

New York State Testing Program 2006: Mathematics, Grades 3-8

Technical Report

**Submitted
December 2006**

**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 (c) 2006 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

SECTION I: INTRODUCTION AND OVERVIEW	2
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	4
TEST TRANSLATIONS	4
CHRONOLOGY	4
SECTION II: TEST DESIGN AND DEVELOPMENT.....	6
TEST DESCRIPTION	6
TEST CONFIGURATION.....	6
TEST BLUEPRINT	7
2006 ITEM MAPPING BY NEW YORK STATE STANDARDS AND STRANDS.....	15
CONTENT RATIONALE	17
ITEM DEVELOPMENT	17
ITEM REVIEW	18
MATERIALS DEVELOPMENT	18
ITEM SELECTION AND TEST CREATION (CRITERIA AND PROCESS)	19
PROFICIENCY AND PERFORMANCE STANDARDS	20
SECTION III: VALIDITY	21
CONTENT VALIDITY	21
CONSTRUCT (INTERNAL STRUCTURE) VALIDITY	22
<i>Internal consistency</i>	22
<i>Unidimensionality</i>	22
<i>Minimization of Bias</i>	24
CONSEQUENTIAL VALIDITY.....	25
SECTION IV: TEST ADMINISTRATION AND SCORING	27
TEST ADMINISTRATION	27
SCORING PROCEDURES OF OPERATIONAL TESTS.....	27
SCORING MODELS	27
SCORING OF CONSTRUCTED RESPONSE ITEMS.....	28
SCORER QUALIFICATIONS AND TRAINING	29
QUALITY CONTROL PROCESS	29
SECTION V: OPERATIONAL TEST DATA COLLECTION AND CLASSICAL ANALYSIS.....	30
DATA COLLECTION	30
DATA PROCESSING	30
SAMPLE CHARACTERISTICS	33
CLASSICAL DATA ANALYSIS	36
<i>Item Difficulty and Response Distribution</i>	37
<i>Point-Biserial Correlation Coefficients</i>	45
<i>Distracter Analysis</i>	45
<i>Test Statistics and Reliability Coefficients</i>	46
<i>Speededness</i>	46

<i>Differential Item Functioning</i>	47
SECTION VI: IRT SCALING	49
IRT MODELS AND RATIONALE FOR USE	49
CALIBRATION SAMPLE	50
CALIBRATION PROCESS	50
ITEM-MODEL FIT	51
LOCAL INDEPENDENCE	52
SCALING	53
<i>Initial Scaling</i>	54
<i>Final Scaling</i>	55
ITEM PARAMETERS	57
TEST CHARACTERISTIC CURVES	63
EQUATING	64
SCORING PROCEDURE	64
RAW SCORE TO SCALE SCORE AND SEM CONVERSION TABLES	66
STANDARD PERFORMANCE INDEX	75
IRT DIF STATISTICS	77
SECTION VII: STANDARD SETTING	80
DESCRIPTION OF STANDARD SETTING PROCESS	80
DESCRIPTION OF THE BOOKMARK METHOD	83
DESCRIPTION OF JUDGE/EXPERT PANELS	83
VERTICALLY MODERATED STANDARDS	83
DEFINITION OF PERFORMANCE LEVELS	84
FINAL CUT SCORES	84
SECTION VIII: RELIABILITY AND STANDARD ERROR OF MEASUREMENT	86
TEST RELIABILITY	86
<i>Reliability for Total Test</i>	86
<i>Reliability for MC items</i>	87
<i>Reliability for CR items</i>	88
<i>Test Reliability for NCLB reporting categories</i>	88
STANDARD ERROR OF MEASUREMENT	94
PERFORMANCE LEVEL CLASSIFICATION CONSISTENCY AND ACCURACY	95
<i>Consistency</i>	96
<i>Accuracy</i>	97
SECTION IX: SUMMARY OF OPERATIONAL TEST RESULTS	98
SCALE SCORE SUMMARY	98
<i>Grade 3</i>	99
<i>Grade 4</i>	101
<i>Grade 5</i>	103
<i>Grade 6</i>	105
<i>Grade 7</i>	107
<i>Grade 8</i>	109
PERFORMANCE LEVEL SUMMARY	111
<i>Grade 3</i>	111
<i>Grade 4</i>	113
<i>Grade 5</i>	114
<i>Grade 6</i>	116
<i>Grade 7</i>	117
<i>Grade 8</i>	119
SECTION X: REFERENCES	121
APPENDICES: APPENDIX A – CRITERIA FOR ITEM ACCEPTABILITY	124

APPENDICES: APPENDIX B – PSYCHOMETRIC GUIDELINES FOR OPERATIONAL ITEM SELECTION.....	126
APPENDICES: APPENDIX C – FACTOR ANALYSIS RESULTS.....	127
APPENDICES: APPENDIX D – DIF STATISTICS.....	143
APPENDICES: APPENDIX E – ITEM MODEL FIT STATISTICS	150
APPENDICES: APPENDIX F – DERIVATION OF THE GENERALIZED SPI PROCEDURE..	158
APPENDICES: APPENDIX G – DERIVATION OF CLASSIFICATION CONSISTENCY AND ACCURACY.....	164
APPENDICES: APPENDIX H – SCALE SCORE FREQUENCY DISTRIBUTIONS	166

List of Tables

TABLE 1. NYSTP MATH 2006 TEST CONFIGURATION	6
TABLE 2. NYSTP MATH 2006 TEST BLUEPRINT	7
TABLE 3A. NYSTP MATH 2006 OPERATIONAL TEST MAP, GRADE 3	9
TABLE 3B. NYSTP MATH 2006 OPERATIONAL TEST MAP, GRADE 4	10
TABLE 3C. NYSTP MATH 2006 OPERATIONAL TEST MAP, GRADE 5	11
TABLE 3D. NYSTP MATH 2006 OPERATIONAL TEST MAP, GRADE 6	12
TABLE 3E. NYSTP MATH 2006 OPERATIONAL TEST MAP, GRADE 7	13
TABLE 3F. NYSTP MATH 2006 OPERATIONAL TEST MAP, GRADE 8.....	14
TABLE 4. NYSTP MATH 2006 STRAND COVERAGE.....	15
TABLE 5. FACTOR ANALYSIS RESULTS FOR MATH TESTS (TOTAL POPULATION)	23
TABLE 6A. NYSTP MATH DATA CLEANING, GRADE 3	31
TABLE 6B. NYSTP MA DATA CLEANING, GRADE 4	31
TABLE 6C. NYSTP MATH DATA CLEANING, GRADE 5	31
TABLE 6D. NYSTP MATH DATA CLEANING, GRADE 6	32
TABLE 6E. NYSTP MATH DATA CLEANING, GRADE 7	32
TABLE 6F. NYSTP MATH DATA CLEANING, GRADE 8.....	32
TABLE 7A. GRADE 3 SAMPLE CHARACTERISTICS (N=183,666)	33
TABLE 7B. GRADE 4 SAMPLE CHARACTERISTICS (N=188,110).....	34
TABLE 7C. GRADE 5 SAMPLE CHARACTERISTICS (N=195,805)	34
TABLE 7D. GRADE 6 SAMPLE CHARACTERISTICS (N=200,301)	35
TABLE 7E. GRADE 7 SAMPLE CHARACTERISTICS (N=205,596).....	35
TABLE 7F. GRADE 8 SAMPLE CHARACTERISTICS (N=208,339).....	36
TABLE 8A. P-VALUES, SCORED RESPONSE DISTRIBUTIONS, AND POINT BISERIALS, GRADE 3	39
TABLE 8B. P-VALUES, SCORED RESPONSE DISTRIBUTIONS, AND POINT BISERIALS, GRADE 4	40
TABLE 8C. P-VALUES, SCORED RESPONSE DISTRIBUTIONS, AND POINT BISERIALS, GRADE 5	41
TABLE 8D. P-VALUES, SCORED RESPONSE DISTRIBUTIONS, AND POINT BISERIALS, GRADE 6	42
TABLE 8E. P-VALUES, SCORED RESPONSE DISTRIBUTIONS, AND POINT BISERIALS, GRADE 7	43
TABLE 8F. P-VALUES, SCORED RESPONSE DISTRIBUTIONS, AND POINT BISERIALS, GRADE 8	44
TABLE 9. NYSTP MATH 2006 TEST FORM STATISTICS AND RELIABILITY	46
TABLE 10. NYSTP MATH 2006 CLASSICAL DIF SAMPLE N-COUNTS	48

TABLE 11. NUMBER OF ITEMS FLAGGED BY SMD AND MANTEL-HAENSZEL DIF METHODS.....	48
TABLE 12. NYSTP MATH 2006 CALIBRATION RESULTS.....	51
TABLE 13. NYSTP MATH 2006 INITIAL TRANSFORMATION CONSTANTS.....	54
TABLE 14. NYSTP MATH 2006 FINAL TRANSFORMATION CONSTANTS	56
TABLE 15A. GRADE 3 2006 OPERATIONAL ITEM PARAMETER ESTIMATES	57
TABLE 15B. GRADE 4 2006 OPERATIONAL ITEM PARAMETER ESTIMATES.....	58
TABLE 15C. GRADE 5 2006 OPERATIONAL ITEM PARAMETER ESTIMATES	59
TABLE 15D. GRADE 6 2006 OPERATIONAL ITEM PARAMETER ESTIMATES	60
TABLE 15E. GRADE 7 2006 OPERATIONAL ITEM PARAMETER ESTIMATES.....	61
TABLE 15F. GRADE 8 2006 OPERATIONAL ITEM PARAMETER ESTIMATES.....	62
TABLE 16. NYSTP MATH 2006 MINIMUM AND MAXIMUM SCALE SCORES	66
TABLE 17A. GRADE 3 RAW SCORE TO SCALE SCORE (WITH STANDARD ERROR)	67
TABLE 17B. GRADE 4 RAW SCORE TO SCALE SCORE (WITH STANDARD ERROR)	68
TABLE 17C. GRADE 5 RAW SCORE TO SCALE SCORE (WITH STANDARD ERROR)	70
TABLE 17D. GRADE 6 RAW SCORE TO SCALE SCORE (WITH STANDARD ERROR)	71
TABLE 17E. GRADE 7 RAW SCORE TO SCALE SCORE (WITH STANDARD ERROR)	72
TABLE 17F. GRADE 8 RAW SCORE TO SCALE SCORE (WITH STANDARD ERROR)	73
TABLE 18. SPI TARGET RANGES.....	76
TABLE 19. NUMBER OF ITEMS FLAGGED FOR DIF BY THE LINN-HARNISCH METHOD .	79
TABLE 20. MEASUREMENT REVIEW MEETING BASED RECOMMENDED IMPACT DATA	80
TABLE 21. PARTICIPANTS BASED CUT SCORES AND ASSOCIATED IMPACT DATA.....	81
TABLE 22. VERTICAL ARTICULATION PANEL BASED CUT SCORES AND ASSOCIATED IMPACT DATA (TABLE LEADER SMOOTHING).....	82
TABLE 23. MEASUREMENT REVIEW FORUM BASED CUT SCORES AND ASSOCIATED IMPACT DATA	82
TABLE 24. FINAL CUT SCORES NYSTP MATH	85
TABLE 25. RELIABILITY AND STANDARD ERROR OF MEASUREMENT FOR THE 2006 NYSTP MATH EXAMS	86
TABLE 26. RELIABILITY AND STANDARD ERROR OF MEASUREMENT FOR THE 2006 NYSTP MATH EXAMS – MC ITEMS ONLY	87
TABLE 27. RELIABILITY AND STANDARD ERROR OF MEASUREMENT FOR THE 2006 NYSTP MATH EXAMS – CR ITEMS ONLY	88
TABLE 28A. GRADE 3 TEST RELIABILITY BY SUBGROUP	89
TABLE 28B. GRADE 4 TEST RELIABILITY BY SUBGROUP	90
TABLE 28C. GRADE 5 TEST RELIABILITY BY SUBGROUP	91
TABLE 28D. GRADE 6 TEST RELIABILITY BY SUBGROUP	92
TABLE 28E. GRADE 7 TEST RELIABILITY BY SUBGROUP	93
TABLE 28F. GRADE 8 TEST RELIABILITY BY SUBGROUP.....	94

TABLE 29. DECISION CONSISTENCY (ALL CUTS).....	96
TABLE 30. DECISION CONSISTENCY (LEVEL III CUT).....	96
TABLE 31. DECISION AGREEMENT (ACCURACY)	97
TABLE 32. MATH GRADES 3-8 SCALE SCORE DISTRIBUTION SUMMARY.....	98
TABLE 33. SCALE SCORE SUMMARY, BY SUBGROUP, GRADE 3.....	100
TABLE 34. SCALE SCORE SUMMARY, BY SUBGROUP, GRADE 4.....	102
TABLE 35. SCALE SCORE SUMMARY, BY SUBGROUP, GRADE 5.....	104
TABLE 36. SCALE SCORE SUMMARY, BY SUBGROUP, GRADE 6.....	106
TABLE 37. SCALE SCORE SUMMARY, BY SUBGROUP, GRADE 7.....	108
TABLE 38. SCALE SCORE SUMMARY, BY SUBGROUP, GRADE 8.....	110
TABLE 39. GRADES 3-8 MATH TEST PERFORMANCE LEVEL DISTRIBUTIONS.....	111
TABLE 40. PERFORMANCE LEVEL SUMMARY, BY SUBGROUP, GRADE 3.....	112
TABLE 41. PERFORMANCE LEVEL SUMMARY, BY SUBGROUP, GRADE 4.....	113
TABLE 42. PERFORMANCE LEVEL SUMMARY, BY SUBGROUP, GRADE 5.....	115
TABLE 43. PERFORMANCE LEVEL SUMMARY, BY SUBGROUP, GRADE 6.....	116
TABLE 44. PERFORMANCE LEVEL SUMMARY, BY SUBGROUP, GRADE 7.....	118
TABLE 45. PERFORMANCE LEVEL SUMMARY, BY SUBGROUP, GRADE 8.....	119

Section I: Introduction and Overview

Introduction

An overview of the New York State Testing Program (NYSTP), Grades 3 through 8, Mathematics (Math) 2006 Operational Tests is provided in this report. The report contains information about operational test development and content, item and test statistics, validity and reliability, differential item functioning studies, test administration and scoring, scaling, and student performance.

Test Purpose

The NYSTP is an assessment system designed to measure concepts, processes, and skills taught in schools in New York. The Math Tests target student progress toward three of the four content standards as described in Section II of this report (Test Design and Development, subsection Content Rationale). The Grades 3-8 Math Tests are written so as to allow all students to have the opportunity to demonstrate their knowledge and skills in these standards. The established cut scores classify student proficiency into one of four levels based on their test performance.

Target Population

Students in New York State public school Grades 3, 4, 5, 6, 7, and 8 (and ungraded students of equivalent age) are the target population for the Grades 3-8 Math Tests. Nonpublic schools may participate in the testing program but the participation is not mandatory for them. In 2006, non public schools participated primarily in the Grades 4 and 8 Tests. Given that non-public schools were not well represented in the testing program, NYSED made a decision to exclude these schools from the data analyses. Public school students must 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 for students with severe disabilities (NYSAA). For more detail on this exemption, please refer to page 2 of the *Mathematics School Administrator's Manual for Public Schools* (SAM, available online at: <http://www.emsc.nysed.gov/3-8/sam/home.htm>)

Test Use and Decisions Based on Assessment

The Grades 3-8 Math Tests are used to measure the extent to which individual students achieve the New York State learning standards in Math, 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 Math 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 Math Tests at each grade level. The scale scores are comparable within each grade level but not across grades because the Grades 3-8 Math 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) of this report. Uses of Grades 3-8 Math Tests scores include: determining student progress within schools and districts, supporting registration of schools and districts, determining eligibility of students for additional instruction time, and providing teachers with indicators of a student's need, or lack of need, for remediation in specific subject area knowledge.

Proficiency Level Cut Score and Classification

The proficiency cut scores (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 Math proficiency cut scores reflect the abilities intended by the New York State Education Department. Performance of students on the Grades 3-8 Math Tests in relation to proficiency level cut scores is reported in a form of Performance Level 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 performance of schools and districts, and the state, is 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 Section VII (Standard Setting) of this report, and in-depth information is provided in the *Bookmark Standard Setting Technical Report 2006 for Grades 3, 4, 5, 6, 7, and 8 Mathematics* and the *NYS MA 2006 Measurement Review Technical Report 2006 for Mathematics*.

Standard Performance Index Scores

Standard Performance Index (SPI) scores are obtained from the Grades 3-8 Math 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) of this report.

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, & National Council on Measurement in Education, 1999, 2002, 2004), 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. Greater detail on testing accommodations can be found in pages 3-5 of the *School Administrator's Manual*.

Test Transcriptions

The tests are transcribed into Braille and Large Type forms, for students that are visually impaired. The students dictate and/or record their responses, and the teachers transcribe student responses onto regular (scannable) answer sheets. The large type forms are created by CTB/McGraw-Hill, and the Braille forms 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 have Library of Congress and Nemeth Code [Braille] Certifications. Braille Publishers, Inc. produced the Braille forms for the previous Grades 4 and 8 tests.

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

Test Translations

Since these are tests of Mathematical ability, the NYSTP 3-8 MA 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 released in each translated language and are available 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.

Chronology

The high level chronology of the test development, administration, and data analysis occurred is outlined below.

1. Test design (2004 - 2005)
 - a. Develop content specifications
 - b. Design test configurations
 - c. Write/Receive approved content standards
 - d. Design test blueprints (targets for test coverage of standards)
2. Item development and field testing (2004-2005)
 - a. Item development (2004)

- b. Field test (May 2005)
 - c. Rangefinding and scoring of field test data (June-July 2005)
 - d. Analyze data from field test (August 2005)
- 3. Operational test construction (September 2005)
- 4. Test administration (March 2006)
- 5. Scoring and data retrieval (March-May 2006)
- 6. Data analysis (May-June 2006)
- 7. Standard setting (July 2006)
- 8. Score reporting (September 2006)

Section II: Test Design and Development

Test Description

The NYSTP 2005-2006 Math operational tests are New York State standards based criterion-referenced exams 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 classrooms during March 2006 over a two-day period for Grades 3, 5, 6, and 7 and 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 (<http://www.nysedregents.org/testing/mathe/06exams/home.htm>). More details on the administration and scoring of these tests can be found in Section IV of this report.

Test Configuration

The operational tests books were administered, in order, on two to three consecutive days, depending on the grade. Table 1 below provides information on the number and type of items in each book as well as testing times. Book 1 contained only MC items. Books 2 and 3 contained only CR items. The *Sample Test Teacher's Directions* (<http://www.emsc.nysed.gov/3-8/math-sample/home.htm>) and the 2006 *School Administrator's Manual* (<http://www.emsc.nysed.gov/3-8/sam/home.htm>) provide more detail on security, scheduling, classroom organization and preparation, test materials, and administration.

