	2PPC	1622.95	26	201713	221.46	537.90	Y
35	2PPC	2296.31	26	201713	314.84	537.90	Y
36	2PPC	481.15	17	201713	79.60	537.90	Y
37	2PPC	460.58	17	201713	76.07	537.90	Y
38	2PPC	1749.41	26	201713	238.99	537.90	Y

**Table E6. Mathematics Grade 8 Item Fit Statistics**

Item	Model	Chi Square	DF	Total N	Obtained Z	Critical Z	Fit OK?
1	3PL	167.71	7	204333	42.95	544.89	Y
2	3PL	45.34	7	204333	10.25	544.89	Y
3	3PL	1608.61	7	204333	428.05	544.89	Y
4	3PL	1229.52	7	204333	326.73	544.89	Y
5	3PL	115.47	7	204333	28.99	544.89	Y
6	3PL	903.88	7	204333	239.70	544.89	Y
7	3PL	123.88	7	204333	31.24	544.89	Y
8	3PL	455.96	7	204333	119.99	544.89	Y
9	3PL	229.03	7	204333	59.34	544.89	Y
10	3PL	307.08	7	204333	80.20	544.89	Y
11	3PL	190.49	7	204333	49.04	544.89	Y
12	3PL	667.27	7	204333	176.46	544.89	Y
13	3PL	206.15	7	204333	53.22	544.89	Y
14	3PL	463.92	7	204333	122.12	544.89	Y
15	3PL	177.88	7	204333	45.67	544.89	Y
16	3PL	400.12	7	204333	105.07	544.89	Y
17	3PL	157.69	7	204333	40.27	544.89	Y
18	3PL	685.03	7	204333	181.21	544.89	Y
19	3PL	138.20	7	204333	35.06	544.89	Y
20	3PL	829.82	7	204333	219.91	544.89	Y
21	3PL	88.07	7	204333	21.67	544.89	Y
22	3PL	212.48	7	204333	54.92	544.89	Y
23	3PL	279.02	7	204333	72.70	544.89	Y
24	3PL	83.96	7	204333	20.57	544.89	Y
25	3PL	129.74	7	204333	32.80	544.89	Y
26	3PL	91.52	7	204333	22.59	544.89	Y
27	3PL	154.75	7	204333	39.49	544.89	Y
28	2PPC	268.76	26	204333	33.66	544.89	Y
29	2PPC	501.17	17	204333	83.03	544.89	Y
30	2PPC	2119.18	17	204333	360.52	544.89	Y
31	2PPC	3648.93	26	204333	502.41	544.89	Y
32	2PPC	440.99	17	204333	72.71	544.89	Y
33	2PPC	2602.63	17	204333	443.43	544.89	Y
34	2PPC	164.14	17	204333	25.23	544.89	Y
35	2PPC	515.20	17	204333	85.44	544.89	Y
36	2PPC	320.98	17	204333	52.13	544.89	Y
37	2PPC	493.68	17	204333	81.75	544.89	Y
38	2PPC	116.31	17	204333	17.03	544.89	Y
39	2PPC	158.30	17	204333	24.23	544.89	Y
40	2PPC	1064.27	26	204333	143.98	544.89	Y
41	2PPC	355.31	26	204333	45.67	544.89	Y
42	2PPC	3558.66	26	204333	489.89	544.89	Y
43	2PPC	4805.48	26	204333	662.79	544.89	N
44	2PPC	520.83	17	204333	86.41	544.89	Y
45	2PPC	254.92	17	204333	40.80	544.89	Y

## **Appendix F—Derivation of the Generalized SPI Procedure**

---

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

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

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

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

for  $0 \leq T_j \leq 1$ ;  $r_j, s_j > 0$ . 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 the maximum score for a standard, will be greater or equal to zero. If correct guessing is possible, as it is with MC items, there will be a non-zero lower limit to  $\hat{T}_j$ , and a three-parameter beta distribution, in which  $\hat{T}_j$  is greater than or equal to this lower limit (Johnson and Kotz, 1979, p. 37) would be more appropriate. The use of the two-parameter beta distribution would tend to underestimate  $T_j$  among very low-performing examinees. While working with tests containing exclusively MC items, Yen found that there does not appear to be a practical importance to this underestimation (Yen, 1987). The impact of any such effect would be reduced as the proportion of CR items in the test increases. The size of this effect, nonetheless, was evaluated using simulations (Yen, Sykes, Ito, and Julian 1997).

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

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

## Appendix G—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 H—Scale Score Frequency Distributions

---

Tables H1–H6 depict the scale score (SS) distributions, by frequency (N-count), percent, cumulative frequency, and cumulative percent for each grade (total population of students from public and charter schools).

**Table H1. Grade 3 Mathematics 2008 SS Frequency Distribution, State**

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
470	222	0.11	222	0.11
548	202	0.10	424	0.21
578	294	0.15	718	0.36
591	397	0.20	1115	0.57
600	486	0.25	1601	0.81
606	597	0.30	2198	1.11
612	707	0.36	2905	1.47
617	712	0.36	3617	1.83
621	835	0.42	4452	2.26
624	905	0.46	5357	2.72
628	955	0.48	6312	3.20
631	1071	0.54	7383	3.74
634	1242	0.63	8625	4.37
637	1335	0.68	9960	5.05
639	1565	0.79	11525	5.84
642	1710	0.87	13235	6.71
644	1917	0.97	15152	7.68
647	2144	1.09	17296	8.77
649	2552	1.29	19848	10.06
652	2802	1.42	22650	11.48
654	3218	1.63	25868	13.11
657	3868	1.96	29736	15.07
659	4337	2.20	34073	17.27
662	5098	2.58	39171	19.85
665	5819	2.95	44990	22.80
668	6771	3.43	51761	26.23
671	8178	4.14	59939	30.38

*(Continued on next page)*

**Table H1. Grade 3 Mathematics 2008 SS Frequency Distribution, State (cont.)**

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
674	9619	4.88	69558	35.25
678	10963	5.56	80521	40.81
682	12996	6.59	93517	47.40
687	15143	7.67	108660	55.07
693	17274	8.75	125934	63.83
700	19399	9.83	145333	73.66
710	20540	10.41	165873	84.07
728	19170	9.72	185043	93.78
770	12263	6.22	197306	100.00

**Table H2. Grade 4 Mathematics 2008 SS Frequency Distribution, State**

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
485	229	0.12	229	0.12
518	186	0.09	415	0.21
545	242	0.12	657	0.33
559	304	0.15	961	0.48
570	353	0.18	1314	0.66
578	437	0.22	1751	0.88
584	445	0.22	2196	1.11
590	475	0.24	2671	1.35
595	541	0.27	3212	1.62
599	607	0.31	3819	1.92
603	661	0.33	4480	2.26
606	667	0.34	5147	2.59
610	726	0.37	5873	2.96
612	742	0.37	6615	3.33
615	799	0.40	7414	3.73
618	915	0.46	8329	4.20
620	999	0.50	9328	4.70
623	1016	0.51	10344	5.21
625	998	0.50	11342	5.71

*(Continued on next page)*

**Table H2. Grade 4 Mathematics 2008 SS Frequency Distribution, State (cont.)**

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
627	1071	0.54	12413	6.25
629	1146	0.58	13559	6.83
631	1201	0.61	14760	7.44
633	1347	0.68	16107	8.11
635	1387	0.70	17494	8.81
637	1444	0.73	18938	9.54
638	1578	0.79	20516	10.34
640	1650	0.83	22166	11.17
642	1755	0.88	23921	12.05
644	1817	0.92	25738	12.97
645	1921	0.97	27659	13.93
647	2030	1.02	29689	14.96
649	2203	1.11	31892	16.07
650	2294	1.16	34186	17.22
652	2473	1.25	36659	18.47
654	2574	1.30	39233	19.76
655	2609	1.31	41842	21.08
657	2702	1.36	44544	22.44
659	2923	1.47	47467	23.91
660	2860	1.44	50327	25.35
662	3157	1.59	53484	26.94
664	3319	1.67	56803	28.61
665	3454	1.74	60257	30.35
667	3615	1.82	63872	32.18
669	3852	1.94	67724	34.12
671	3941	1.99	71665	36.10
673	4280	2.16	75945	38.26
674	4390	2.21	80335	40.47
676	4525	2.28	84860	42.75
678	5028	2.53	89888	45.28
681	5106	2.57	94994	47.85
683	5377	2.71	100371	50.56

*(Continued on next page)*

**Table H2. Grade 4 Mathematics 2008 SS Frequency Distribution, State (cont.)**

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
685	5803	2.92	106174	53.49
687	6140	3.09	112314	56.58
690	6418	3.23	118732	59.81
693	6728	3.39	125460	63.20
696	7163	3.61	132623	66.81
699	7429	3.74	140052	70.55
703	8016	4.04	148068	74.59
707	8100	4.08	156168	78.67
712	8533	4.30	164701	82.97
718	8603	4.33	173304	87.30
726	8344	4.20	181648	91.51
736	7393	3.72	189041	95.23
755	6097	3.07	195138	98.30
800	3371	1.70	198509	100.00