Table 1. NYSTP Math 2006 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 Math 2006 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		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 Math tests assess student performance on the content and process strands of New York Mathematics Learning Standard 3. The test items assess a variety of performance indicators in these strands. Each question is aligned to 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 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. NYSTP Math 2006 Test Blueprint

Grade	Total Points	Content Strand	Target # Points	Selected # Points	Target % of test	Selected % of test
3	39	Number Sense and Operations	19	17	48%	44%
		Algebra	5	6	13%	15%
		Geometry	5	5	13%	13%
		Measurement	5	5	13%	13%
		Statistics and Probability	5	6	13%	15%

(Continued on next page)

Table 2. NYSTP Math 2006 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	32	45%	46%
		Algebra	10	10	14%	14%
		Geometry	8	8	12%	11%
		Measurement	12	12	17%	17%
		Statistics and Probability	8	8	12%	11%
5	46	Number Sense and Operations	18	16	39%	35%
		Algebra	5	5	11%	11%
		Geometry	12	12	25%	26%
		Measurement	6	7	14%	15%
		Statistics and Probability	5	6	11%	13%
6	49	Number Sense and Operations	18	20	37%	41%
		Algebra	9	9	19%	18%
		Geometry	8	8	17%	16%
		Measurement	6	6	11%	12%
		Statistics and Probability	8	6	16%	12%
7	50	Number Sense and Operations	15	13	30%	26%
		Algebra	6	4	12%	8%
		Geometry	7	9	14%	18%
		Measurement	7	9	14%	18%
		Statistics and Probability	15	15	30%	30%
8	69	Number Sense and Operations	8	9	11%	13%
		Algebra	30	29	44%	42%
		Geometry	24	21	35%	30%
		Measurement	7	10	10%	14%

Table 3a. NYSTP Math 2006 Operational Test Map, Grade 3

Item#	Item Type	Answer Key	Max Points	Content Strand	Performance Indicator
1	MC	B	1	4	3.M02
2	MC	H	1	1	3.N03
3	MC	D	1	4	3.M07
4	MC	F	1	2	3.A01
5	MC	C	1	1	3.N02
6	MC	G	1	1	3.N18
7	MC	C	1	1	3.N03
8	MC	J	1	1	3.N16
9	MC	B	1	1	3.N19
10	MC	H	1	4	3.M02
11	MC	A	1	1	3.N18
12	MC	J	1	1	3.N24
13	MC	C	1	4	3.M07
14	MC	H	1	1	3.N21
15	MC	B	1	1	3.N07
16	MC	G	1	1	3.N25
17	MC	D	1	1	3.N13
18	MC	J	1	4	3.M09
19	MC	C	1	1	3.N22
20	MC	G	1	1	3.N10
21	MC	A	1	2	3.A01
22	MC	H	1	3	3.G03
23	MC	C	1	1	3.N25
24	MC	G	1	2	3.A02
25	MC	D	1	5	3.S07
26	SR	n/a	2	1	3.N18
27	SR	n/a	2	3	3.G01
28	ER	n/a	3	5	3.S07
29	ER	n/a	3	2	3.A02
30	SR	n/a	2	5	3.S05
31	SR	n/a	2	3	3.G01

Table 3b. NYSTP Math 2006 Operational Test Map, Grade 4

Item#	Item Type	Answer Key	Max Points	Content Strand	Performance Indicator
1	MC	B	1	1	4.N02
2	MC	H	1	4	4.M02
3	MC	B	1	1	3.N20
4	MC	J	1	1	4.N15
5	MC	B	1	1	4.N26
6	MC	H	1	1	3.N14
7	MC	A	1	4	4.M08
8	MC	J	1	1	4.N18
9	MC	C	1	4	4.M02
10	MC	G	1	1	4.N18
11	MC	D	1	1	4.N18
12	MC	J	1	2	3.A01
13	MC	B	1	5	4.S06
14	MC	H	1	2	4.A04
15	MC	B	1	1	4.N14
16	MC	G	1	5	4.S05
17	MC	C	1	2	4.A04
18	MC	J	1	1	4.N06
19	MC	C	1	4	4.M09
20	MC	G	1	2	4.A04
21	MC	C	1	3	4.G03
22	MC	F	1	4	4.M04
23	MC	B	1	1	4.N22
24	MC	H	1	3	4.G02
25	MC	D	1	1	4.N13
26	MC	J	1	1	3.N15
27	MC	A	1	1	4.N14
28	MC	F	1	4	4.M01
29	MC	D	1	2	4.A04
30	MC	F	1	3	4.G01
31	SR	n/a	2	4	4.M08
32	ER	n/a	3	3	4.G03
33	SR	n/a	2	1	4.N18
34	SR	n/a	2	4	4.M08
35	SR	n/a	2	1	4.N21
36	SR	n/a	2	1	4.N17
37	SR	n/a	2	1	4.N20
38	ER	n/a	3	5	4.S03
39	SR	n/a	2	2	4.A01
40	SR	n/a	2	1	4.N14

(Continued on next page)

Table 3b. NYSTP Math 2006 Operational Test Map, Grade 4 (cont.)

Item#	Item Type	Answer Key	Max Points	Content Strand	Performance Indicator
41	SR	n/a	2	1	4.N14
42	SR	n/a	2	1	4.N14
43	SR	n/a	2	1	4.N20
44	SR	n/a	2	4	4.M08
45	ER	n/a	3	2	4.A04
46	SR	n/a	2	3	4.G01
47	ER	n/a	3	5	4.S06
48	SR	n/a	2	1	4.N18

Table 3c. NYSTP Math 2006 Operational Test Map, Grade 5

Item#	Item Type	Answer Key	Max Points	Content Strand	Performance Indicator
1	MC	C	1	1	5.N03
2	MC	H	1	4	5.M01
3	MC	B	1	2	5.A08
4	MC	H	1	2	5.A07
5	MC	B	1	3	5.G11
6	MC	G	1	1	5.N19
7	MC	A	1	1	5.N05
8	MC	H	1	1	5.N23
9	MC	C	1	1	5.N16
10	MC	F	1	3	5.G03
11	MC	D	1	1	5.N16
12	MC	G	1	4	5.M07
13	MC	D	1	1	5.N19
14	MC	F	1	1	5.N22
15	MC	A	1	5	4.S04
16	MC	J	1	3	5.G04
17	MC	D	1	2	5.A06
18	MC	H	1	1	5.N21
19	MC	A	1	5	5.S02
20	MC	J	1	4	5.M05
21	MC	B	1	1	5.N07
22	MC	J	1	5	5.S03
23	MC	A	1	4	5.M08
24	MC	G	1	2	4.A02
25	MC	B	1	2	5.A07
26	MC	F	1	3	5.G10
27	SR	n/a	2	1	5.N17

(Continued on next page)

Table 3c. NYSTP Math 2006 Operational Test Map, Grade 5 (cont.)

Item#	Item Type	Answer Key	Max Points	Content Strand	Performance Indicator
28	ER	n/a	3	3	5.G11
29	ER	n/a	3	3	5.G04
30	SR	n/a	2	1	5.N26
31	SR	n/a	2	1	5.N16
32	SR	n/a	2	3	5.G08
33	ER	n/a	3	5	5.S02
34	ER	n/a	3	4	5.M03

Table 3d. NYSTP Math 2006 Operational Test Map, Grade 6

Item#	Item Type	Answer Key	Max Points	Content Strand	Performance Indicator
1	MC	D	1	1	6.N04
2	MC	F	1	3	5.G12
3	MC	C	1	4	6.M05
4	MC	G	1	5	5.S05
5	MC	A	1	1	6.N18
6	MC	H	1	3	6.G06
7	MC	B	1	4	6.M03
8	MC	G	1	1	6.N21
9	MC	C	1	1	6.N02
10	MC	F	1	2	6.A02
11	MC	D	1	4	6.M03
12	MC	G	1	1	6.N14
13	MC	B	1	5	6.S05
14	MC	H	1	1	6.N11
15	MC	D	1	1	6.N22
16	MC	H	1	2	6.A06
17	MC	B	1	1	6.N19
18	MC	G	1	2	6.A06
19	MC	C	1	1	6.N25
20	MC	H	1	3	6.G01
21	MC	B	1	3	6.G06
22	MC	H	1	5	5.S06
23	MC	D	1	2	5.A03
24	MC	H	1	2	6.A01
25	MC	B	1	1	6.N14
26	SR	n/a	2	3	6.G02
27	SR	n/a	2	1	6.N11
28	SR	n/a	2	3	5.G14

(Continued on next page)

Table 3d. NYSTP Math 2006 Operational Test Map, Grade 6 (cont.)

Item#	Item Type	Answer Key	Max Points	Content Strand	Performance Indicator
29	ER	n/a	3	4	6.M01
30	SR	n/a	2	2	5.A04
31	SR	n/a	2	2	6.A01
32	SR	n/a	2	1	6.N24
33	ER	n/a	3	1	6.N12
34	ER	n/a	3	5	5.S05
35	ER	n/a	3	1	6.N26

Table 3e. NYSTP Math 2006 Operational Test Map, Grade 7

Item#	Item Type	Answer Key	Max Points	Content Strand	Performance Indicator
1	MC	D	1	1	7.N12
2	MC	F	1	3	7G.03
3	MC	C	1	1	7.N11
4	MC	J	1	1	7.N09
5	MC	C	1	5	7.S04
6	MC	J	1	4	7.M02
7	MC	C	1	5	7.S06
8	MC	H	1	2	7.A01
9	MC	B	1	1	7.N11
10	MC	H	1	1	7.N08
11	MC	B	1	2	7.A01
12	MC	J	1	5	6.S11
13	MC	B	1	1	7.N06
14	MC	F	1	2	7.A01
15	MC	B	1	1	7.N12
16	MC	J	1	4	7.M02
17	MC	B	1	3	7G.03
18	MC	J	1	1	7.N05
19	MC	B	1	3	7G.01
20	MC	J	1	4	7.M03
21	MC	B	1	5	7.S06
22	MC	H	1	1	7.N18
23	MC	A	1	2	6.A03
24	MC	H	1	4	7.M04
25	MC	D	1	5	6.S09
26	MC	H	1	5	7.S08
27	MC	B	1	3	7G.03
28	MC	H	1	1	7.N07

(Continued on next page)

Table 3e. NYSTP Math 2006 Operational Test Map, Grade 7 (cont.)

Item#	Item Type	Answer Key	Max Points	Content Strand	Performance Indicator
29	MC	B	1	3	7G.01
30	MC	G	1	5	7.S09
31	SR	n/a	2	5	6.S04
32	SR	n/a	2	3	7G.02
33	ER	n/a	3	1	7.N12
34	ER	n/a	3	4	7.M08
35	SR	n/a	2	4	7.M04
36	SR	n/a	2	3	7G.04
37	ER	n/a	3	5	6.S03
38	ER	n/a	3	5	7.S08

Table 3f. NYSTP Math 2006 Operational Test Map, Grade 8

Item#	Item Type	Answer Key	Max Points	Content Strand	Performance Indicator
1	MC	C	1	2	8.A10
2	MC	J	1	3	8.G01
3	MC	C	1	3	8.G03
4	MC	H	1	1	8.N04
5	MC	D	1	2	8.A08
6	MC	G	1	2	7.A02
7	MC	C	1	2	8.A06
8	MC	G	1	3	8.G07
9	MC	C	1	2	8.A07
10	MC	G	1	3	8.G02
11	MC	B	1	2	7.A04
12	MC	H	1	4	8.M01
13	MC	B	1	1	8.N05
14	MC	H	1	4	7.M01
15	MC	A	1	2	8.A01
16	MC	H	1	2	7.A10
17	MC	C	1	3	8.G01
18	MC	F	1	2	7.A04
19	MC	C	1	2	8.A02
20	MC	H	1	3	8.G05
21	MC	D	1	3	8.G07
22	MC	J	1	2	8.A01
23	MC	C	1	2	7.A10
24	MC	J	1	2	7.A04
25	MC	B	1	3	8.G03

(Continued on next page)

Table 3f. NYSTP Math 2006 Operational Test Map, Grade 8 (cont.)

Item#	Item Type	Answer Key	Max Points	Content Strand	Performance Indicator
26	MC	G	1	2	8.A01
27	MC	D	1	3	7.G08
28	SR	n/a	2	1	8.N04
29	ER	n/a	3	3	8.G05
30	SR	n/a	2	4	7.M01
31	ER	n/a	3	2	7.A04
32	SR	n/a	2	3	7.G08
33	SR	n/a	2	2	7.A10
34	SR	n/a	2	4	7.M06
35	SR	n/a	2	3	7.G09
36	ER	n/a	3	2	8.A12
37	SR	n/a	2	3	8.G10
38	SR	n/a	2	2	7.A04
39	ER	n/a	3	1	8.N04
40	SR	n/a	2	4	7.M05
41	SR	n/a	2	2	7.A04
42	ER	n/a	3	2	7.A10
43	ER	n/a	3	3	8.G07
44	SR	n/a	2	1	8.N06
45	SR	n/a	2	4	8.M01

2006 Item Mapping by New York State Standards and Strands**Table 4. NYSTP Math 2006 Strand Coverage**

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

(Continued on next page)

Table 4. NYSTP Math 2006 Strand Coverage (cont.)

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

Note: Three items from grade 7 (#4, 11, and 15) and one item from grade 8 (#17) were suppressed from scoring due to item exposure and do not contribute to nor detract from student scores.

Content Rationale

In August 2004, CTB/McGraw-Hill facilitated specifications meetings in Albany, NY 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
- how much emphasis to place on each assessable performance indicator
- what were the limitations, if any, to be applied to the assessable performance indicators
- what were some general examples of items that could be used
- finalization of the test blueprint for each grade

The committees, selected from around the state for their grade-level expertise, were grouped by grade band (i.e., 3/4, 5/6, 7/8) and met for four days. Upon completion of the committee meetings, NYSED reviewed the committees' determinations and approved them, with minor adjustments when necessary to maintain consistency across the grades. In January 2005, a second specifications meeting, again with New York State educators from around the state, was held in order to review changes made to the New York State Mathematics Learning Standard 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 comprised of New York State educators were selected for their content and grade-level expertise for Item Review. 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 (for constructed-response items)
- the appropriateness of the correct response and distracters, in the case of 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. In December, 2005, Field Test enrollment was conducted online, accompanied by active recruitment letters. Participation was determined by the State matrix, and follow-up phone calls from CTB/McGraw-Hill reminded schools, as needed, to fulfill their enrollment requirements. After the enrollment period closed, final enrollments were tallied and reviewed for demographic characteristics. Test forms were assigned to participants at the school (grade) level while balancing the demographic statistics across forms, in order to proactively sample the students. The Field Tests were administered to students across New York State during the week of May 23, 2005. 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 June 2005 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. 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.

CTB/McGraw-Hill also developed a *Guide to the Grades 3-8 Testing Program*, which consisted of several sections: an *Introduction to the Grades 3-8 Testing Program* (posted at <http://emsc33.nysed.gov/3-8/intro.pdf>) as well as a sample test (which mirrored the operational test), a *Teacher's Directions* (<http://www.emsc.nysed.gov/3-8/math-sample/home.htm>), and a *Scoring Guide* for each grade (posted at <http://emsc33.nysed.gov/3-8/intro.pdf>). This *Guide* was also printed and delivered to schools.

Item Selection and Test Creation (criteria and process)

The first operational Grades 3-8 Math Tests were administered in March 2006. 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 Math 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. Final approval of the selected items was given by NYSED. Two criteria governed the item selection process. The first of these was to meet the content specifications provided by the New York State Department of Education. Next, within the limits set by these requirements, developers selected items with the best psychometric characteristics from the field test item pool.

Item selection for the operational tests was facilitated using the proprietary program ITEMWIN (Burket, 1988). ITEMWIN 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, & Burket, 1989).

ITEMWIN has three parts. The first part selects a working item pool of manageable size from the larger pool. The second part of the program uses this selected item pool to perform the final test selection. In the third part of the program, a table shows both 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 (see below), 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, they were reviewed for alignment with the test design, Blueprint, and the Research guidelines for item selection (see Appendix B).

When approved internally, preliminary selections were sent to NYSED staff for their review. NYSED staff (including their Content and Research representative experts) traveled to CTB/McGraw-Hill in Monterey in September 2005 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. After approval by NYSED, the tests were produced and administered in March 2006.

In addition to the test books, CTB/McGraw-Hill produced two *School Administrator's Manuals* (one for public schools (<http://www.emsc.nysed.gov/3-8/sam/math06p.pdf>) and one for nonpublic schools (<http://www.emsc.nysed.gov/3-8/sam/math06np.pdf>) and *Teacher's Directions* for each grade (<http://www.emsc.nysed.gov/3-8/directions/home.htm>) so that the tests were administered in a standardized fashion across the state.

Proficiency and Performance Standards

Proficiency cut score recommendations and the drafting of performance standards occurred at the NYSTP Math 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. Section VII of this report, Standard Setting, provides an overview of the method, participants, achievement levels, and results (impact). For specific detail, please refer to the *Bookmark Standard Setting Technical Report 2006 for Grades 3, 4, 5, 6, 7, and 8 Mathematics* and *NYS 2006 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 on-going 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 element for validity. A test can not be valid if it is not also reliable.

The American Educational Research Association (AERA), American Psychological Association (APA), and the 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 always refers to the degree to which that evidence supports the inferences that are made from the scores. The inferences regarding specific uses of a test are validated, not the test itself.

Content Validity

Generally, achievement tests are used for student level outcomes, either (1) making predictions about students, or (2) describing students' performance (Mehrens & 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, the NYSTP documents student performance in the area of Math as defined by the New York State Math 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 to 4 in Section II). The test development process requires specific attention to content representation and the balance thereof within each test form. New York State educators were involved in test constructions in various test development stages. For example, they reviewed field tests for their alignment with test blueprint. They also participated in a process of establishing scoring rubrics for constructed response items. Section II (Test Design and Development) of this report contains more information

specific to item review process. An independent study of alignment between the New York State curriculum and the New York State Grades 3-8 Math 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 Math Tests is supported by several types of evidence that can be obtained from the Math 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, are reliable and consistent. Reliability coefficients of the tests for total populations and subgroups of students are presented in Section VIII (Reliability and Standard Error of Measurement). For the total populations the reliability coefficients ranged from 0.89 to 0.96 and for all subgroups, the reliability coefficients are greater than 0.80. Overall, high internal consistency of New York State Math 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 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. It was found that all items in grades 3 and 5 and over 90% of the items on other Math tests display good item-model fit, which provides solid evidence for the appropriateness of IRT models used to calibrate and scale the test data. Another evidence for the efficacy of modeling ability is provided by demonstrating that the questions on New York State Math tests are related. What relates the questions is most parsimoniously claimed to be the common ability acquired by students studying the subject. Factor analysis of the test data is one way of modeling the common ability. This analysis may show that there is a single or main factor that can account for much of the relationship between the test questions. A large first component would provide evidence of the latent ability which is the primary cognitive behavior students have in common with respect to the particular questions asked. A large main factor found from a factor analysis of an achievement test suggests a univocal ability construct that may be considered to be what the questions were designed to have in common, i.e., Mathematics ability.

To demonstrate the common factor (ability) underlying student responses to Math 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 Math 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 which are appropriate only for MC items). The study was done 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. Figures 1, 5, 9, 13, 17 and 20 in Appendix C provide scree plots (Cattell, 1966) of eigenvalues that demonstrate essential unidimensionality of the trait measured by each test.

It was found that more than one factor with eigenvalue greater than 1.0 was present in each data set which would suggest the presence of small additional factors. However the ratio of the variance accounted for by the first factor to the remaining factors is sufficiently large to support the claim that these tests are essentially unidimensional. These ratios showed that the first eigenvalues were at least 5 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 Math 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, below.

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

Grade	Initial Eigenvalues			
	Component	Total	% of Variance	Cumulative %
3	1	7.88	25.41	25.41
	2	1.21	3.89	29.30
	3	1.07	3.45	32.74
4	1	13.20	27.49	27.49
	2	1.49	3.11	30.60
	3	1.22	2.53	33.14
	4	1.00	2.09	35.23
5	1	9.00	26.47	26.47
	2	1.36	3.99	30.46
	3	1.04	3.06	33.52

(Continued on next page)

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

Grade	Initial Eigenvalues			
	Component	Total	% of Variance	Cumulative %
6	1	9.60	27.43	27.43
	2	1.53	4.36	31.80
	3	1.09	3.12	34.92
7	1	8.22	23.50	23.50
	2	1.52	4.33	27.83
	3	1.14	3.25	31.07
	4	1.02	2.92	33.99
8	1	14.04	31.90	31.90
	2	1.30	2.95	34.86
	3	1.10	2.50	37.36
	4	1.00	2.28	39.64

This evidence supports the claim that there is a construct ability underlying the items/tasks in each Math 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 Math 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 Math tests for the analyzed subgroups. Factor analysis results for LEP, SWD and students using accommodations are provided in Table C1 of Appendix C in this report.