**Table H3. Grade 5 Mathematics 2008 SS Frequency Distribution, State**

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
495	248	0.12	248	0.12
522	219	0.11	467	0.23
548	334	0.17	801	0.40
565	469	0.24	1270	0.64
577	624	0.31	1894	0.95
586	750	0.38	2644	1.33
595	960	0.48	3604	1.81
602	1086	0.54	4690	2.35
608	1310	0.66	6000	3.01
614	1519	0.76	7519	3.77
619	1869	0.94	9388	4.71
623	2065	1.04	11453	5.74
628	2356	1.18	13809	6.92
631	2687	1.35	16496	8.27

*(Continued on next page)*

**Table H3. Grade 5 Mathematics 2008 SS Frequency Distribution, State (cont.)**

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
635	2914	1.46	19410	9.73
638	3075	1.54	22485	11.27
642	3318	1.66	25803	12.94
645	3559	1.78	29362	14.72
647	3939	1.97	33301	16.69
650	4081	2.05	37382	18.74
653	4251	2.13	41633	20.87
655	4727	2.37	46360	23.24
658	4853	2.43	51213	25.67
661	5192	2.60	56405	28.28
663	5463	2.74	61868	31.02
666	5715	2.87	67583	33.88
668	6080	3.05	73663	36.93
671	6358	3.19	80021	40.12
673	6694	3.36	86715	43.47
676	7062	3.54	93777	47.01
679	7566	3.79	101343	50.81
682	7834	3.93	109177	54.73
685	8315	4.17	117492	58.90
688	8754	4.39	126246	63.29
692	9324	4.67	135570	67.96
696	9976	5.00	145546	72.96
701	10423	5.23	155969	78.19
707	10555	5.29	166524	83.48
714	10513	5.27	177037	88.75
725	9828	4.93	186865	93.68
744	7953	3.99	194818	97.67
780	4656	2.33	199474	100.00

**Table H4. Grade 6 Mathematics 2008 SS Frequency Distribution, State**

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
500	1230	0.61	1230	0.61
561	911	0.45	2141	1.06
580	1184	0.59	3325	1.65
591	1354	0.67	4679	2.32
599	1457	0.72	6136	3.04
605	1571	0.78	7707	3.82
610	1589	0.79	9296	4.61
614	1694	0.84	10990	5.45
618	1851	0.92	12841	6.37
622	1935	0.96	14776	7.33
625	2018	1.00	16794	8.33
628	2154	1.07	18948	9.39
631	2353	1.17	21301	10.56
633	2371	1.18	23672	11.74
636	2592	1.28	26264	13.02
638	2732	1.35	28996	14.37
641	2768	1.37	31764	15.75
643	3018	1.50	34782	17.24
645	3153	1.56	37935	18.81
647	3386	1.68	41321	20.48
650	3667	1.82	44988	22.30
652	3887	1.93	48875	24.23
654	4153	2.06	53028	26.29
656	4391	2.18	57419	28.46
658	4635	2.30	62054	30.76
661	4910	2.43	66964	33.20
663	5179	2.57	72143	35.76
665	5523	2.74	77666	38.50
668	5753	2.85	83419	41.35
670	6230	3.09	89649	44.44
672	6447	3.20	96096	47.64
675	6494	3.22	102590	50.86
678	7050	3.49	109640	54.35
681	7164	3.55	116804	57.90

*(Continued on next page)*

**Table H4. Grade 6 Mathematics 2008 SS Frequency Distribution, State (cont.)**

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
684	7527	3.73	124331	61.64
687	7649	3.79	131980	65.43
690	8267	4.10	140247	69.53
694	8400	4.16	148647	73.69
698	8768	4.35	157415	78.04
703	9079	4.50	166494	82.54
709	9337	4.63	175831	87.17
717	8867	4.40	184698	91.56
729	8032	3.98	192730	95.54
749	5999	2.97	198729	98.52
780	2990	1.48	201719	100.00