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 SES (socioeconomic status) bias. All materials were written and reviewed to conform to the company'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 specifications and carefully checked by groups of trained New York State educators.

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

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

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

The second step 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 filed test materials was reviewed by at least these same people.

In the third procedure, New York State educators who reviewed all field test materials. These professionals were asked to consider and comment on the appropriateness of language, subject matter, and representation of people.

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 & Mille, 1979; Jensen, 1980). Thus, empirical studies were conducted.

As a 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 amount of DIF in all test items: standardized mean difference, Mantel-Haenszel (described in Section V – Data Collection and Classical Analysis), and Linn-Harnisch (described in Section VI – IRT Scaling). 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 items that were deemed free of bias were included in the operational tests

Consequential Validity

The *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999) addressed the concept of consequential validity in testing indicating that when educational testing programs are mandated by school, district, state or other authorities, the ways in which test results are intended to be used should be clearly described. It is the responsibility of those who mandate the use of tests to monitor their impact and to identify and minimize potential negative consequences. Consequences resulting from the uses of the test, both intended and unintended, should also be examined by the test user. Efforts should be made to document the provision of instruction in tested content and skills.

Consequential validity is often referred to as the social consequences of using a particular test for a particular purpose. The use of a test is said to have consequential validity to the extent that society benefits from that use of the test. Consequential validity is relevant to test use and score interpretation and is not directly related to test properties. For this

reason, it is not straightforward to measure/collect evidence on the consequential aspects of validity. The test data alone do not provide sufficient evidence of this type of validity. Evaluation of consequential evidence may for instance involve examining variation in school performance in terms of contextual and evidential variables. Information on teachers' instruction and classroom assessment practices is very important in understanding the success or failure of accountability systems and reform efforts. Teacher surveys or focus groups can be used to collect information regarding the use of the tests and how the tests impacted the curriculum and instruction. A better understanding of the extent to which performance gains on assessments reflect improved instruction and student learning rather than more superficial interventions such as narrow test preparation activities would also provide evidence of consequential validity. Because 2006 is the baseline year of the new York State Grades 3-8 testing program and there is no history of student performance in grades 3, 5, 6 and 7 no score gain analyses can be conducted for these grades based on 2006 test data. Grade 4 and 8 assessments were administered in the past but no direct equating of 2006 to 2005 assessments was conducted.

Given the limitations of the first year test data, it is proposed to revisit the issue of consequential validity with the test scores in year 2007 and beyond, when data from more than one administration are available for analysis. Longitudinal test data along with additional information collected from New York State educators (e.g., information on understanding of learning standards, motivation and effort to adapt the curriculum and instruction to content standards, instructional practices, classroom assessment format and content, use and nature of test assessment preparation activities, professional development) will allow for meaningful analyses and interpretation of the score gain and uniformity of standards, learning expectations, and consequences for all students.

Section IV: Test Administration and Scoring

Listed below are brief summaries of New York State test administration and scoring procedures. For a greater understanding of the paragraphs below, please review the *New York State Scoring Leader Handbooks* and SAM (*School Administrator's Manual*). In addition, please refer to Scoring Site Operations Manual (2006) posted at <http://www.emsc.nysed.gov/3-8/archived.htm#scoring>.

Test Administration

NYSTP Grades 3-8 Math Tests were administered at the classroom level, during March, 2006. The testing window for grades 3, 4, and 5 was March 6th through 10th, 2006. The testing window for grades 6, 7, and 8 was January 13th through 17th. The make-up test administration window was March 13th through 17th for grades 3-5 and from March 20-24 for grades 6-8. The make up testing window allowed for students that 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” encompassed scorers at every site.

Scoring Models

For the 2005-06 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 Math Tests. Schools were able to score these tests regionally, district-wide, or individually. Schools were required to enter one of the following “scoring model codes” on student answer sheets:

1. Regional scoring – The 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) – in this model 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 (S/CDN) provided districts with technical support and advice in making this decision. In addition, please refer to the following link for a brief comparison between regional/district scoring and local scoring (see Attachment C at: <http://www.emsc.nysed.gov/3-8/update-rev-dec05.htm>).

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 with guidance from NYSED and New York State teachers. In Summer of 2005, Handscoring met with groups of teachers from across the state in Rangefinding sessions. Sets of actual 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 NYSED policies. Handscoring created Scoring Guides based on Rangefinding decisions and conferences with NYSED. Handscoring also aided in the creation of a DVD, which explained each section of the Scoring Guides in greater detail. 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. Handscoring staff also conducted training sessions in New York City to better equip teachers and administrators with enhanced knowledge of scoring principles and criteria.

At this time, scoring is 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, Math FAQs (at: <http://emsc33.nysed.gov/3-8/faq.htm>), and a DVD, which highlighted important elements of the Scoring Guides. Each test booklet was scored by 3 separate Scoring Committee Members, who scored 3 distinct sections of the test book. After each test book was completed, the Table Facilitator or Mathematics Leader conducted a “read-behind” of approximately 12 sets of booklets 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 Quality Control Process subsection).

Scorer Qualifications and Training

The scoring of the operational test was conducted by pre-qualified administrators and teachers. Trainers used the Scoring Guides to train Scoring Committee Members on the criteria for scoring constructed response items. 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 are evenly dispersed. Teams were broken down into groups of three to ensure that a variety of scorers touch each book. If a scorer and facilitator could not reach a decision on a paper after reviewing the Scoring Guides, Math FAQs, and DVD, they called the New York State Helpline, a call center established to aid teachers and administrators during Operational scoring. The Helpline staff consisted of previously trained and prepared CTB/McGraw-Hill Handscoring personnel who answered questions by phone, fax, or email. When a member of the staff was unable to resolve an issue, they deferred to NYSED for a scoring decision. After complete books were scored, the table facilitator conducted a “read-behind” of approximately 12 completed sets of books per hour to verify accuracy of scoring. 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 appropriately. To affirm that all schools across the state adhered to scoring guidelines and policies, approximately 5 percent of the schools results are audited each year by an outside vendor.

Section V: Operational Test Data Collection and Classical Analysis

Data Collection

Operational test data were collected in several phases. During Phase 1 a sample of approximately 80% of the student test records were extracted from the Data Repository, checked by NYSED for representativeness, and delivered to CTB/McGraw-Hill. These data were used for integrity checks (data present in defined fields was in-range). Phase 2 involved extraction of close to 100% of the student test records from the Data Repository in May 2006. These data were used for classical item analysis and calibrations.

Not all test data were uploaded to the 100% Data files. For example, only public schools data were submitted to the Data Repository. Nonpublic schools were delivered in separate files to CTB/McGraw-Hill (grades 4 and 8 only) by NYSED and were not used for any data analysis. Any erroneous student records (pending resolution), or data late from school districts, was not released by the Data Repository in the 100% files, and arrived in separate 'straggler' files. Students affected by these exceptions were not included in CTB/McGraw-Hill Research's classical and IRT analyses; however, all students that participated in the NYSTP Math operational exams received scores and test results.

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. EDITCHECKER verifies that the data fields are in-range (as defined), that students' identifying information is present, and that the data is acceptable for delivery to CTB/McGraw-Hill Research. NYSED and the Data Repository were provided with the results of the checking, and some edits to the initial data were made; however, CTB Research performs data cleaning to the delivered data and excludes some student cases in order to obtain a sample of the utmost integrity. It should be noted that the major groups of cases excluded from the data set were: out-of-grade students, Limited English Proficiency (LEP) students, students whose response would not produce a valid score and students from non-public schools. Of these groups, the largest one was LEP students. This decision was based on a belief that limited English proficiency of these students might interfere with their performance on the test. Research suggests that inclusion of small special populations (e.g., students with disabilities, LEP) has little or no effect on calibration results (Karkee, Lewis, Barton, & Haug, 2002). A list of the data cleaning procedures conducted by Research and accompanying case counts is presented in Tables 6a-6f, below.

Table 6a. NYSTP Math Data Cleaning, Grade 3

Dataset	Exclusion Rule	N. Deleted	N. Cases Remain
Grade 3 100%			202,069
Grade 3 100%	Out of grade	994	201,075
Grade 3 100%	Ungraded out-of-range	0	201,075
Grade 3 100%	Duplicate student ID	0	201,075
Grade 3 100%	Duplicate string	0	201,075
Grade 3 100%	LEP = Yes & Test Language = English	16,052	185,023
Grade 3 100%	Out-of-range response	0	185,023
Grade 3 100%	Invalid Score	118	184,905
Grade 3 100%	Non-Public	1,239	183,666

Table 6b. NYSTP MA Data Cleaning, Grade 4

Dataset	Exclusion Rule	N. Deleted	N. Cases Remain
Grade 4 100%			202,093
Grade 4 100%	Out of grade	1,036	201,057
Grade 4 100%	Ungraded out-of-range	0	201,057
Grade 4 100%	Duplicate student ID	0	201,057
Grade 4 100%	Duplicate string	8	201,049
Grade 4 100%	LEP = Yes & Test Language = English	12,424	188,625
Grade 4 100%	Out-of-range response	0	188,625
Grade 4 100%	Invalid Score	192	188,433
Grade 4 100%	Non-Public	323	188,110

Table 6c. NYSTP Math Data Cleaning, Grade 5

Dataset	Exclusion Rule	N. Deleted	N. Cases Remain
Grade 5 100%			209,173
Grade 5 100%	Out of grade	1,299	207,874
Grade 5 100%	Ungraded out-of-range	0	207,874
Grade 5 100%	Duplicate student ID	2	207,872
Grade 5 100%	Duplicate string	6	207,866
Grade 5 100%	LEP = Yes & Test Language = English	10,801	197,065
Grade 5 100%	Out-of-range response	0	197,065
Grade 5 100%	Invalid Score	121	196,944
Grade 5 100%	Non-Public	1,139	195,805

Table 6d. NYSTP Math Data Cleaning, Grade 6

Dataset	Exclusion Rule	N. Deleted	N. Cases Remain
Grade 6 100%			211,392
Grade 6 100%	Out of grade	1,227	210,165
Grade 6 100%	Ungraded out-of-range	0	210,165
Grade 6 100%	Duplicate student ID	2	210,163
Grade 6 100%	Duplicate string	6	210,157
Grade 6 100%	LEP = Yes & Test Language = English	8,647	201,510
Grade 6 100%	Out-of-range response	0	201,510
Grade 6 100%	Invalid Score	236	201,274
Grade 6 100%	Non-Public	973	200,301

Table 6e. NYSTP Math Data Cleaning, Grade 7

Dataset	Exclusion Rule	N. Deleted	N. Cases Remain
Grade 7 100%			217,394
Grade 7 100%	Out of grade	654	216,740
Grade 7 100%	Ungraded out-of-range	0	216,740
Grade 7 100%	Duplicate student ID	2	216,738
Grade 7 100%	Duplicate string	6	216,732
Grade 7 100%	LEP = Yes & Test Language = English	9,610	207,122
Grade 7 100%	Out-of-range response	0	207,122
Grade 7 100%	Invalid Score	387	206,735
Grade 7 100%	Non-Public	1,139	205,596

Table 6f. NYSTP Math Data Cleaning, Grade 8

Dataset	Exclusion Rule	N. Deleted	N. Cases Remain
Grade 8 100%			219,254
Grade 8 100%	Out of grade	443	218,811
Grade 8 100%	Ungraded out-of-range	0	218,811
Grade 8 100%	Duplicate student ID	4	218,807
Grade 8 100%	Duplicate string	4	218,803
Grade 8 100%	LEP = Yes & Test Language = English	9,245	209,558
Grade 8 100%	Out-of-range response	0	209,558
Grade 8 100%	Invalid Score	694	208,864
Grade 8 100%	Non-Public	525	208,339

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 variable as it was found that the New York State population is fairly evenly split by gender categories (Males and Females).

Table 7a. Grade 3 Sample Characteristics (N=183,666)

	Demographic Category	N-count	Percent of total N
NRC	New York City	61402	33.43
	Big 4 Cities	7274	3.96
	Urban-suburban	14293	7.78
	Rural	11293	6.15
	Average Need	57797	31.47
	Low Need	29265	15.93
	Charter	1679	0.91
	(unassigned)	663	0.36
Ethnicity	Asian	11219	6.25
	Black or African-American	37646	20.97
	Hispanic or Latino	26895	14.98
	American Indian or Alaska Native	963	0.52
	Native Hawaiian/Other Pacific Islander	37	0.02
	White	101955	55.51
	Blank (no response)	13	0.01

Table 7b. Grade 4 Sample Characteristics (N=188,110)

Demographic Category		N-count	Percent of total N
NRC	New York City	63585	33.80
	Big 4 Cities	6942	3.69
	Urban-suburban	14503	7.71
	Rural	11511	6.12
	Average Need	59182	31.46
	Low Need	30420	16.17
	Charter	1353	0.72
	(unassigned)	614	0.33
Ethnicity	Asian	12533	6.66
	Black or African-American	36961	19.65
	Hispanic or Latino	32551	17.30
	American Indian or Alaska Native	950	0.51
	Native Hawaiian/Other Pacific Islander	46	0.02
	White	105065	55.85
	Blank (no response)	4	0.00

Table 7c. Grade 5 Sample Characteristics (N=195,805)

Demographic Category		N-count	Percent of total N
NRC	New York City	66274	33.85
	Big 4 Cities	7405	3.78
	Urban-suburban	14781	7.55
	Rural	11870	6.06
	Average Need	61718	31.52
	Low Need	30960	15.81
	Charter	2087	1.07
	(unassigned)	710	0.36
Ethnicity	Asian	12967	6.62
	Black or African-American	39061	19.95
	Hispanic or Latino	34642	17.69
	American Indian or Alaska Native	1017	0.52
	Native Hawaiian/Other Pacific Islander	50	0.03
	White	108064	55.19
	Blank (no response)	4	0.00

Table 7d. Grade 6 Sample Characteristics (N=200,301)

Demographic Category		N-count	Percent of total N
NRC	New York City	66895	33.40
	Big 4 Cities	7695	3.84
	Urban-suburban	15491	7.73
	Rural	12662	6.32
	Average Need	64178	32.04
	Low Need	31229	15.59
	Charter	1392	0.69
	(unassigned)	759	0.38
Ethnicity	Asian	12753	6.37
	Black or African-American	40152	20.05
	Hispanic or Latino	35503	17.72
	American Indian or Alaska Native	1110	0.55
	Native Hawaiian/Other Pacific Islander	41	0.02
	White	110741	55.29
	Blank (no response)	1	0.00

Table 7e. Grade 7 Sample Characteristics (N=205,596)

Demographic Category		N-count	Percent of total N
NRC	New York City	66891	32.54
	Big 4 Cities	8800	4.28
	Urban-suburban	15881	7.72
	Rural	13396	6.52
	Average Need	67133	32.65
	Low Need	31360	15.25
	Charter	1111	0.54
	(unassigned)	1024	0.50
Ethnicity	Asian	12399	6.03
	Black or African-American	41927	20.39
	Hispanic or Latino	35653	17.34
	American Indian or Alaska Native	1084	0.53
	Native Hawaiian/Other Pacific Islander	44	0.02
	White	114488	55.69
	Blank (no response)	1	0.00

Table 7f. Grade 8 Sample Characteristics (N=208,339)

Demographic Category		N-count	Percent of total N
NRC	New York City	67939	32.61
	Big 4 Cities	9243	4.44
	Urban-suburban	15748	7.56
	Rural	13400	6.43
	Average Need	68552	32.90
	Low Need	31400	15.07
	Charter	798	0.38
	(unassigned)	1259	0.60
Ethnicity	Asian	12501	6.00
	Black or African-American	42006	20.16
	Hispanic or Latino	35505	17.04
	American Indian or Alaska Native	1045	0.50
	Native Hawaiian/Other Pacific Islander	32	0.02
	White	117248	56.28
	Blank (no response)	2	0.00

Classical Data Analysis

Classical data analysis of the Grades 3-8 Math 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 was to occur with an item (i.e. a printing error or potentially correct distracter), item analysis is the stage that errors should be flagged and evaluated for rectification (suppression, credit, or other acceptable solution). Analyses of test level data comprise the second element of classical data analysis. These include examination of the raw score statistics (mean and standard deviation) and test reliability measures (Cronbach's alpha and Feldt-Raju coefficient). Assessment of test speededness is another important element of classical analysis. Additionally, classical Differential Item Functioning (DIF) analysis is conducted at this stage. DIF analysis includes computation of standardized mean differences and Mantel-Haenszel statistics for New York State items to identify potential item bias. All classical data analysis results contribute information on the validity and reliability of the tests (also see Sections III and VIII of this report).

Item Rescoring and Suppression

At the first stage of items analysis two Math items needed to be rescored and four items were suppressed from subsequent analyses.

Rescored Items

Grade 4, Math Book 1, Item # 2 was rescored for students taking the Chinese language version of this test. Students taking Chinese version of this test received a score of 1 on this item regardless of what they bubbled. The reason for item rescore was translation error.

Grade 5, Math Book 1, Item # 2 was rescored for students taking the Braille version of the test. Students taking the Braille version of this test should received a score of 1 on this item regardless of what they bubbled. The reason for item rescore was translation error.

Suppressed items

Grade 7, Math Book 1, items 4, 11 and 15 were suppressed from scoring due to accidental item exposure. Two of these items were measuring Number Sense and one was measuring Algebra. These items appeared in a Sample Test and students had opportunity to study them before operational test administration. All suppressed items were MC items. As a result of this suppression, the Grade 7 Math test was reduced from 38 to 35 items. The maximum test raw score decreased from 50 to 47 score points.

Grade 8, Math Book 1, item 17 (an MC item that was measuring Geometry) was suppressed from scoring due to accidental item exposure. This item appeared in a Sample Test and students had opportunity to study it before operational test administration. Forty four items remained in Grade 8 Math test after suppressing the exposed item. The maximum test raw score decreased from 69 to 68 score points for this grade.

Based on the test blueprint, the exclusion of the three items from grade 7 and the one item from grade 8 Mathematics tests did not affect the test content in any important way. The discrepancy between the target (blueprint) percent of score points and actual percent of score points is about 7% for Number Sense (grade 7), 6% for Algebra (grade 7), and 6% for Geometry (grade 8). The discrepancies between the target percent of score points and actual percent of score points for all other Content Strands are 5% or less. These small differences between target and actual percent score points indicate that suppressing previously exposed items has negligible impact on the alignment of test content maps to the test blueprint.

It should be noted that the results of the data analysis in subsequent sections of the report reflect the scoring adjustments described in this section.

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 percent of students that did not attempt the item. For MC items, “% at 0” represents the percent of students that double-bubbled responses, and other “PCT sel” categories represent the percent 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 “PCT

sel” categories depict the percent of students that 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 that 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 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 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 0.294 to 0.964. For grade 3, the item p-values were between 0.490 and 0.940 with a mean of 0.78. For grade 4, the item p-values were between 0.400 and 0.927 with a mean of 0.73. For grade 5, the item p-values were between 0.374 and 0.964 with a mean of 0.68. For grade 6, the item p-values were between 0.294 and 0.905 with a mean of 0.62. For grade 7, the item p-values were between 0.388 and 0.960 with a mean of 0.62. For grade 8, the item p-values were between 0.333 and 0.832 with a mean of 0.59. These statistics are also provided in Table 9, along with other classical test summary statistics.

Table 8a. P-values, Scored Response Distributions, and Point Bisorials, Grade 3

Item	N-count	P-value	% Omit	% at 0	PCT Sel Option 1	PCT Sel Option 2	PCT Sel Option 3	PCT Sel Option 4	Pbis Option 1	Pbis Option 2	Pbis Option 3	Pbis Option 4	Pbis Key
1	183666	0.891	0.02	0.04	6.57	*89.08	2.96	1.33	-0.20	0.29	-0.13	-0.17	0.29
2	183666	0.842	0.06	0.03	3.16	2.08	*84.12	10.55	-0.19	-0.17	0.51	-0.41	0.51
3	183666	0.864	0.07	0.04	3.49	3.26	6.83	*86.30	-0.28	-0.22	-0.24	0.44	0.44
4	183666	0.863	0.08	0.03	*86.23	3.42	5.98	4.26	0.44	-0.26	-0.25	-0.22	0.44
5	183666	0.940	0.04	0.03	1.05	2.79	*93.99	2.11	-0.20	-0.16	0.33	-0.21	0.33
6	183666	0.739	0.12	0.04	3.61	*73.81	12.01	10.42	-0.17	0.50	-0.26	-0.34	0.50
7	183666	0.600	0.06	0.06	13.19	6.99	*59.98	19.72	-0.24	-0.24	0.39	-0.13	0.39
8	183666	0.815	0.11	0.06	7.77	7.09	3.52	*81.45	-0.26	-0.32	-0.24	0.51	0.51
9	183666	0.904	0.10	0.02	4.06	*90.36	2.97	2.50	-0.31	0.43	-0.19	-0.21	0.43
10	183666	0.845	0.07	0.03	2.48	6.05	*84.45	6.94	-0.19	-0.08	0.32	-0.26	0.32
11	183666	0.704	0.13	0.04	*70.33	6.44	10.68	12.39	0.52	-0.18	-0.26	-0.35	0.52
12	183666	0.490	0.13	0.04	22.11	18.67	10.08	*48.98	-0.27	0.01	-0.23	0.36	0.36
13	183666	0.874	0.12	0.02	2.54	4.02	*87.30	6.01	-0.22	-0.24	0.47	-0.31	0.47
14	183666	0.847	0.10	0.03	5.10	5.14	*84.62	5.02	-0.24	-0.27	0.50	-0.31	0.50
15	183666	0.843	0.10	0.09	11.54	*84.26	1.94	2.08	-0.18	0.36	-0.24	-0.29	0.36
16	183666	0.588	0.22	0.03	32.43	*58.68	4.16	4.48	-0.28	0.47	-0.20	-0.27	0.47
17	183666	0.888	0.11	0.05	5.52	4.29	1.31	*88.72	-0.34	-0.23	-0.14	0.45	0.45
18	183666	0.809	0.11	0.04	13.38	2.51	3.20	*80.76	-0.30	-0.21	-0.18	0.42	0.42
19	183666	0.727	0.20	0.03	2.67	16.61	*72.53	7.97	-0.17	-0.33	0.52	-0.30	0.52
20	183666	0.828	0.17	0.04	9.43	*82.68	5.05	2.64	-0.33	0.43	-0.16	-0.19	0.43
21	183666	0.847	0.23	0.03	*84.55	3.26	6.66	5.26	0.44	-0.27	-0.24	-0.23	0.44
22	183666	0.893	0.21	0.02	1.98	1.91	*89.13	6.74	-0.25	-0.19	0.37	-0.21	0.37
23	183666	0.709	0.33	0.02	7.87	11.20	*70.65	9.93	-0.27	-0.15	0.43	-0.25	0.43
24	183666	0.819	0.46	0.02	10.49	*81.52	4.52	2.99	-0.28	0.45	-0.24	-0.20	0.45
25	183666	0.860	0.64	0.02	10.96	1.16	1.80	*85.42	-0.43	-0.22	-0.20	0.53	0.53
26	183666	0.915	0.05	3.35	10.33	86.27							
27	183666	0.621	0.12	9.24	57.31	33.34							
28	183666	0.739	0.06	0.99	31.47	12.37	55.12						
29	183666	0.765	0.07	0.58	14.00	40.74	44.61						
30	183666	0.836	0.17	9.54	13.66	76.63							
31	183666	0.646	0.19	16.89	36.96	45.95							

Table 8b. P-values, Scored Response Distributions, and Point Bisorials, Grade 4

Item	N-count	P-value	% Omit	% at 0	PCT Sel Option 1	PCT Sel Option 2	PCT Sel Option 3	PCT Sel Option 4	Pbis Option 1	Pbis Option 2	Pbis Option 3	Pbis Option 4	Pbis Key
1	188110	0.914	0.01	0.02	2.34	*91.39	4.96	1.29	-0.22	0.36	-0.23	-0.15	0.36
2	188110	0.758	0.03	0.03	3.38	8.99	*75.72	11.87	-0.09	-0.25	0.38	-0.23	0.38
3	188110	0.901	0.05	0.02	4.91	*90.01	3.17	1.84	-0.26	0.42	-0.19	-0.27	0.42
4	188110	0.839	0.07	0.03	3.41	4.86	7.75	*83.88	-0.32	-0.28	-0.26	0.52	0.52
5	188110	0.848	0.06	0.02	7.23	*84.73	4.85	3.12	-0.22	0.37	-0.21	-0.18	0.37
6	188110	0.604	0.11	0.03	23.84	9.85	*60.31	5.85	-0.44	-0.07	0.50	-0.15	0.50
7	188110	0.714	0.07	0.04	*71.38	4.65	10.43	13.43	0.54	-0.18	-0.19	-0.43	0.54
8	188110	0.844	0.13	0.04	3.01	7.77	4.79	*84.27	-0.32	-0.30	-0.20	0.49	0.49
9	188110	0.685	0.05	0.03	13.59	12.02	*68.50	5.81	-0.15	-0.34	0.45	-0.22	0.45
10	188110	0.869	0.06	0.03	5.89	*86.87	2.05	5.10	-0.30	0.35	-0.15	-0.12	0.35
11	188110	0.676	0.32	0.07	13.97	12.06	6.20	*67.38	-0.32	-0.24	-0.17	0.50	0.50
12	188110	0.708	0.11	0.03	10.63	7.08	11.48	*70.68	-0.21	-0.15	-0.24	0.40	0.40
13	188110	0.842	0.08	0.02	4.98	*84.16	4.81	5.94	-0.16	0.43	-0.21	-0.32	0.43
14	188110	0.880	0.07	0.01	3.69	4.08	*87.97	4.18	-0.29	-0.26	0.50	-0.27	0.50
15	188110	0.745	0.10	0.02	6.76	*74.45	11.17	7.51	-0.20	0.51	-0.25	-0.35	0.51
16	188110	0.400	0.19	0.04	15.82	*39.88	19.55	24.53	-0.05	0.27	-0.04	-0.23	0.27
17	188110	0.631	0.10	0.04	15.77	14.61	*63.01	6.48	-0.25	-0.11	0.38	-0.22	0.38
18	188110	0.782	0.22	0.04	8.00	9.20	4.56	*77.99	-0.20	-0.19	-0.17	0.35	0.35
19	188110	0.620	0.10	0.02	27.60	8.96	*61.90	1.42	-0.32	-0.25	0.48	-0.15	0.48
20	188110	0.840	0.13	0.02	5.18	*83.89	4.15	6.63	-0.21	0.43	-0.21	-0.28	0.43
21	188110	0.927	0.14	0.03	2.55	2.49	*92.57	2.23	-0.24	-0.22	0.39	-0.19	0.39
22	188110	0.776	0.18	0.03	*77.42	4.36	8.66	9.35	0.34	-0.23	-0.09	-0.23	0.34
23	188110	0.736	0.34	0.03	7.83	*73.39	9.49	8.92	-0.23	0.51	-0.26	-0.31	0.51
24	188110	0.837	0.20	0.03	6.55	3.73	*83.57	5.91	-0.13	-0.10	0.29	-0.25	0.29
25	188110	0.697	0.32	0.07	5.68	12.67	11.78	*69.48	-0.19	-0.18	-0.20	0.37	0.37
26	188110	0.695	0.34	0.03	11.30	12.19	6.92	*69.23	-0.14	-0.31	-0.17	0.42	0.42
27	188110	0.505	0.56	0.05	*50.25	13.93	13.93	21.28	0.23	-0.18	-0.09	-0.05	0.23
28	188110	0.668	0.45	0.03	*66.55	13.02	16.30	3.65	0.33	-0.19	-0.13	-0.23	0.33
29	188110	0.586	0.65	0.05	18.02	18.12	4.95	*58.21	-0.38	-0.25	-0.11	0.54	0.54
30	188110	0.807	0.75	0.02	*80.07	5.11	4.83	9.22	0.34	-0.16	-0.14	-0.24	0.34
31	188110	0.771	0.04	12.43	20.86	66.66							
32	188110	0.908	0.10	2.66	3.85	11.80	81.60						
33	188110	0.512	0.23	24.02	49.24	26.51							
34	188110	0.712	0.07	12.27	32.95	54.71							
35	188110	0.768	0.12	20.34	5.75	73.80							
36	188110	0.627	0.16	30.65	13.21	55.98							
37	188110	0.586	0.26	31.68	19.26	48.80							
38	188110	0.785	0.10	2.62	11.91	32.83	52.54						
39	188110	0.710	0.99	19.57	18.34	61.10							
40	188110	0.837	0.04	9.15	14.21	76.61							
41	188110	0.822	0.06	12.72	10.09	77.12							
42	188110	0.553	0.09	31.85	25.54	42.52							
43	188110	0.595	0.16	28.89	23.08	47.87							
44	188110	0.805	0.09	8.83	21.27	69.82							
45	188110	0.832	0.10	6.07	7.77	16.59	69.46						
46	188110	0.526	0.14	17.24	60.15	22.47							
47	188110	0.760	0.12	11.72	6.82	23.16	58.19						
48	188110	0.659	0.33	20.99	25.97	52.71							

Table 8c. P-values, Scored Response Distributions, and Point Bisorials, Grade 5

Item	N-count	P-value	% Omit	% at 0	PCT Sel Option 1	PCT Sel Option 2	PCT Sel Option 3	PCT Sel Option 4	Pbis Option 1	Pbis Option 2	Pbis Option 3	Pbis Option 4	Pbis Key
1	195805	0.883	0.03	0.02	4.40	3.75	*88.32	3.48	-0.26	-0.20	0.43	-0.25	0.43
2	195805	0.816	0.02	0.02	6.54	9.48	*81.56	2.38	-0.18	-0.18	0.32	-0.17	0.32
3	195805	0.964	0.01	0.01	1.68	*96.36	0.81	1.13	-0.16	0.24	-0.12	-0.14	0.24
4	195805	0.602	0.05	0.02	35.82	1.61	*60.16	2.35	-0.52	-0.15	0.57	-0.07	0.57
5	195805	0.495	0.05	0.03	1.29	*49.50	3.32	45.81	-0.10	0.23	-0.13	-0.16	0.23
6	195805	0.622	0.20	0.02	13.56	*62.08	13.94	10.20	-0.21	0.52	-0.29	-0.27	0.52
7	195805	0.701	0.03	0.02	*70.03	25.72	1.81	2.38	0.40	-0.33	-0.12	-0.15	0.40
8	195805	0.444	0.19	0.02	6.45	38.55	*44.32	10.46	-0.16	-0.28	0.42	-0.12	0.42
9	195805	0.788	0.13	0.02	3.75	12.68	*78.72	4.71	-0.20	-0.45	0.54	-0.16	0.54
10	195805	0.514	0.29	0.02	*51.23	24.27	6.92	17.27	0.50	-0.35	-0.16	-0.16	0.50
11	195805	0.772	0.16	0.02	6.07	9.55	7.12	*77.08	-0.29	-0.21	-0.26	0.48	0.48
12	195805	0.374	0.09	0.02	23.26	*37.39	8.46	30.77	-0.10	0.41	-0.11	-0.27	0.41
13	195805	0.687	0.13	0.02	17.23	9.74	4.29	*68.60	-0.24	-0.40	-0.18	0.53	0.53
14	195805	0.899	0.08	0.01	*89.79	3.81	3.13	3.18	0.40	-0.24	-0.24	-0.18	0.40
15	195805	0.586	0.09	0.02	*58.50	5.73	9.72	25.94	0.50	-0.13	-0.12	-0.41	0.50
16	195805	0.879	0.07	0.02	0.81	10.77	0.52	*87.82	-0.13	-0.22	-0.10	0.27	0.27
17	195805	0.745	0.12	0.02	15.75	5.90	3.84	*74.38	-0.44	-0.17	-0.14	0.53	0.53
18	195805	0.890	0.11	0.01	1.40	7.19	*88.87	2.42	-0.17	-0.33	0.40	-0.13	0.40
19	195805	0.696	0.12	0.02	*69.53	19.70	4.33	6.31	0.46	-0.26	-0.24	-0.24	0.46
20	195805	0.511	0.32	0.02	15.25	19.53	13.95	*50.94	-0.27	-0.27	-0.13	0.50	0.50
21	195805	0.889	0.24	0.02	4.77	*88.65	3.96	2.37	-0.17	0.31	-0.19	-0.15	0.31
22	195805	0.635	0.43	0.03	19.31	6.09	10.95	*63.19	-0.37	-0.19	-0.18	0.51	0.51
23	195805	0.706	0.33	0.02	*70.39	17.22	7.15	4.88	0.35	-0.17	-0.22	-0.19	0.35
24	195805	0.625	0.34	0.03	16.53	*62.27	10.96	9.88	-0.32	0.46	-0.25	-0.09	0.46
25	195805	0.704	0.49	0.02	8.47	*70.07	3.65	17.30	-0.25	0.49	-0.23	-0.29	0.49
26	195805	0.823	0.55	0.02	*81.86	5.05	8.00	4.54	0.31	-0.15	-0.21	-0.14	0.31
27	195805	0.767	0.11	17.21	12.20	70.47							
28	195805	0.857	0.13	4.24	9.16	11.70	74.76						
29	195805	0.453	0.19	45.79	7.74	10.98	35.30						
30	195805	0.623	0.15	28.77	17.84	53.24							
31	195805	0.659	0.08	6.15	55.94	37.83							
32	195805	0.459	0.30	40.77	26.32	32.61							
33	195805	0.669	0.24	7.91	24.04	27.22	40.59						
34	195805	0.767	0.19	5.90	15.88	20.43	57.61						

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

Item	N-count	P-value	% Omit	% at 0	PCT Sel Option 1	PCT Sel Option 2	PCT Sel Option 3	PCT Sel Option 4	Pbis Option 1	Pbis Option 2	Pbis Option 3	Pbis Option 4	Pbis Key
1	200301	0.860	0.10	0.01	1.45	2.18	10.32	*85.95	-0.17	-0.16	-0.35	0.43	0.43
2	200301	0.831	0.02	0.01	*83.08	15.88	0.70	0.32	0.33	-0.29	-0.12	-0.09	0.33
3	200301	0.720	0.08	0.01	18.58	2.40	*71.92	7.00	-0.22	-0.19	0.40	-0.26	0.40
4	200301	0.762	0.05	0.02	12.85	*76.21	5.45	5.43	-0.28	0.47	-0.23	-0.24	0.47
5	200301	0.294	0.15	0.02	*29.38	23.84	32.95	13.67	0.47	-0.06	-0.10	-0.40	0.47
6	200301	0.843	0.05	0.01	5.51	4.40	*84.21	5.82	-0.27	-0.26	0.40	-0.13	0.40
7	200301	0.668	0.07	0.01	28.07	*66.76	1.96	3.12	-0.47	0.52	-0.14	-0.09	0.52
8	200301	0.797	0.06	0.01	16.24	*79.65	2.33	1.71	-0.26	0.32	-0.15	-0.08	0.32
9	200301	0.602	0.11	0.02	24.12	7.34	*60.13	8.29	-0.39	-0.10	0.40	0.00	0.40
10	200301	0.832	0.13	0.01	*83.09	5.35	3.87	7.56	0.43	-0.24	-0.15	-0.30	0.43
11	200301	0.711	0.08	0.02	0.97	8.92	18.99	*71.02	-0.13	-0.34	-0.30	0.50	0.50
12	200301	0.752	0.05	0.01	19.12	*75.19	3.50	2.12	-0.24	0.36	-0.21	-0.18	0.36
13	200301	0.543	0.14	0.02	7.67	*54.19	16.61	21.38	-0.11	0.48	-0.18	-0.36	0.48
14	200301	0.458	0.10	0.01	9.91	22.72	*45.75	21.50	-0.21	-0.41	0.57	-0.12	0.57
15	200301	0.729	0.11	0.01	5.79	6.68	14.62	*72.80	-0.25	-0.14	-0.33	0.48	0.48
16	200301	0.905	0.07	0.01	6.91	1.92	*90.42	0.67	-0.38	-0.17	0.45	-0.12	0.45
17	200301	0.576	0.15	0.02	34.19	*57.51	4.49	3.65	-0.18	0.29	-0.12	-0.18	0.29
18	200301	0.451	0.36	0.02	26.20	*44.94	18.34	10.15	-0.26	0.35	-0.13	-0.03	0.35
19	200301	0.603	0.10	0.01	14.20	15.43	*60.20	10.05	-0.13	-0.39	0.43	-0.08	0.43
20	200301	0.633	0.22	0.02	14.26	8.20	*63.12	14.18	-0.28	-0.22	0.47	-0.19	0.47
21	200301	0.793	0.16	0.02	14.93	*79.22	3.20	2.48	-0.34	0.41	-0.15	-0.12	0.41
22	200301	0.605	0.22	0.02	14.86	9.19	*60.39	15.32	-0.38	-0.06	0.52	-0.27	0.52
23	200301	0.839	0.24	0.02	4.90	5.05	6.11	*83.67	-0.18	-0.27	-0.30	0.47	0.47
24	200301	0.388	0.27	0.01	50.48	6.04	*38.73	4.46	-0.14	-0.24	0.32	-0.14	0.32
25	200301	0.876	0.28	0.01	7.98	*87.32	3.45	0.96	-0.23	0.36	-0.25	-0.14	0.36
26	200301	0.720	0.12	21.07	13.86	64.94							
27	200301	0.628	0.27	32.45	9.29	57.99							
28	200301	0.477	0.50	40.45	23.11	35.95							
29	200301	0.704	0.27	12.40	16.50	18.25	52.58						
30	200301	0.754	0.25	16.62	15.79	67.34							
31	200301	0.409	0.95	51.77	13.59	33.70							
32	200301	0.720	0.31	13.18	29.54	56.98							
33	200301	0.374	0.56	46.39	18.82	9.88	24.35						
34	200301	0.586	0.38	24.48	18.83	12.57	43.75						
35	200301	0.395	0.50	27.33	45.01	8.55	18.61						