**Table H5. Grade 7 Mathematics 2008 SS Frequency Distribution, State**

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
500	769	0.37	769	0.37
523	539	0.26	1308	0.63
561	668	0.32	1976	0.95
579	877	0.42	2853	1.37
590	1032	0.49	3885	1.86
598	1219	0.58	5104	2.45
604	1377	0.66	6481	3.11
610	1501	0.72	7982	3.82
615	1605	0.77	9587	4.59
619	1974	0.95	11561	5.54
622	2172	1.04	13733	6.58
626	2373	1.14	16106	7.72
629	2637	1.26	18743	8.98
632	2862	1.37	21605	10.35
635	3024	1.45	24629	11.80
638	3301	1.58	27930	13.38
640	3633	1.74	31563	15.12

*(Continued on next page)*

**Table H5. Grade 7 Mathematics 2008 SS Frequency Distribution, State (cont.)**

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
643	3772	1.81	35335	16.93
645	4154	1.99	39489	18.92
647	4294	2.06	43783	20.98
650	4471	2.14	48254	23.12
652	4635	2.22	52889	25.34
654	4926	2.36	57815	27.70
657	5060	2.42	62875	30.13
659	5567	2.67	68442	32.80
661	5594	2.68	74036	35.48
663	5935	2.84	79971	38.32
666	6093	2.92	86064	41.24
668	6269	3.00	92333	44.24
671	6598	3.16	98931	47.40
673	6750	3.23	105681	50.64
676	6991	3.35	112672	53.99
678	7259	3.48	119931	57.47
681	7478	3.58	127409	61.05
684	7561	3.62	134970	64.67
688	7745	3.71	142715	68.38
691	8020	3.84	150735	72.23
695	8019	3.84	158754	76.07
699	8299	3.98	167053	80.05
704	8431	4.04	175484	84.09
710	8294	3.97	183778	88.06
718	8029	3.85	191807	91.91
728	7305	3.50	199112	95.41
747	5983	2.87	205095	98.28
800	3599	1.72	208694	100.00

**Table H6. Grade 8 Mathematics 2008 SS Frequency Distribution, State**

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
480	604	0.29	604	0.29
538	522	0.25	1126	0.54
559	763	0.36	1889	0.90
571	920	0.44	2809	1.34
580	1076	0.51	3885	1.85
587	1171	0.56	5056	2.40
592	1292	0.61	6348	3.02
597	1326	0.63	7674	3.65
601	1371	0.65	9045	4.30
604	1511	0.72	10556	5.02
608	1552	0.74	12108	5.76
610	1551	0.74	13659	6.50
613	1715	0.82	15374	7.31
616	1667	0.79	17041	8.10
618	1705	0.81	18746	8.92
620	1856	0.88	20602	9.80
622	1969	0.94	22571	10.73
624	2061	0.98	24632	11.71
626	2074	0.99	26706	12.70
628	2208	1.05	28914	13.75
630	2176	1.03	31090	14.79
631	2258	1.07	33348	15.86
633	2308	1.10	35656	16.96
635	2421	1.15	38077	18.11
636	2555	1.22	40632	19.32
638	2572	1.22	43204	20.55
640	2633	1.25	45837	21.80
641	2761	1.31	48598	23.11
643	2709	1.29	51307	24.40
644	2855	1.36	54162	25.76
646	2971	1.41	57133	27.17
647	2968	1.41	60101	28.58
648	2990	1.42	63091	30.01
650	3269	1.55	66360	31.56