Table 8e. P-values, Scored Response Distributions, and Point Bisorials, Grade 7

Item	N-count	P-value	% Omit	% at 0	PCT Sel Option 1	PCT Sel Option 2	PCT Sel Option 3	PCT Sel Option 4	Pbis Option 1	Pbis Option 2	Pbis Option 3	Pbis Option 4	Pbis Key
1	205596	0.854	0.06	0.01	7.18	3.97	3.42	*85.36	-0.23	-0.25	-0.15	0.38	0.38
2	205596	0.945	0.04	0.01	*94.42	2.94	1.31	1.28	0.29	-0.18	-0.17	-0.15	0.29
3	205596	0.694	0.09	0.01	4.48	20.49	*69.30	5.64	-0.20	-0.21	0.35	-0.16	0.35
5	205596	0.829	0.20	0.01	2.47	7.20	*82.73	7.39	-0.17	-0.22	0.44	-0.31	0.44
6	205596	0.825	0.05	0.01	1.29	6.20	10.00	*82.45	-0.13	-0.23	-0.22	0.36	0.36
7	205596	0.935	0.02	0.01	1.22	3.07	*93.51	2.17	-0.14	-0.13	0.22	-0.11	0.22
8	205596	0.727	0.08	0.01	13.63	4.13	*72.63	9.52	-0.34	-0.16	0.49	-0.24	0.49
9	205596	0.649	0.14	0.01	11.63	*64.85	5.15	18.21	-0.36	0.29	-0.22	0.06	0.29
10	205596	0.649	0.07	0.01	6.88	23.37	*64.88	4.78	-0.15	-0.17	0.32	-0.19	0.32
12	205596	0.874	0.05	0.02	3.29	3.18	6.11	*87.35	-0.25	-0.19	-0.29	0.44	0.44
13	205596	0.651	0.05	0.02	4.55	*65.10	8.70	21.58	-0.20	0.48	-0.14	-0.36	0.48
14	205596	0.504	0.08	0.02	*50.38	36.57	4.97	7.98	0.45	-0.28	-0.16	-0.21	0.45
16	205596	0.514	0.09	0.01	18.66	4.95	24.98	*51.32	-0.40	-0.20	-0.07	0.47	0.47
17	205596	0.842	0.03	0.01	13.15	*84.21	1.79	0.81	-0.34	0.39	-0.13	-0.11	0.39
18	205596	0.624	0.09	0.02	5.60	4.57	27.36	*62.36	-0.09	-0.22	-0.23	0.35	0.35
19	205596	0.432	0.19	0.02	19.26	*43.14	25.17	12.22	-0.23	0.36	-0.11	-0.13	0.36
20	205596	0.771	0.08	0.02	5.93	9.26	7.72	*77.00	-0.24	-0.28	-0.21	0.47	0.47
21	205596	0.960	0.04	0.00	0.95	*95.96	1.29	1.75	-0.14	0.27	-0.17	-0.16	0.27
22	205596	0.649	0.17	0.01	5.60	25.51	*64.82	3.89	-0.19	-0.38	0.50	-0.15	0.5
23	205596	0.736	0.11	0.01	*73.55	9.03	13.08	4.22	0.41	-0.22	-0.23	-0.21	0.41
24	205596	0.468	0.19	0.02	10.05	25.69	*46.74	17.32	-0.20	-0.18	0.42	-0.18	0.42
25	205596	0.460	0.16	0.02	15.97	29.36	8.59	*45.90	-0.13	-0.10	-0.17	0.28	0.28
26	205596	0.833	0.13	0.01	1.80	7.62	*83.15	7.28	-0.13	-0.02	0.23	-0.24	0.23
27	205596	0.887	0.11	0.02	8.69	*88.62	2.07	0.50	-0.15	0.25	-0.21	-0.11	0.25
28	205596	0.786	0.15	0.01	11.63	1.57	*78.44	8.19	-0.22	-0.17	0.43	-0.31	0.43
29	205596	0.600	0.28	0.02	19.32	*59.80	9.58	11.01	0.00	0.29	-0.22	-0.25	0.29
30	205596	0.644	0.26	0.02	8.28	*64.28	7.97	19.19	-0.21	0.46	-0.19	-0.28	0.46
31	205596	0.630	0.32	17.30	39.17	43.21							
32	205596	0.472	0.57	29.62	45.85	23.96							
33	205596	0.547	0.80	23.66	25.09	13.61	36.84						
34	205596	0.388	0.62	36.68	31.57	9.40	21.74						
35	205596	0.444	1.37	43.63	22.43	32.57							
36	205596	0.576	0.54	34.37	15.69	49.40							
37	205596	0.425	0.39	5.65	71.69	11.35	10.91						
38	205596	0.481	0.45	30.46	22.05	19.38	27.66						

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

Item	N-count	P-value	% Omit	% at 0	PCT Sel Option 1	PCT Sel Option 2	PCT Sel Option 3	PCT Sel Option 4	Pbis Option 1	Pbis Option 2	Pbis Option 3	Pbis Option 4	Pbis Key
1	208339	0.626	0.24	0.03	13.65	9.15	*62.44	14.50	-0.08	-0.24	0.37	-0.23	0.37
2	208339	0.790	0.04	0.02	7.28	4.81	8.89	*78.96	-0.22	-0.20	-0.21	0.39	0.39
3	208339	0.646	0.16	0.02	24.80	6.93	*64.51	3.59	-0.26	-0.25	0.43	-0.15	0.43
4	208339	0.589	0.12	0.01	23.77	15.57	*58.79	1.73	-0.17	-0.34	0.42	-0.08	0.42
5	208339	0.526	0.14	0.01	17.76	4.53	24.99	*52.56	-0.37	-0.22	-0.18	0.54	0.54
6	208339	0.631	0.04	0.02	0.99	*63.12	33.81	2.03	-0.12	0.43	-0.37	-0.13	0.43
7	208339	0.333	0.18	0.01	8.52	52.97	*33.20	5.12	-0.27	-0.07	0.35	-0.25	0.35
8	208339	0.615	0.10	0.02	4.44	*61.44	6.70	27.30	-0.15	0.44	-0.24	-0.27	0.44
9	208339	0.662	0.21	0.01	6.32	6.19	*66.04	21.22	-0.16	-0.27	0.49	-0.31	0.49
10	208339	0.585	0.08	0.02	10.05	*58.44	7.85	23.56	-0.20	0.54	-0.14	-0.39	0.54
11	208339	0.832	0.12	0.01	7.19	*83.10	3.34	6.24	-0.16	0.34	-0.20	-0.21	0.34
12	208339	0.820	0.14	0.02	3.17	7.77	*81.91	7.00	-0.17	-0.21	0.40	-0.27	0.40
13	208339	0.698	0.12	0.01	11.88	*69.71	6.35	11.93	-0.17	0.38	-0.23	-0.19	0.38
14	208339	0.683	0.07	0.01	2.12	11.83	*68.27	17.70	-0.16	-0.12	0.27	-0.16	0.27
15	208339	0.712	0.07	0.01	*71.14	17.14	8.33	3.31	0.41	-0.28	-0.18	-0.17	0.41
16	208339	0.564	0.14	0.01	21.45	13.07	*56.34	8.99	-0.15	-0.26	0.36	-0.11	0.36
18	208339	0.723	0.14	0.01	*72.15	5.68	13.24	8.78	0.53	-0.24	-0.26	-0.33	0.53
19	208339	0.639	0.10	0.01	4.44	4.94	*63.89	26.62	-0.23	-0.29	0.56	-0.36	0.56
20	208339	0.712	0.10	0.01	18.30	7.23	*71.17	3.18	-0.22	-0.30	0.44	-0.22	0.44
21	208339	0.682	0.09	0.02	5.54	6.12	20.09	*68.14	-0.18	-0.17	-0.33	0.46	0.46
22	208339	0.685	0.17	0.02	4.20	19.41	7.83	*68.37	-0.24	-0.30	-0.27	0.51	0.51
23	208339	0.734	0.14	0.02	6.14	6.99	*73.34	13.38	-0.21	-0.28	0.50	-0.29	0.50
24	208339	0.588	0.35	0.02	7.31	14.39	19.35	*58.59	-0.15	-0.26	-0.18	0.41	0.41
25	208339	0.765	0.18	0.03	4.84	*76.41	11.27	7.28	-0.14	0.45	-0.30	-0.25	0.45
26	208339	0.532	0.20	0.02	30.30	*53.10	9.92	6.46	-0.15	0.42	-0.31	-0.19	0.42
27	208339	0.609	0.29	0.03	18.69	7.77	12.54	*60.68	-0.32	-0.27	-0.20	0.54	0.54
28	208339	0.491	0.37	36.43	28.67	34.54							
29	208339	0.642	0.61	20.03	11.44	23.87	44.05						
30	208339	0.664	0.32	21.78	23.41	54.49							
31	208339	0.422	2.99	39.96	17.66	12.92	26.47						
32	208339	0.505	1.29	30.92	35.93	31.85							
33	208339	0.581	0.84	12.42	58.31	28.43							
34	208339	0.660	1.11	19.31	28.57	51.01							
35	208339	0.587	1.65	27.77	25.77	44.82							
36	208339	0.434	1.96	39.34	17.28	14.01	27.41						
37	208339	0.687	0.68	17.18	27.91	54.24							
38	208339	0.551	1.96	33.33	21.33	43.39							
39	208339	0.513	0.88	29.38	18.76	19.28	31.71						
40	208339	0.496	2.56	28.96	40.28	28.20							
41	208339	0.550	2.34	34.00	19.89	43.77							
42	208339	0.590	1.44	10.68	28.93	31.39	27.56						
43	208339	0.559	1.04	15.88	25.47	32.27	25.35						
44	208339	0.545	2.00	36.00	17.11	44.89							
45	208339	0.632	2.04	28.57	14.91	54.48							

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 that responded correctly also tend to do well on the overall test. For incorrect answer options (distracters), the point biserial should be negative, which indicates that students who scored lower on the overall test had a tendency to pick a distracter. No item answer keys were flagged for point biserials on any of the Grades 3-8 Math Tests. Point biserials for correct answer options (pbis*) on the tests ranged from 0.22 to 0.57. For grade 3, the pbis* were between 0.29 and 0.53. For grade 4, the pbis* were between 0.23 and 0.54. For grade 5, the pbis* were between 0.23 and 0.57. For grade 6, pbis* were between 0.29 and 0.57. For grade 7, the pbis* were between 0.22 and 0.52. For grade 8, the pbis* were between 0.27 and 0.56.

Distracter Analysis

Item distracters provide additional information on student performance on test questions. Two types of information on item distracters are available from New York State test data: information on proportion of students selecting incorrect item response options and the point biserial coefficient of distracters (discrimination power of incorrect answer choice). The proportions of students selecting incorrect responses while responding to MC items are provided in Tables 8a-8f of this report. 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 Math items except 3 items: 5 and 24 on the grade 6 test and item 7 on the grade 8 test.

Approximately 29% of students answered grade 6 item 5 correctly while close to 33% of students selected a single incorrect option 3. Answer choices on this item were examined and no content/key problem was identified. This item was also found to have a good discrimination power with a point biserial of 0.47. Approximately 39% of students answered grade 6 item 24 correctly while close to 50% of students selected a single incorrect option 1. Answer choices on this item were examined and no content/key problem was identified. This item was also found to have a good discrimination power with a point biserial of 0.32. Approximately 33% of students answered grade 8 item 7 correctly while close to 53% of students selected a single incorrect option 2. Answer choices on this item were examined and no content/key problem was identified. This item was also found to have a good discrimination power with a point biserial of 0.35.

As mentioned in the Point Biserial Correlations subsection, items are flagged if the point biserial of any distracter is positive. One grade 3 item was flagged for positive point biserial values on a distracter (incorrect) answer option (item 12, 0.01). One grade 7 item

was flagged for positive point biserial values on distracter (incorrect) answer options (item 9, 0.06). No test items were flagged for point biserials of both a distracter and the correct answer option. None of the point biserials of distracter options on any 2006 NYSTP Math test exceeded the point biserial of the corresponding answer key option. There were no flags for point biserials of distracters in grades 4, 5, 6, and 8.

Test Statistics and Reliability Coefficients

Test statistics including raw score mean and standard deviation are presented in Table 9, below. 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 Math 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 from 0.89 to 0.95. Feldt-Raju reliability coefficients ranged from 0.89 to 0.96. 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 met or exceeded 0.80, across statistics, which is a good indication that the NYSTP 3-8 Math Tests are acceptably reliable. High reliability indicates that scores are consistent and not unduly influenced by random error. More information on test reliability and standard error of measurement is provided in Section VIII (Reliability) of this report.

Table 9. NYSTP Math 2006 Test Form Statistics and Reliability

Grade	Max RS	RS Mean	RS SD	P-value Mean	Minimum P-value	Maximum P-value	Cronbach Alpha	Feldt-Raju Alpha
3	39	30.53	6.88	0.78	0.49	0.94	0.89	0.89
4	70	51.07	14.09	0.73	0.40	0.93	0.94	0.94
5	46	31.44	9.62	0.68	0.37	0.96	0.90	0.91
6	49	30.59	11.16	0.62	0.29	0.91	0.91	0.92
7	47	29.03	9.60	0.62	0.39	0.96	0.89	0.91
8	68	40.01	17.14	0.59	0.33	0.83	0.95	0.96

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, we want all scores to be based on actual student performance, and 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 Math 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 & Blestein, 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 .10 and .19, inclusive. A large amount of practically significant DIF is represented by an SMD with an absolute value of .20 or greater. Then, the Mantel-Haenszel method was 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 method has a critical value of 6.63 (degrees of freedom = 1 for MC items; alpha = .01), and is compared to its corresponding Delta-value (significant when absolute value of Delta > 1.50) to factor in effect size (Zwick, Donoghue, & 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 & Jones, 1993).

Classical DIF analyses were conducted on subgroups of Need Resource Category (focal group: High Needs; reference group: Low Needs), gender (focal group: Female; reference group: Male), ethnicity (focal groups: Black or African-American, Hispanic or Latino, and Asian); reference group: White) and test language (focal group: Spanish; reference group: English). The minimum sample size for a focal group (the subgroup to be compared to the reference, or 'majority' group) in these analyses was 500. A random sample of 7,000 student records was used to compute DIF. If a focal group's case count fell below 500, the group was augmented with extra cases from the dataset. Table 10 shows the percent of items exhibiting DIF. For details on DIF items, please refer to Appendix D

Table 10. NYSTP Math 2006 Classical DIF Sample N-Counts

Grade	Ethnicity				Gender		Need Resource Category		Test Language	
	Black or African-American	Hispanic or Latino	Asian	White	Female	Male	High	Low	Spanish	English
3	1523	1346	500	3960	3580	3749	3848	3406	500	6846
4	1381	1400	503	3984	3607	3661	3819	3394	500	6881
5	1457	1404	504	3930	3661	3634	3830	3364	500	6893
6	1467	1480	503	3932	3616	3766	3931	3374	500	6912
7	1488	1406	505	4032	3609	3822	3958	3397	500	6899
8	1447	1395	505	4070	3707	3710	3898	3447	500	6897

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	14
5	12
6	11
7	18
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

IRT Models and Rationale for Use

Item response theory (IRT) allows comparisons among items and scale scores, even those from different test forms, by using a common scale for all items and examinees (i.e., as if there were a hypothetical test that contained items from all forms). The three-parameter logistic (3PL) model (Lord & Novick, 1968; Lord, 1980) was used to analyze item responses on the multiple choice items. For analysis of the constructed response 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 multiple choice 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 (Lord & Novick, 1968; Lord, 1980) was used in the analysis of MC items. In this model, the probability that a student with ability θ responds correctly to item i is

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

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

For analysis of the constructed response items, the two-parameter partial credit (2PPC) model (Muraki, 1992; Yen, 1993) was used. The 2PPC model is a special case of Bock's (1972) nominal model. Bock's model states that the probability of an examinee with ability θ having a score $(k - 1)$ at the k -th level of the j -th item is

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

where

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

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

$$C_{jk} = -\sum_{i=0}^{k-1} \gamma_{ji}, \quad \text{where } \gamma_{j0} = 0,$$

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

Calibration Sample

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

Calibration Process

The IRT model parameters were estimated using CTB/McGraw-Hill's PARDUX software (Burket, 2002). PARDUX estimates parameters simultaneously for MC and CR items using marginal maximum likelihood procedures implemented via the EM (expectation-maximization) 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 Math tests did not incur anything problematic during item calibration. The number of estimation cycles was set to 50 for all grades. The estimated parameters were in the original theta metric and all of the items were well within the prescribed parameter ranges. The b ('difficulty') parameter ranges were reasonable, with a skew that reflects the

generally high p-values present in the NYSTP 2006 Math item analysis. When the PARDUX program encounters difficulty estimating the c ('guessing') parameter, it assigns a default c parameter value of 0.2000. While it is perfectly normal to expect some default c estimates, a reasonableness check is conducted to make sure that there are not an excessive amount of test items with default c parameter. For the Grades 3-8 Math Tests, all calibration estimation results are reasonable.

Table 12. NYSTP Math 2006 Calibration Results

Grade	Largest 'a' parameter	'b' parameter range		# items with Default 'c'	Theta Mean	Theta Standard Deviation	N students
3	2.181	-3.432	0.499	3	0.07	1.301	183666
4	2.137	-3.352	1.870	0	-0.06	1.174	188110
5	2.397	-4.014	1.193	1	0.03	1.185	195805
6	2.439	-3.759	2.321	0	-0.02	1.184	200301
7	2.383	-4.363	1.708	4	-0.10	1.164	205596
8	2.812	-1.639	2.651	0	0.02	1.175	208339

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_{lj} was assumed to have approximately a chi-square distribution with the following degree of freedom:

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

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

To adjust for differences in degrees of freedom among items, Q_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 obtained Z_{Q1} was greater than the value of Z_{Q1} critical. If the obtained Z_{Q1} was less than Z_{Q1} critical the items were rated as having acceptable fit. It should be noted that most items in the NYSTP 2006 Math test demonstrated good model fit, further supporting use of the chosen models. No items in grades 3 or 5 exhibited poor item-model fit statistics. The following items exhibited misfit: grade 4 item 37 ($Z_{Q1} = 643.34$, Z_{Q1} critical = 496.10), grade 6 items 22 ($Z_{Q1} = 571.46$, Z_{Q1} critical = 527.90) and 30 ($Z_{Q1} = 770.12$, Z_{Q1} critical = 526.56), grade 7 item 35 ($Z_{Q1} = 693.99$, Z_{Q1} critical = 539.08), and grade 8 items 33 ($Z_{Q1} = 1323.18$, Z_{Q1} critical = 546.46) and 42 ($Z_{Q1} = 1956.51$, Z_{Q1} critical = 543.14). Fit statistics and status for all items in the Grades 3-8 Math 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, a student's response on one item is not dependent upon their response to another item. Statistically speaking, when a student's ability is accounted for, their response to each item is statistically independent.

One way to measure the statistical independence of items within a test is via the Q_3 statistic (Yen, 1984). This statistic was obtained by correlating differences between students' observed and expected responses for pairs of items after taking into account overall test performance. The Q_3 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 3-8 Math Tests and only a few pairs of items were found to be locally dependent. Grade 3 items 6 & 11 and items 4 & 21 (both MC items from the same Content Strand and PI) were found to be locally dependent ($Q_3 = 0.237$ and 0.240 , respectively). Grade 4 items 7 & 15, MC items from different Content Strands and PIs, were found to be locally dependent ($Q_3 = 0.240$). Grade 6 items 7 & 11 (both MC items from the same Content Strand and PI) were found to be locally dependent ($Q_3 = 0.251$). Grade 7 items 13 & 18 (both MC items from the same Content Strand) were found to be locally dependent ($Q_3 = 0.232$). Grade 8 items 28 & 39 (an SR and ER item from the same Content Strand and PI) were found to be locally dependent ($Q_3 = 0.232$). No items from grade 5 were found to be locally dependent. The frequency and magnitude of these statistics were not sufficient to warrant concern.

Scaling

The scaling of the Grades 3-8 Math Tests was conducted in two stages: initial scaling during which preliminary item parameters were estimated and preliminary scoring tables were developed, and final scaling during which the tests were rescaled to align the Level III cut across grades and final scoring tables were developed. Preliminary item parameters were used to evaluate items, order items in terms of their difficulty for the purpose of standard setting. Preliminary scoring tables were used to produce scale score frequency distribution used for impact data during the standard setting process. Final item parameters were used to produce final raw score to scale score conversion tables.

Initial Scaling

Temporary and arbitrary transformation constants were used to transform the New York State Math item parameters in the original theta metric estimated during the item calibration process to the scale score metric. These constants are presented in Table 13.

Table 13. NYSTP Math 2006 Initial Transformation Constants

Grade	<i>M1</i>	<i>M2</i>
3	30	450
4	30	500
5	30	550
6	30	600
7	30	650
8	30	700

The item parameters in a scale score (SS) metric were obtained using the following procedures implemented by the PARDUX program:

$$\begin{aligned}A_{ss} &= a_{\theta} / M1 \\B_{ss} &= M1 * b_{\theta} + M2 \\F_{ss} &= f_{\theta} / M1 \\G_{ss} &= g_{\theta} + (f_{\theta} / M1) * M2 \\C_{ss} &= c_{\theta}\end{aligned}$$

where:

A_{ss} is a discrimination parameter in scale score metric for MC items

B_{ss} is a difficulty parameter in scale score metric for MC items

F_{ss} is a discrimination parameter in scale score metric for CR items

G_{ss} is a difficulty for category m_j in scale score metric for CR items

a_{θ} is a discrimination parameter in the original theta metric for MC items

b_{θ} is a difficulty parameter in the original theta metric for MC items

f_{θ} is a discrimination parameter in the original theta metric for CR items

g_{θ} is a difficulty level for category m_j in the original theta metric for CR items

C_{ss} and c_{ss} is a guessing parameter in the original theta metric

In the 2PPC model, f (alpha) and g (gamma) are analogous to b and a , where alpha is the discrimination parameter and gamma over alpha (g/f) is the location where adjacent trace lines cross on the ability scale. Because of the different metrics used, the 3PL (multiple-choice) parameters b and a are not directly comparable to the 2PPC parameters f and g , however they can be converted to a common metric. The two metrics are related by $b = g/f$ and $a = f / 1.7$ (Burket, 2002). As a result of this procedure, the MC and CR items are placed on the same scale. Note that for the 2PPC model there are $m_j - 1$ (where m_j is a score level j) independent g 's and one f , for a total of m_j independent parameters estimated for each item while there is one a and one b per item in the 3PL model.

The scale score parameters were used to produce temporary Raw Score to Scale Score conversion tables for Standard Setting. Detailed process of scoring table development is presented in Scoring Method subsection.

Final Scaling

It was decided by NYSED to establish a single ‘Meets Learning Standards’ cut score (or the minimum scale score needed to demonstrate proficiency) across grades. Although the scales are distinct and unique, each student was deemed proficient if they met or exceeded the cut score of 650 (also called the Level III cut). In order to maintain the psychometric properties of the scales, and avoid undue influence on the standard setting process, rescaling was conducted after standard setting. In a process of rescaling the Level III cut scores established during the standard setting were rescale to 650 and a common standard deviation of 40 was set across grades using the following process:

- 1. The Level III cut score from the Bookmark Standard Setting was standardized with respect to the temporary test mean.**

$$X = \frac{(Cut_{Old} - Mean_{Old})}{SD_{Old}}$$

where

X is a standardized value

Cut_{Old} is a Level III cut from the Standard Setting (on a temporary scale)

$Mean_{Old}$ is test mean on a temporary scale

SD_{Old} is a test standard deviation on a temporary scale

- 2. The standardized value (X) was used to calculate the new mean with respect to the new cut (650).**

$$Mean_{new} = Cut_{new} - SD_{new} * X$$

where

$Mean_{new}$ is a test mean on the final scale

Cut_{new} is 650 (on the final scale)

SD_{new} is 40 (on the final scale)

- 3. The scaling constants K_1 and K_2 were calculated using new and old test means and standard deviations.**

$$K_1 = \frac{SD_{New}}{SD_{Old}}$$

$$K_2 = (Mean_{New} - \frac{SD_{New}}{SD_{Old}} Mean_{Old})$$

4. Final transformation parameters $M1$ and $M2$ were derived.

$$M1_{new} = K_1 * M1_{Old}$$

$$M2_{new} = K_1 * M2_{Old} + K_2.$$

where

$M1_{Old}$ and $M2_{Old}$ are temporary transformation parameters (as presented in Table 13)

The final transformation parameters $M1_{new}$ and $M2_{new}$ were used to transform item parameters obtained in a calibration process into the final scale score metric. The transformation process was described in details in Initial Scaling section. Table 14 presents the final transformation parameters for New York State Grades 3-8 Math Tests.

Table 14. NYSTP Math 2006 Final Transformation Constants

Grade	$M1_{new}$	$M2_{new}$
3	30.8862	678.8271
4	33.3671	681.1426
5	33.6465	666.8232
6	33.2183	658.8582
7	33.2414	656.6483
8	33.1181	653.3118

Following rescaling of the Level III cut, the remaining proficiency cuts (Level II and Level IV) set during the Standard Setting were adjusted accordingly using the same final transformation constants (from Table 14) and the following procedure:

1. Temporary Level II and Level IV cut scores in scale score metric were transformed back to the original theta.

$$Cut_{\theta} = (Cut_{Old} - M2_{Old}) / M1_{Old} \text{ where}$$

Cut_{θ} is the cut score (Level II or Level IV) in a theta metric, and

Cut_{Old} is the temporary cut score (Level II or Level IV) in a scale score metric.

2. The cut scores in the original theta metric were transformed to the final scale score metric using the final transformation constants.

$$Cut_{New} = Cut_{\theta} * M1_{new} + M2_{new} \text{ where}$$

Cut_{New} is the final cut score (Level II or Level IV) in a scale score metric.

This procedure of cut score transformation preserved the standard setting impact data associated with the Math proficiency cut scores.

Item Parameters

As previously discussed, the item parameters were estimated by the software PARDUX (Burket, 2002) and were rescaled after standard setting to allow for NYSED to implement 650 as the Level III cut score across all grades. The item parameters were rescaled using the procedure and final scaling constants presented in the Final Scaling section. Again, PARDUX was used to perform these transformations. The final item parameters (after rescaling) are presented in Tables 15a-15f. Descriptions of what each of the parameter variables mean is presented in the subsection depicting the IRT models and rationale.

Table 15a. Grade 3 2006 Operational Item Parameter Estimates

Item	Max Pts	a par/ alpha	b par/ gamma1	c par/ gamma2	gamma3
1	1	0.01833	602.1477	0.0749	
2	1	0.03816	641.2468	0.1512	
3	1	0.03129	634.0283	0.1787	
4	1	0.03154	635.1752	0.2000	
5	1	0.02552	599.7139	0.1062	
6	1	0.03358	654.4322	0.1042	
7	1	0.02077	666.7701	0.0499	
8	1	0.03863	647.4078	0.1835	
9	1	0.03207	621.8970	0.1131	
10	1	0.01821	623.2629	0.2000	
11	1	0.03823	660.6219	0.1105	
12	1	0.02664	689.8462	0.1438	
13	1	0.03294	629.1395	0.0708	
14	1	0.03603	637.9692	0.1114	
15	1	0.02104	624.5256	0.1062	
16	1	0.03919	676.7733	0.1488	
17	1	0.03241	625.9854	0.0924	
18	1	0.02465	634.9525	0.0481	
19	1	0.03658	657.7173	0.1191	
20	1	0.02519	631.3907	0.0427	
21	1	0.03113	638.5721	0.2000	
22	1	0.02411	613.9880	0.0513	
23	1	0.03064	662.9405	0.2028	
24	1	0.02618	634.0276	0.0227	
25	1	0.04155	639.4362	0.1335	
26	2	0.03411	20.6935	20.6732	
27	2	0.02795	16.6440	19.6544	
28	3	0.03089	16.3752	21.4484	19.6202
29	3	0.02550	13.1147	15.8113	17.3034
30	2	0.03048	19.4687	18.8135	
31	2	0.02707	17.1302	18.2557	

Table 15b. Grade 4 2006 Operational Item Parameter Estimates

Item	Max Pts	a par/ alpha	b par/ gamma1	c par/ gamma2	gamma3
1	1	0.02333	608.7155	0.2139	
2	1	0.01936	646.0087	0.1898	
3	1	0.02550	610.9609	0.0551	
4	1	0.03496	639.8271	0.1832	
5	1	0.01888	614.7547	0.0447	
6	1	0.02960	672.1873	0.1072	
7	1	0.03364	658.2379	0.1222	
8	1	0.02877	631.1812	0.0830	
9	1	0.02331	658.1154	0.1094	
10	1	0.01833	609.2901	0.0987	
11	1	0.03767	670.4069	0.2259	
12	1	0.03674	677.3143	0.3944	
13	1	0.02358	629.8813	0.1719	
14	1	0.03341	627.2582	0.1260	
15	1	0.03060	652.9692	0.1329	
16	1	0.02899	715.2779	0.2225	
17	1	0.02421	677.8253	0.2493	
18	1	0.02105	652.9091	0.3325	
19	1	0.03159	674.0847	0.1690	
20	1	0.02376	630.8004	0.1719	
21	1	0.02653	606.8107	0.1719	
22	1	0.01607	633.3177	0.1349	
23	1	0.03304	657.9217	0.1809	
24	1	0.01476	612.7578	0.1508	
25	1	0.02701	674.6913	0.3465	
26	1	0.02613	667.0628	0.2486	
27	1	0.03408	713.4182	0.3604	
28	1	0.02266	678.7620	0.3309	
29	1	0.03531	674.3895	0.0888	
30	1	0.01894	641.7502	0.2853	
31	2	0.04627	29.3534	30.1367	
32	3	0.02300	14.0206	13.8197	13.5445
33	2	0.03095	19.8598	22.0110	
34	2	0.04285	26.8479	28.6626	
35	2	0.02581	18.1697	14.9421	
36	2	0.02978	20.5413	18.9323	
37	2	0.02156	14.8495	13.8466	
38	3	0.02941	17.0292	18.3176	19.5851
39	2	0.02670	17.6103	16.9651	
40	2	0.04245	26.7429	26.8413	

(Continued on next page)

Table 15b. Grade 4 2006 Operational Item Parameter Estimates (cont.)

Item	Max Pts	a par/ alpha	b par/ gamma1	c par/ gamma2	gamma3
41	2	0.03109	20.3625	18.9311	
42	2	0.03660	24.5013	24.6935	
43	2	0.04585	30.4909	30.7674	
44	2	0.04900	30.5103	31.8320	
45	3	0.01931	12.1140	11.9831	11.6698
46	2	0.02915	18.0866	21.1132	
47	3	0.03308	21.6906	20.6333	21.6550
48	2	0.04907	31.9119	32.8345	

Table 15c. Grade 5 2006 Operational Item Parameter Estimates

Item	Max Pts	a par/ alpha	b par/ gamma1	c par/ gamma2	gamma3
1	1	0.03021	611.7228	0.0889	
2	1	0.01990	631.9633	0.3450	
3	1	0.02457	570.7358	0.1410	
4	1	0.03908	659.9354	0.0791	
5	1	0.01177	696.0367	0.2000	
6	1	0.03537	660.1602	0.1332	
7	1	0.02131	645.6419	0.1576	
8	1	0.03739	685.5920	0.1634	
9	1	0.04191	639.2383	0.1625	
10	1	0.03420	672.1648	0.1046	
11	1	0.02837	634.4818	0.1022	
12	1	0.02812	690.2507	0.0825	
13	1	0.03158	647.8414	0.0773	
14	1	0.02829	604.1620	0.0429	
15	1	0.02857	660.5015	0.0707	
16	1	0.01619	592.7010	0.1776	
17	1	0.03118	638.4744	0.0598	
18	1	0.02856	610.0720	0.1284	
19	1	0.02605	647.8609	0.1404	
20	1	0.04186	675.4148	0.1516	
21	1	0.01909	593.2549	0.0786	
22	1	0.03457	659.7062	0.1520	
23	1	0.02160	655.3851	0.2943	
24	1	0.03472	666.0500	0.2304	
25	1	0.03540	653.6561	0.2141	
26	1	0.01781	621.3362	0.2527	

(Continued on next page)

Table 15c. Grade 5 2006 Operational Item Parameter Estimates (cont.)

Item	Max Pts	a par/ alpha	b par/ gamma1	c par/ gamma2	gamma3
27	2	0.03970	25.6936	24.6437	
28	3	0.01855	10.9168	11.7715	10.4910
29	3	0.01855	13.9310	12.0708	11.5055
30	2	0.03129	20.8061	19.9516	
31	2	0.03819	22.2423	26.0790	
32	2	0.02568	17.3444	17.2291	
33	3	0.02544	15.0844	16.5879	16.8087
34	3	0.03038	18.0795	19.4685	19.3423

Table 15d. Grade 6 2006 Operational Item Parameter Estimates

Item	Max Pts	a par/ alpha	b par/ gamma1	c par/ gamma2	gamma3
1	1	0.02909	608.6798	0.1129	
2	1	0.01893	609.1229	0.2236	
3	1	0.02060	629.8523	0.1099	
4	1	0.03133	634.5873	0.2235	
5	1	0.04310	686.5806	0.0632	
6	1	0.02554	611.9234	0.1677	
7	1	0.03758	648.4501	0.1846	
8	1	0.03201	649.0620	0.4999	
9	1	0.01758	642.7670	0.0199	
10	1	0.03116	624.5643	0.2903	
11	1	0.03612	643.8658	0.2164	
12	1	0.02331	639.8581	0.3255	
13	1	0.02469	655.6049	0.0443	
14	1	0.04318	667.7565	0.0766	
15	1	0.02780	636.0981	0.1787	
16	1	0.04225	606.5135	0.1629	
17	1	0.04105	682.3153	0.4067	
18	1	0.02734	683.7633	0.2087	
19	1	0.02259	655.4062	0.1787	
20	1	0.03188	654.6285	0.2096	
21	1	0.03617	642.9141	0.4255	
22	1	0.02635	646.4304	0.0276	
23	1	0.03097	614.5873	0.0978	
24	1	0.04223	691.1914	0.2215	
25	1	0.02399	601.7759	0.1648	
26	2	0.03180	20.6177	19.4210	

(Continued on next page)

Table 15d. Grade 6 2006 Operational Item Parameter Estimates (cont.)

Item	Max Pts	a par/ alpha	b par/ gamma1	c par/ gamma2	gamma3
27	2	0.04048	27.1510	25.0438	
28	2	0.02770	18.5279	18.1186	
29	3	0.02859	17.6139	18.3659	17.9640
30	2	0.01783	11.4646	10.2830	
31	2	0.03257	22.5782	21.0413	
32	2	0.03079	18.7104	19.6542	
33	3	0.03548	23.8947	24.3219	23.3558
34	3	0.02435	15.7715	16.3371	15.0818
35	3	0.04017	25.3356	28.4585	26.7978

Table 15e. Grade 7 2006 Operational Item Parameter Estimates

Item	Max Pts	a par/ alpha	b par/ gamma1	c par/ gamma2	gamma3
1	1	0.02392	602.4435	0.1694	
2	1	0.02675	575.2502	0.2000	
3	1	0.01705	631.1523	0.1694	
5	1	0.03012	615.9310	0.2051	
6	1	0.02269	614.4258	0.2745	
7	1	0.01696	554.9368	0.2000	
8	1	0.03910	640.5286	0.2714	
9	1	0.01348	637.9583	0.1694	
10	1	0.02143	657.3926	0.3306	
12	1	0.03264	601.5996	0.0674	
13	1	0.03332	648.1782	0.2212	
14	1	0.03436	665.8060	0.1764	
16	1	0.02358	655.2125	0.0448	
17	1	0.02297	598.2627	0.0307	
18	1	0.02213	655.8610	0.2679	
19	1	0.04218	680.4636	0.2285	
20	1	0.03038	626.4383	0.1885	
21	1	0.02908	568.3716	0.1694	
22	1	0.03426	646.4219	0.1859	
23	1	0.02239	625.1539	0.1340	
24	1	0.03696	673.3446	0.2054	
25	1	0.01345	674.8019	0.1063	
26	1	0.01196	581.5527	0.2000	
27	1	0.01575	575.4078	0.2000	
28	1	0.03225	632.4098	0.3296	
29	1	0.02162	670.7078	0.3480	

(Continued on next page)

Table 15e. Grade 7 2006 Operational Item Parameter Estimates (cont.)

Item	Max Pts	a par/ alpha	b par/ gamma1	c par/ gamma2	gamma3
30	1	0.02785	645.4647	0.1724	
31	2	0.02289	13.7618	14.9829	
32	2	0.04410	27.8323	30.2147	
33	3	0.02938	18.5234	19.7696	18.6820
34	3	0.02655	17.1967	18.7677	17.1558
35	2	0.04083	27.0105	26.9863	
36	2	0.03295	21.8401	20.6722	
37	3	0.02277	11.9386	16.8447	15.4809
38	3	0.03575	23.0368	23.5810	23.7559

Table 15f. Grade 8 2006 Operational Item Parameter Estimates

Item	Max Pts	a par/ alpha	b par/ gamma1	c par/ gamma2	gamma3
1	1	0.03141	664.5044	0.3626	
2	1	0.02482	624.4440	0.2694	
3	1	0.02636	649.2514	0.2338	
4	1	0.03517	664.6324	0.3071	
5	1	0.04005	660.4163	0.1538	
6	1	0.04994	663.5501	0.3808	
7	1	0.04134	691.0345	0.1752	
8	1	0.02385	648.1544	0.1563	
9	1	0.03524	647.9830	0.2395	
10	1	0.03610	652.9907	0.1541	
11	1	0.03005	634.1697	0.4994	
12	1	0.02577	615.9004	0.2194	
13	1	0.01964	634.0898	0.1849	
14	1	0.01497	646.8560	0.3126	
15	1	0.03420	650.9235	0.3899	
16	1	0.04707	674.9749	0.3748	
18	1	0.04057	638.2175	0.2150	
19	1	0.04466	649.0950	0.2025	
20	1	0.02327	628.3431	0.0985	
21	1	0.02741	640.3220	0.1880	
22	1	0.03332	640.0059	0.1670	
23	1	0.03638	637.0397	0.2387	
24	1	0.03795	666.5538	0.3261	
25	1	0.03099	632.4268	0.2713	
26	1	0.03072	667.1426	0.2312	

(Continued on next page)

Table 15f. Grade 8 2006 Operational Item Parameter Estimates (cont.)

Item	Max Pts	a par/ alpha	b par/ gamma1	c par/ gamma2	gamma3
27	1	0.03718	649.9465	0.1515	
28	2	0.03802	24.6473	25.1813	
29	3	0.03350	21.5487	20.8715	21.6524
30	2	0.02360	14.9559	14.6696	
31	3	0.04693	30.8976	31.3120	31.1526
32	2	0.05451	34.7139	36.5705	
33	2	0.03525	20.8242	24.1243	
34	2	0.03375	20.9876	21.6438	
35	2	0.04408	28.1364	28.6556	
36	3	0.04711	30.9588	31.2690	31.2708
37	2	0.04287	26.5032	27.5021	
38	2	0.05582	36.0824	36.3965	
39	3	0.03939	25.4196	25.7102	26.0079
40	2	0.04404	27.8911	29.8270	
41	2	0.06207	40.1747	40.4823	
42	3	0.03053	18.1478	19.7195	20.5830
43	3	0.02725	16.7220	17.4967	18.5146
44	2	0.03796	25.0593	24.2495	
45	2	0.02935	19.3669	18.0621	

Test Characteristic Curves

Test Characteristic Curves (TCCs) provide an overview of the test in IRT SS metric. The TCCs were generated using final operational item parameters for all test items. TCCs are the summation of all the Item Characteristic Curves (ICCs), for items which contribute to the Operational Scale Score. Standard Error (SE) Curves graphically show the amount of measurement error at different ability levels. TCCs and SE Curves are presented on the next page, in Figures 1 through 6. These curves provided target psychometric properties for selection of 2007 operational test forms. During selection of the 2007 test forms, consideration was given to proper alignment of the baseline (2006) TCC and SE curves and 2007 TCC and SE curves.