*(Continued on next page)*

**Table H6. Grade 8 Mathematics 2008 SS Frequency Distribution, State (cont.)**

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
651	3315	1.58	69675	33.14
653	3328	1.58	73003	34.72
654	3510	1.67	76513	36.39
656	3652	1.74	80165	38.13
657	3646	1.73	83811	39.86
659	3619	1.72	87430	41.58
660	3823	1.82	91253	43.40
662	4035	1.92	95288	45.32
663	4015	1.91	99303	47.23
665	4165	1.98	103468	49.21
667	4143	1.97	107611	51.18
668	4357	2.07	111968	53.25
670	4495	2.14	116463	55.39
672	4628	2.20	121091	57.59
674	4694	2.23	125785	59.82
676	4863	2.31	130648	62.13
678	4867	2.31	135515	64.45
680	5154	2.45	140669	66.90
682	5348	2.54	146017	69.44
685	5341	2.54	151358	71.98
687	5673	2.70	157031	74.68
690	5853	2.78	162884	77.47
694	5904	2.81	168788	80.27
697	5960	2.83	174748	83.11
702	6149	2.92	180897	86.03
707	6250	2.97	187147	89.01
713	6077	2.89	193224	91.90
721	5775	2.75	198999	94.64
733	5108	2.43	204107	97.07
754	4043	1.92	208150	98.99
775	2115	1.01	210265	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 PARDEX and BIGSTEPS*.
- Fitzpatrick, A. R. and M. W. Julian. 1996. *Two studies comparing the parameter estimates produced by PARDEX and PARSCLE*. Monterey, CA: CTB/McGraw-Hill.
- Fitzpatrick, A. R., V. Link, W. M. Yen, G. Burket, K. Ito, and R. Sykes. 1996. Scaling performance assessments: A comparison between one-parameter and two-parameter partial credit models. *Journal of Educational Measurement* 33: 291–314.
- Green, D. R., W. M. Yen, and G.R. Burket. 1989. Experiences in the application of item response theory in test construction. *Applied Measurement in Education* 2: 297–312.
- Hambleton, R. K., B. E. Clauser, K. M. Mazor, and R. W. Jones. 1993. Advances in the detection of differentially functioning test items. *European Journal of Psychological Assessment* 9 (1): 1–18.
- Huynh, H. and C. Schneider. 2004. Vertically moderated standards as an alternative to vertical scaling: Assumptions, practices, and an odyssey through NAEP. Paper presented at the National Conference on Large-Scale Assessment, Boston, MA, June 21.
- Jensen, A. R. 1980. *Bias in mental testing*. New York: Free Press.
- Johnson, N. L. and S. Kotz. 1970. *Distributions in statistics: continuous univariate distributions*, Vol. 2. New York: John Wiley.
- Kim, D. 2004. *WLCLASS* [Computer program].
- Kolen, M. J. and R. L. Brennan. 1995. *Test equating. Methods and practices*. New York, NY: Springer-Verlag.
- Lee, W., B. A. Hanson, and R. L. Brennan. 2002. Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement* 26: 412–432.

- Linn, R. L. 1991. Linking results of distinct assessments. *Applied Measurement in Education* 6 (1): 83–102.
- Linn, R. L., and D. Harnisch. 1981. Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement*, 18: 109–118.
- Livingston, S. A. and C. Lewis. 1995. Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement* 32: 179–197.
- Lord, F. M. 1980. *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Lord, F. M., and M. R. Novick. 1968. *Statistical theories of mental test scores*. Menlo Park, CA: Addison-Wesley.
- Mehrens, W. A., and I. J. Lehmann. 1991. *Measurement and Evaluation in Education and Psychology*, 3rd ed. New York: Holt, Rinehart, and Winston.
- Muraki, E. 1992. A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement*, 16: 159–176.
- Muraki, E., and R. D. Bock. 1991. *PARSCALE: Parameter scaling of rating data* [Computer program]. Chicago: Scientific Software, Inc.
- Novick, M. R. and P. H. Jackson. 1974. *Statistical methods for educational and psychological research*. New York: McGraw-Hill.
- Qualls, A. L. 1995. Estimating the reliability of a test containing multiple-item formats. *Applied Measurement in Education*, 8: 111–120.
- Reckase, M.D. 1979. Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics* 4: 207–230.
- Sandoval, J. H., and M. P. Mille. 1979 *Accuracy of judgments of WISC-R item difficulty for minority groups*. Paper presented at the annual meeting of the American Psychological Association. New York, August.
- Stocking, M. L. and F. M. Lord. 1983. Developing a common metric in item response theory. *Applied Psychological Measurement*, 7: 201–210.
- Thissen, D. 1982. Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika* 47: 175–186.
- Wang, T., M. J. Kolen, and D. J. Harris. 2000. Psychometric properties of scale scores and performance levels for performance assessment using polytomous IRT. *Journal of Educational Measurement*, 37: 141–162.
- Wright, B. D. and J. M. Linacre. 1992. *BIGSTEPS Rasch Analysis* [Computer program]. Chicago: MESA Press.
- Yen, W. M. 1997. The technical quality of performance assessments: Standard errors of percents of students reaching standards. *Educational Measurement: Issues and Practice*: 5–15.
- Yen, W. M. 1993. Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30: 187–213.
- Yen, W. M. 1984. Obtaining maximum likelihood trait estimates from number correct scores for the three-parameter logistic model. *Journal of Educational Measurement* 21: 93–111.
- Yen, W. M. 1981. Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, 5: 245–262.

- Yen, W. M., R. C. Sykes, K. Ito, and M. Julian. 1997 *A Bayesian/IRT index of objective performance for tests with mixed-item types*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago: March.
- Zwick, R., J. R. Donoghue, and A. Grima. 1993. Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement* 36: 225–33.