Figure 1. Grade 3 2006 OP TCC and SE

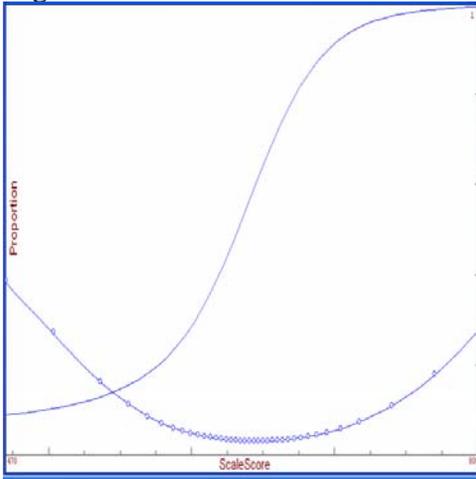


Figure 2. Grade 4 2006 OP TCC and SE

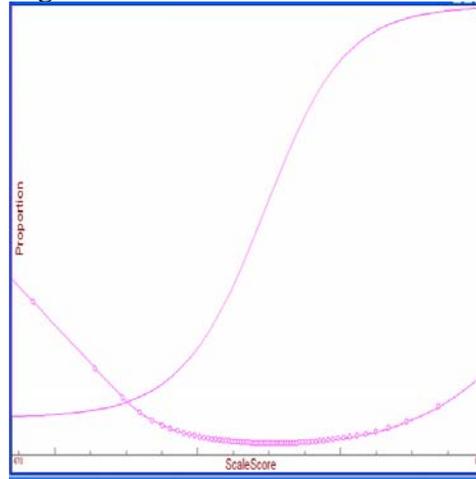


Figure 3. Grade 5 2006 OP TCC and SE

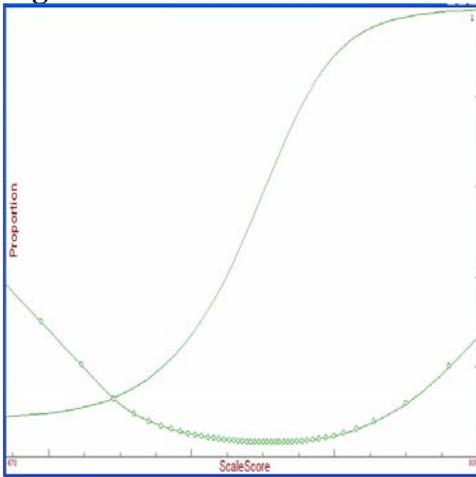


Figure 4. Grade 6 2006 OP TCC and SE

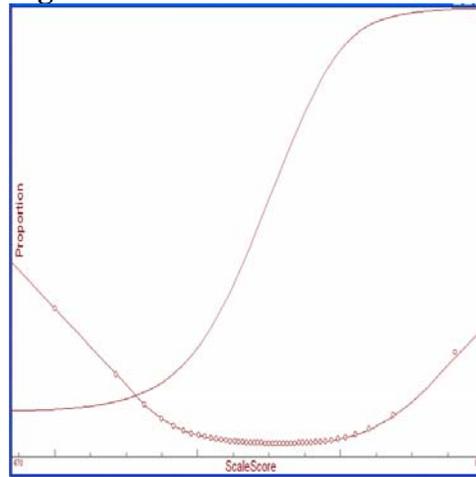


Figure 5. Grade 7 2006 OP TCC and SE

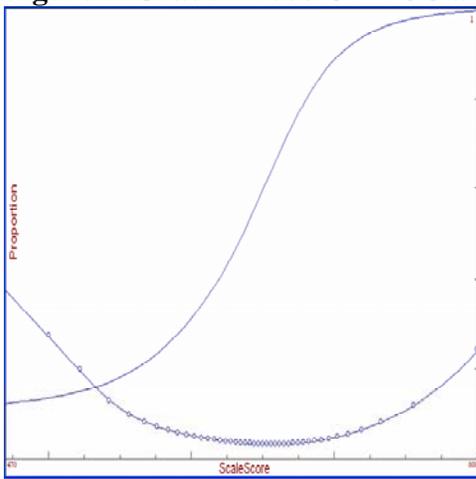
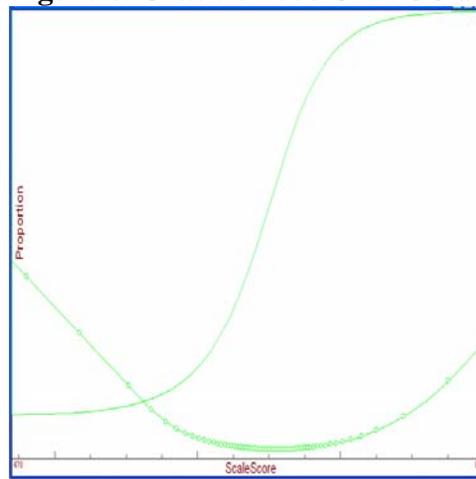


Figure 6. Grade 8 2006 OP TCC and SE



Equating

The Grades 3-8 Math testing program is considered to be a new family of tests on new scales with new proficiency and content standards set in 2006. Therefore, no direct equating between years 2005 and 2006 was performed for the grades 4 and 8 assessments. It should be noted that there is no history for the grades 3, 5, 6 and 7 Math state assessments. The new Math assessments (administered in 2007 and beyond) will be equated to the 2006 baseline year during live data calibrations using a TCC equating method (Stocking & Lord, 1983) and implemented in PARDUX.

The 2006 operational item parameters were used to scale and equate the 2003, 2005 and 2006 field test (FT) items eligible for selections of 2007 (and future) operational test forms. The 2006 MC item parameters were used as anchor parameters to equate the 2006 FT items to the 2006 scale via common examinees that were administered both the 2006 operational test and the 2006 field test. The 2005 field test items were equated to the 2006 scale via common item set. The 2006 operational MC items that were initially administered during the 2005 field test constituted the anchor set for this equating. Finally, small subsets of 2003 field test items for grades 4 and 8 were also placed on the 2006 scale. This equating was done in two steps. First, common items between the 2006 OP and 2005 FT and common examinees between the 2005 FT and 2005 OP were used to form a link between the 2006 and 2005 OP. This operation placed 2005 OP test items on 2006 scale. Next, common items between 2005 OP, 2006 OP and 2003 FT were used as anchor items to place the 2003 FT items on the 2005 scale (now same as 2006 scale). Only MC items were used as anchors. A Stocking and Lord TCC equating method implemented in PARDUX was employed to equate the 2003, 2005 and 2006 FT items to the 2006 operational scale. A detailed description and discussion of the FT equating procedures is provided in the separate NYSTP Grades 3-8 Math Field Test Report.

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 Math Tests. An inverse TCC method was employed. The scoring tables were created using CTB/McGraw-Hill's proprietary FLUX program. The inverse of the test characteristic curve procedure produces trait values based on unweighted raw scores. These estimates show negligible bias for tests with maximum possible raw scores of at least 30 points. All New York State Math tests have a maximum raw score higher than 30 points. In the inverse TCC method, a student's trait

estimate is taken to be the trait value which has an expected raw score equal to the student's observed raw score. It was found that for tests containing all MC items, the inverse of the TCC is an excellent first-order approximation to the number correct Maximum Likelihood Estimates (MLE) showing negligible bias for tests of at least 30 items. For tests with a mixture of MC and CR items, the MLE and TCC estimates are even more similar (Yen, 1984).

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

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

where

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

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

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

After the Raw Score to Scale Score conversion tables are produced, some adjustments to the lowest and highest obtainable scale scores are typically necessary to obtain a smooth transition between the lowest obtainable scale score (LOSS) and the penultimate LOSS and between the highest obtainable scale score (HOSS) and the penultimate HOSS. The preliminary LOSS and HOSS are automatically set by FLUX, but most of the time they need to be manually adjusted to obtain better psychometric properties of the scoring table and/or score distribution. There are no strict statistical procedures for LOSS and HOSS adjustment and CTB/McGraw-Hill developed a guideline for this procedure.

The following scoring table properties were taken into consideration while setting HOSS:

- The HOSS must be greater than the SS ($n-1$) that is the Scale Score associated with the number correct score for one item wrong (n is the maximum number of raw score points on a test)
- The HOSS should be low enough that the Standard Error (SE) for HOSS $< 10 \times$ Minimum (SE)
- The HOSS gap should be in the same ballpark as the Penultimate HOSS gap

It is usually more difficult to set LOSS values than HOSS values because LOSS values have much higher standard errors. The following scoring table properties were taken into consideration while setting LOSS:

- The LOSS should be high enough that the SE for Loss $< 15 \times$ Min (SE); this criterion can be difficult to meet for some tests
- The LOSS gap should be in the same ballpark as the Penultimate LOSS gap

Adjustments to LOSS and HOSS values were made to meet listed above specifications. The adjustments included manual changes to the LOSS and HOSS. After each change the scoring tables were regenerated and their properties evaluated. Various scale ranges were examined and the most appropriate scale score ranges to maintain psychometric

properties of the scales were identified. The LOSS and HOSS values are presented in Table 16, below.

Table 16. NYSTP Math 2006 Minimum and Maximum Scale Scores

Grade	LOSS	HOSS
3	470	770
4	485	800
5	495	780
6	500	780
7	500	800
8	480	775

Raw Score to Scale Score and SEM Conversion tables

The scale score (SS) is the basic score for the New York State tests. It is used to derive other scores that describe test performance, such as the four performance levels and 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. It should be noted that the Level III cut (650) set during the Standard Setting process does not always appear in Math scoring tables. This cut score established during the standard setting was based on a combination of information from Ordered Item Booklet (item content and item parameters) and scale score frequency distributions based on preliminary scoring tables. The adjustment to the cut scores during the Measurement Review meeting were based on scale score frequency distributions only. In cases where the adjustments were based on the scale score frequency distribution alone, the transformed cut score value (650) appears in a scoring table. In cases where the Level III cut was set based on the Ordered Item Booklet and the corresponding item parameter did not appear in the preliminary scoring table (not all item parameter values from the Ordered Item Booklet appear as ability estimates in scoring tables), the transformed cut is still 650, but does not appear in the final scoring table.

The Standard Error (SE) of a scale score indicates the precision with which the ability is estimated and it 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 SS 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 17a. Grade 3 Raw Score to Scale Score (with Standard Error)

Raw Score	Scale Score	Standard Error
0	470	113
1	470	113
2	470	113
3	470	113
4	503	80
5	536	47
6	556	33
7	569	25
8	579	21
9	587	18
10	594	15
11	599	14
12	604	13
13	609	12
14	613	11
15	617	11
16	621	10
17	624	10
18	628	10
19	631	9
20	634	9
21	637	9
22	640	9
23	644	9
24	647	9
25	650	9
26	653	9
27	657	9
28	660	9
29	664	10
30	668	10
31	672	11
32	676	11
33	682	12
34	688	13
35	695	15

(Continued on next page)

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

Raw Score	Scale Score	Standard Error
36	704	17
37	717	22
38	740	32
39	770	52

Table 17b. Grade 4 Raw Score to Scale Score (with Standard Error)

Weighted Raw Score	Scale Score	Standard Error
0	485	100
1	485	100
2	485	100
3	485	100
4	485	100
5	485	100
6	485	100
7	528	56
8	547	37
9	559	27
10	568	22
11	575	19
12	581	17
13	586	15
14	590	14
15	594	13
16	598	12
17	601	12
18	604	11
19	607	11
20	610	10
21	613	10
22	615	10
23	618	9
24	620	9
25	622	9
26	624	9
27	626	8
28	628	8
29	630	8
30	632	8
31	634	8

(Continued on next page)

Table 17b. Grade 4 Raw Score to Scale Score (with Standard Error) (cont.)

Weighted Raw Score	Scale Score	Standard Error
32	636	8
33	638	8
34	640	8
35	641	8
36	643	8
37	645	7
38	647	7
39	649	7
40	650	7
41	652	7
42	654	7
43	656	7
44	658	7
45	659	7
46	661	7
47	663	8
48	665	8
49	667	8
50	669	8
51	671	8
52	673	8
53	676	8
54	678	8
55	680	8
56	683	9
57	685	9
58	688	9
59	691	9
60	695	10
61	698	10
62	702	11
63	707	12
64	712	12
65	718	14
66	725	15
67	734	17
68	747	22
69	769	31
70	800	54

Table 17c. Grade 5 Raw Score to Scale Score (with Standard Error)

Raw Score	Scale Score	Standard Error
0	495	88
1	495	88
2	495	88
3	495	88
4	495	88
5	523	60
6	546	37
7	560	28
8	570	23
9	578	20
10	586	18
11	592	16
12	597	15
13	603	14
14	607	13
15	611	13
16	615	12
17	619	11
18	623	11
19	626	11
20	629	10
21	633	10
22	636	10
23	639	10
24	642	10
25	644	9
26	647	9
27	650	9
28	653	9
29	656	9
30	659	9
31	661	9
32	664	9
33	667	10
34	671	10
35	674	10
36	677	10
37	681	11
38	685	11
39	689	12
40	694	13

(Continued on next page)

Table 17c. Grade 5 Raw Score to Scale Score (with Standard Error) (cont.)

Raw Score	Scale Score	Standard Error
41	700	14
42	706	15
43	715	18
44	728	23
45	750	35
46	780	59

Table 17d. Grade 6 Raw Score to Scale Score (with Standard Error)

Raw Score	Scale Score	Standard Error
0	500	97
1	500	97
2	500	97
3	500	97
4	500	97
5	500	97
6	543	54
7	563	34
8	575	25
9	583	20
10	590	17
11	596	15
12	601	14
13	605	13
14	609	12
15	613	11
16	616	11
17	620	11
18	623	10
19	626	10
20	629	10
21	632	9
22	634	9
23	637	9
24	640	9
25	642	9
26	645	9
27	647	9
28	650	9
29	652	9
30	655	9

(Continued on next page)

Table 17d. Grade 6 Raw Score to Scale Score (with Standard Error) (cont.)

Raw Score	Scale Score	Standard Error
31	657	9
32	660	9
33	663	9
34	665	9
35	668	9
36	671	9
37	674	9
38	676	9
39	679	9
40	683	9
41	686	10
42	690	10
43	694	11
44	698	11
45	704	12
46	710	14
47	720	18
48	737	27
49	780	68

Table 17e. Grade 7 Raw Score to Scale Score (with Standard Error)

Raw Score	Scale Score	Standard Error
0	500	80
1	500	80
2	500	80
3	500	80
4	500	80
5	500	80
6	500	80
7	522	58
8	542	38
9	556	29
10	567	24
11	576	21
12	584	19
13	590	17
14	597	16
15	602	15
16	607	14

(Continued on next page)

Table 17e. Grade 7 Raw Score to Scale Score (with Standard Error) (cont.)

Raw Score	Scale Score	Standard Error
17	612	13
18	616	12
19	620	12
20	624	11
21	628	11
22	631	10
23	635	10
24	638	10
25	641	10
26	644	10
27	647	10
28	650	9
29	653	9
30	656	9
31	659	9
32	662	9
33	665	10
34	668	10
35	671	10
36	675	10
37	678	10
38	682	11
39	686	11
40	691	12
41	696	13
42	702	14
43	710	16
44	719	19
45	733	24
46	756	35
47	800	71

Table 17f. Grade 8 Raw Score to Scale Score (with Standard Error)

Weighted Raw Score	Scale Score	Standard Error
0	480	119
1	480	119
2	480	119
3	480	119
4	480	119
5	480	119

(Continued on next page)

Table 17f. Grade 8 Raw Score to Scale Score (with Standard Error) (cont.)

Weighted Raw Score	Scale Score	Standard Error
6	480	119
7	517	82
8	552	48
9	567	32
10	578	24
11	585	19
12	591	17
13	596	15
14	601	13
15	605	12
16	608	11
17	611	10
18	614	10
19	617	9
20	619	9
21	621	8
22	624	8
23	626	8
24	628	8
25	630	7
26	631	7
27	633	7
28	635	7
29	636	7
30	638	7
31	640	6
32	641	6
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	655	6
43	657	6
44	658	6
45	660	6
46	661	6
47	663	6

(Continued on next page)

Table 17f. Grade 8 Raw Score to Scale Score (with Standard Error) (cont.)

Weighted Raw Score	Scale Score	Standard Error
49	666	6
50	668	6
51	669	6
52	671	6
53	673	7
54	675	7
55	677	7
56	679	7
57	681	7
58	684	8
59	686	8
60	689	9
61	693	9
62	697	10
63	701	11
64	707	12
65	715	15
66	725	19
67	744	28
68	775	51

Standard Performance Index

The Standard Performance Index (SPI) reported for each objective measured by the Grades 3-8 Math 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 2006 Grades 3-8 Math 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 18 presents SPI target ranges. The objectives in these tables are denoted as follows: 1 – Number Sense and Operations, 2 – Algebra, 3 – Geometry, 4 – Measurement, and 5 – Statistics and Probability.

Table 18. SPI Target Ranges

Grade	Objective	No. Items	Total Points	Level III cut SPI target range
3	1	16	17	54-70
	2	4	6	62-75
	3	3	5	50-61
	4	5	5	69-82
	5	3	6	55-71
4	1	23	32	46-58
	2	7	10	55-65
	3	5	8	69-75
	4	9	12	49-62
	5	4	8	54-66
5	1	13	16	58-72
	2	5	5	53-69
	3	7	12	47-58
	4	5	7	51-63
	5	4	6	46-62
6	1	14	20	43-57
	2	7	9	54-65
	3	6	8	58-73
	4	4	6	59-77
	5	4	6	46-65
7	1	9	11	57-71
	3	7	9	57-68
	4	6	9	38-54
	5	10	15	57-67
8	1	5	9	43-57
	2	20	29	45-58
	3	13	20	56-68
	4	6	10	59-69

It should be noted that SPI scores typically require at least 4 score points within any Content Strand. Excluding items 4 and 15 from the grade 7 test and item 17 from the

grade 8 test still left at least 4 score points in those strands for the SPI computation. However, excluding item 11 from the grade 7 test resulted in having only 3 items (and 3 score points) in the Algebra Content Strand. Because SPI scores based on less than 4 points tend to be less reliable (Yen, Sykes, Ito & Julian, 1997), a decision was made not to compute the SPI scores for Grade 7 Algebra in 2006.

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 Math 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.

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 Math Tests (Linn & Harnisch, 1981). In this procedure, item parameters (discrimination, location, and guessing) and the scale score (θ) for each examinee were estimated for the three-parameter logistic model or the two parameter partial credit model in the case of constructed-response 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 (θ) scale. The expected proportion correct for each group based on the model prediction is compared to the observed (actual) proportion correct obtained by the group. The proportion of people in decile g who are expected to answer item i correctly is

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

where n_g is the number of examinees in decile g . To compute the proportion of people expected to answer item i correctly (over all deciles) for a group (e.g., Asian) the formula is given by:

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

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

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

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

$$D_i = O_i - P_i.$$

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 (θ) scale. The decile group difference (D_{ig}) can be either positive or negative. When the difference (D_{ig}) is

greater than or equal to 0.100, or less than or equal to -0.100, the item is flagged for potential DIF.

The following groups were analyzed using the IRT based DIF analysis: Female, Male, Asian, Black or African-American, Hispanic or Latino, White, High-Needs Districts (by NRC code), Low-Needs Districts (by NRC code), Chinese language test version, and Spanish language test version. Note that the N-counts of the following groups were insufficient for analysis in any grade: Native Hawaiian/Other Pacific Islander, Russian, Haitian-Creole, and Korean. The Linn-Harnisch DIF computation procedure does not require large samples, but a minimum sample of 200 cases per focal group is generally recommended. Note that the N-count for the Grade 3 Chinese language test version was 181 and the results need to be interpreted with caution. The N counts for all other groups were over 200. Most of the items flagged by IRT DIF were items from Chinese and Spanish language versions of the test. It should be noted that the 2005 Math field tests were not offered in any of the translation languages and no data were available for initial DIF analyses (before operational form selection). Also, as indicated in the classical DIF analysis section, items flagged for DIF do not necessarily display bias. As shown in Table 19, 7 items were flagged for DIF in grade 3, 19 items were flagged in the grade 4 test, 6 items were flagged in the grade 5 test, 9 items were flagged on the grade 6, 14 items were flagged on the grade 7 test, and 18 items were flagged on the grade 8 test the by Linn-Harnisch method. A detailed list of flagged items including DIF direction and magnitude is presented in Appendix E

Table 19. Number of Items Flagged for DIF by the Linn-Harnisch Method

Grade	Number of Flagged Items
3	7
4	19
5	6
6	9
7	14
8	18

Section VII: Standard Setting

This section provides some basic information on the process and results from the process of determining performance categories and establishing cut scores. Standard setting for the Grades 3-8 Math Tests occurred in Albany from July 18th through July 21st 2006. Prior to this meeting a Measurement Review meeting attended by representatives of the State and CTB/McGraw-Hill was held in Albany on December 1st, 2005. Participants were recruited from across the State of New York for the Measurement Review. The same participants met again after the Standard Setting for the Measurement Review Forum. The second meeting was held in Albany on July 24th 2006. This section briefly describes the model, participants, achievement levels and results from the standard setting and adjustments from the Measurement Review Forum. Standard setting technical reports, with greater detail on the elements presented here and additional information on validity, security, quality control, training of and evaluations from participants, and detailed results were published separately and provided to NYSED. Please refer to the *Bookmark Standard Setting Technical Report 2006 for Grades 3, 4, 5, 6, 7, and 8 Mathematics* and *NYS 2006 Measurement Review Technical Report 2006 for Mathematics* for more details.

Description of Standard Setting Process

The NYSTP Math Standard Setting was a multi-step process during which New York State educators and policy makers set the new performance standards for the Grades 3-8 Math Tests. The Standard Setting process involved the following stages:

1. Measurement Review Forum – The fifteen participants recruited by NYSED for the Measurement Review Forum were policy makers and educators in the New York State educational system. The purpose of this meeting was to review the 2005 performance standards for grades 4 and 8 and to recommend ideal impact data for the New York State Testing Program (NYSTP) new set assessments in Mathematics (Grades 3-8 Math Tests). For more details on Measurement Review Forum process and outcomes, refer to the *Measurement Review Technical Report 2005 for English/Language Arts and Mathematics*. The impact data recommended by Measurement Review participants are presented in Table 20.

Table 20. Measurement Review Meeting based Recommended Impact Data

Grade	% of Students in Each Performance Level				% Level III and IV
	Level I	Level II	Level III	Level IV	
3	8.0	14.9	50.3	26.8	77.1
4	6.6	13.1	52.1	28.2	80.3
5	8.0	15.3	50.7	26.0	76.7
6	9.1	17.0	49.9	24.0	73.9
7	10.0	19.1	49.1	21.8	70.9
8	11.3	20.5	48.1	20.1	68.2

2. Standard Setting Committee (all standard setting participants) – At this stage the New York State educators examined test items for content and recommended content-based (and impact data supported) cut scores. Participants in each grade participated in 3 or 4 rounds of activities in which they recommended three cut scores (*Partially Meeting Learning Standards, Meeting Learning Standards, and Meeting Learning Standards with Distinction*), which defined four performance levels: *Not Meeting Learning Standards, Partially Meeting Learning Standards, Meeting Learning Standards, and Meeting Learning Standards with Distinction*. Participants were recruited from across New York to recommend cut scores. Each grade had approximately 25 participants. The standard setting participants were involved in setting standards for two grades. The grade groups were Grades 3 and 4, Grades 5 and 6, and Grades 7 and 8. The participants went through 3 or 4 rounds of bookmark placements while setting performance cuts. The Bookmark placement was always an individual activity followed by group discussions and impact data presentation. Several types of impact data was shown to participants: data from previous Bookmark placement (voting) rounds, data from other grades, and historical data for grades 4 and 8. The impact data were presented as a ‘reality check.’ The final cut scores set by standard setting participants along with corresponding impact data are shown in Table 21.

Table 21. Participants based Cut Scores and Associated Impact Data

Grade	Level II Cut Score	Level III Cut Score	Level IV Cut Score	% of Students in Each Performance Level				% Level III and IV
				Level I	Level II	Level III	Level IV	
3	410	430	476	11.4	17.0	54.6	17.0	71.6
4	463	488	518	14.8	24.3	34.9	26.0	60.9
5	513	549	577	13.7	36.2	31.0	19.1	50.1
6	581	604	634	29.4	27.4	30.0	13.2	43.2
7	617	653	681	17.8	39.8	27.0	15.4	42.4
8	680	702	730	26.3	26.4	29.0	18.3	47.3

3. Vertical Articulation Panel (table leaders) – At this stage table leaders discussed final recommendations from standard setting groups and examined the impact data for logical progression from grade to grade. Based on the test content and impact data they adjusted cut scores to allow for logical progression (smooth trend) of impact data across grades (see Table 22).

Table 22. Vertical Articulation Panel based Cut Scores and Associated Impact Data (Table Leader Smoothing)

Grade	Level II Cut Score	Level III Cut Score	Level IV Cut Score	% of Students in Each Performance Level				% Level III and IV
				Level I	Level II	Level III	Level IV	
3	410	438	476	11.4	25.0	46.6	17.0	63.6
4	463	488	529	14.9	24.3	45.9	15.0	60.8
5	515	549	580	15.5	34.4	35.2	15.0	50.2
6	573	604	634	20.6	36.2	30.0	13.2	43.3
7	621	653	683	20.4	37.3	30.2	12.2	42.4
8	678	707	739	24.6	34.7	28.4	12.3	40.7

4. Measurement Review Forum – The same participants who attended the Measurement Review Forum in December 2005 were invited again to study the data presented during the workshop and to draw upon their experience working with students, schools, and school systems around the State of New York. Nine participants reviewed NYSTP Math grades 4 and 8 historical results, along with results from the Bookmark Procedure (stage 2 of Standard Setting) and the Vertical Articulation Panel (stage 3 of Standard Setting), and then recommended ways for further data smoothing. The recommended cuts and impact data from the Measurement Review Forum is presented in Table 23.

Table 23. Measurement Review Forum based Cut Scores and Associated Impact Data

Grade	Level II Cut Score	Level III Cut Score	Level IV Cut Score	% of Students in Each Performance Level				% Level III and IV
				Level I	Level II	Level III	Level IV	
3	397	422	474	6.3	13.1	55.4	25.2	80.6
4	447	472	519	7.4	14.6	52.1	26.0	78.1
5	508	535	579	10.3	21.2	49.3	19.2	68.5
6	562	592	634	13.3	26.2	47.3	13.2	60.6
7	609	644	683	13.1	31.1	43.6	12.2	55.8
8	667	697	744	14.8	31.1	43.8	10.3	54.1

5. NYSED final recommendation – at this stage NYSED reviewed historical results, results from Vertical Articulation Panel and the Measurement Forum and accepted the recommendation of the Measurement Forum without further adjustments.

Description of the Bookmark Method

The Bookmark method (Mitzel, Lewis, Patz, & Green, 2001) was used at Standard Setting to set cut scores. In the Bookmark method, an ordered item booklet (OIB) is produced which reorders the test items in order of difficulty. CR items appear multiple times in the OIB, with placement corresponding to the difficulty of obtaining each score point above zero. One item appears on each page. Participants conceptualized what point a minimally proficient student would successfully reach in the OIB and put a ‘Bookmark’ in that place. Participants were seated in groups at tables, and each table discussed their personal judgment (Bookmark results) and developed a table consensus. Then, results for each table were shared with the room along with impact data (percent of students in each performance level should the cut scores be applied). Through a balance of discussion and several rounds of adjustments, a final consensus of the location for each cut point was determined. Participants then filled out evaluations on the experience and reconvened in grade span groupings for descriptor writing. (Descriptors are written statements that describe the specific knowledge, skills and abilities a student must demonstrate to be classified into each performance achievement level.) The table leaders were convened to vertically articulate the impact data across grades; all participants understood that such smoothing would be conducted on the final round cuts (only table leaders were trained in articulation, but all participants were informed of this as part of the standard setting process). The articulated cut points and impact data were then forwarded to the Measurement Review Committee for State review and approval.

Description of Judge/Expert Panels

The panels were comprised of participants who were recruited from across New York State. A total of 75 New York State educators participated in the standard setting for the Grades 3-8 Math Tests. The majority of participants had Master’s degrees and over a decade of teaching experience. Each grade level had approximately 25 participants. The same groups of participants worked on establishing cut scores on adjacent grades 3 and 4, 5 and 6, and 7 and 8. The participants established cut scores for grades 4, 6 and 8 first and then for grades 3, 5 and 7. The participants for each grade were split into four tables (groups) that were balanced in regards to demographic statistics (i.e., school size and geographic location). Each table had a table leader, who monitored the group discourse. All participants were given extensive Bookmark training prior to the activity and had ample opportunity to familiarize themselves with the materials, data, process, and target student definitions. The Standard Setting Technical Report includes a survey of “Evaluation Results” that give information on the previous educational experience, diversity, and self-assessed confidence that the participants were well qualified and trained to validate the standard setting.

Vertically Moderated Standards

The New York State Math performance standards were set to satisfy the concept of vertical moderation. Vertical moderation of standards provides grade-to-grade comparability through consistency in setting cut scores for proficiency levels. In this approach, a smooth and rational pattern of percent of students falling into each

proficiency category was established during the standard setting process. There are two primary conditions that must be met to establish vertically moderated standards (VMS). First, a set of common policy definitions for the achievement levels needs to be used for all grades. Second, a consistent trend line needs to be imposed on the percentage of students in proficiency levels across grades (Huynh & Schneider, 2004). The Grades 3-8 Math Tests and test data satisfy both conditions. First, definitions for performance levels are comparable across grades for *Not Meeting Learning Standards*, *Partially Meeting Learning Standards*, *Meeting Learning Standards*, and *Meeting Learning Standards with Distinction* categories. Second, as shown in Table 23 below, there is a smooth decreasing trend of percent of students in Level III and Level IV categories. In the VMS approach, student growth could then be measured from year to year by measuring a student's progress relative to proficiency. In other words, a student's yearly progress is defined in terms of adequate end of year performance that allows the student to successfully meet the challenges in the next grade

Definition of Performance Levels

The standard setting participants wrote performance descriptors on the last day of the standard setting, using the items in the ordered item booklet as the evidence for their statements. The descriptors went through an editing process at CTB/McGraw-Hill (for style and consistency). When the final cut scores were established, the content grade specialists at CTB/McGraw-Hill adjusted the position of the descriptors if necessitated by an adjustment in cut score. The descriptors were written for the following performance levels:

Not Meeting Learning Standards (Level I) - Student performance does not demonstrate an understanding of the mathematics content expected at this grade level.

Partially Meeting Learning Standards (Level II) - Student performance demonstrates a partial understanding of the mathematics content expected at this grade level.

Meeting Learning Standards (Level III) - Student performance demonstrates an understanding of the mathematics content expected at this grade level.

Meeting Learning Standards with Distinction (Level IV) - Student performance demonstrates a thorough understanding of the mathematics content expected at this grade level.

Final Cut scores

As described in Section VI of this report, after Standard Setting each grade's data were rescaled such that the Level III cut equals 650. For details, please see the subsection Scaling. The final cut scores on the final scale are presented in Table 24.

Table 24. Final Cut Scores NYSTP Math

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

Section VIII: Reliability and Standard Error of Measurement

This section presents specific information on various test reliability statistics and standard errors of measurement, 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 standard error of measurement. For the Grades 3-8 MA Tests, we calculated two types of reliability statistics: Cronbach's Alpha (Cronbach, 1951) and Feldt-Raju (alpha) (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 (alpha) measures are appropriate for tests of multiple item formats (multiple-choice and constructed response). Please note that the Feldt-Raju statistics in Section IV are based upon the classical analysis and calibration sample, whereas the statistics below are based on the population.

Reliability for Total Test

Overall test reliability is a very good indication of each exam's internal consistency. Included in Table 25 are the case counts (N), number of test items (# Items), Cronbach's Alpha and associated Standard Error of Measurement (SEM), and Feldt-Raju Alpha and associated SEM obtained for the total Math tests.

Table 25. Reliability and Standard Error of Measurement for the 2006 NYSTP

Math Exams

Grade	N	# Items	# RS points	Cronbach's Alpha	SEM of Cronbach's	Feldt-Raju Alpha	SEM of Feldt-Raju
3	201908	31	39	0.89	2.35	0.90	2.27
4	202695	48	70	0.94	3.60	0.95	3.39
5	209200	34	46	0.90	3.07	0.91	2.89
6	211376	35	49	0.91	3.36	0.93	3.10

(Continued on next page)

Table 25. Reliability and Standard Error of Measurement for the 2006 NYSTP**Math Exams (cont.)**

Grade	N	# Items	# RS points	Cronbach's Alpha	SEM of Cronbach's	Feldt-Raju Alpha	SEM of Feldt-Raju
7	217225	35	47	0.89	3.17	0.91	2.91
8	219294	44	68	0.95	4.00	0.96	3.62

All of the coefficients for total test reliability are in the range of 0.89-0.96, which indicates 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 Alpha were computed separately for multiple choice and constructed-response items 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 26 presents reliabilities for the MC subsets.

Table 26. Reliability and Standard Error of Measurement for the 2006 NYSTP**Math Exams – MC Items Only**

Grade	N	# Items	Cronbach's Alpha	SEM of Cronbach's	Feldt-Raju Alpha	SEM of Feldt-Raju
3	201908	25	0.88	1.73	0.88	1.72
4	202695	30	0.88	2.10	0.88	2.08
5	209200	26	0.87	1.97	0.88	1.95
6	211376	25	0.86	1.97	0.87	1.96
7	217225	27	0.84	2.03	0.84	2.02
8	219294	26	0.87	2.12	0.87	2.11

Reliability for CR items

Reliability coefficients were also computed for the subsets of CR items. It should be noted that the Grades 3-8 MA Tests include 6 to 18 CR items depending on grade level. The results are presented in Table 27.

Table 27. Reliability and Standard Error of Measurement for the 2006 NYSTP

Math Exams – CR Items Only

Grade	N	# Items	# RS Points	Cronbach's Alpha	SEM of Cronbach's	Feldt-Raju Alpha	SEM of Feldt-Raju
3	201908	6	14	0.70	1.50	0.71	1.45
4	202695	18	40	0.91	2.70	0.92	2.66
5	209200	8	20	0.79	2.18	0.81	2.12
6	211376	10	24	0.86	2.47	0.86	2.40
7	217225	8	20	0.83	2.16	0.84	2.07
8	219294	18	42	0.93	3.05	0.94	2.92

Note: Results should be interpreted with caution for grades 3, 5 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, Disability status (all students with disabilities (SWD) together), and all students using test accommodations (SUA). For LEP students, reliability coefficients were computed for the following subgroups: students taking the English version of the Math test, and students taking Math tests in each of the five non-English languages the test was translated into (Chinese, Haitian-Creole, Korean, Russian and Spanish). As shown in Tables 28a-28f, the estimated reliabilities for subgroups were close in magnitude to the test reliability estimates of the population. Cronbach's Alpha reliability coefficients across subgroups were equal to or greater than 0.85, with the exception of the grade 3 Low Needs district subgroup for which the reliability coefficient was 0.83. 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 the same group (grade 3, Low Needs districts) for which the Feldt-Raju Alpha was 0.84. Overall, the New York State Math tests were found to have very good test internal consistency (reliability) for analyzed subgroups of examinees.

Table 28a. Grade 3 Test Reliability by Subgroup

Group	Subgroup	N	Cronbach's Alpha	SEM of Cronbach's	Feldt-Raju Alpha	SEM of Feldt-Raju
State	All Students	201908	0.89	2.35	0.90	2.27
Gender	Female	98465	0.89	2.35	0.90	2.27
	Male	103443	0.90	2.35	0.90	2.26
Ethnicity	Asian	14367	0.87	2.01	0.88	1.91
	Black or African-American	39984	0.90	2.56	0.90	2.48
	Hispanic or Latino	42108	0.89	2.51	0.90	2.42
	American Indian or Alaska Native	1016	0.89	2.52	0.89	2.44
	Native Hawaiian/Other Pacific Islander	40	0.89	2.28	0.90	2.16
	White	104380	0.86	2.21	0.87	2.15
	NRC	New York City	73451	0.91	2.42	0.91
Four Big Cites		8221	0.90	2.67	0.90	2.59
High Needs Urban-Suburban		16116	0.88	2.44	0.89	2.38
High Needs Rural		11560	0.86	2.39	0.87	2.33
Average Needs		59750	0.86	2.26	0.87	2.19
Low Needs		30045	0.83	2.07	0.84	2.01
Charter		2050	0.86	2.53	0.87	2.47
SWD	All codes	26796	0.90	2.69	0.91	2.61
SUA	All codes	38106	0.90	2.67	0.91	2.59
LEP	English	16193	0.89	2.63	0.90	2.54
	Chinese	183	0.85	2.38	0.87	2.21
	Haitian-Creole	52	0.90	2.80	0.91	2.67
	Korean	68	0.88	2.19	0.89	2.07
	Russian	50	0.93	2.74	0.94	2.54
	Spanish	3353	0.89	2.74	0.90	2.65
	All Translations	3706	0.90	2.72	0.90	2.62

Table 28b. Grade 4 Test Reliability by Subgroup

Group	Subgroup	N	Cronbach's Alpha	SEM of Cronbach's	Feldt-Raju Alpha	SEM of Feldt-Raju
State	All Students	202695	0.94	3.60	0.95	3.39
Gender	Female	98692	0.94	3.61	0.94	3.41
	Male	104003	0.94	3.58	0.95	3.37
Ethnicity	Asian	14482	0.93	3.10	0.94	2.94
	Black or African-American	38555	0.94	3.88	0.94	3.66
	Hispanic or Latino	41541	0.94	3.84	0.94	3.62
	American Indian or Alaska Native	971	0.94	3.79	0.94	3.58
	Native Hawaiian/Other Pacific Islander	50	0.95	3.39	0.96	3.07
	White	107092	0.93	3.39	0.94	3.24
	NRC	New York City	72565	0.94	3.73	0.95
NRC	Four Big Cites	8022	0.94	3.92	0.95	3.68
	High Needs Urban-Suburban	16090	0.94	3.75	0.94	3.54
	High Needs Rural	11741	0.93	3.69	0.93	3.52
	Average Needs	60888	0.93	3.46	0.93	3.31
	Low Needs	31122	0.91	3.12	0.92	3.00
	Charter	1590	0.93	3.85	0.94	3.66
	SWD	All codes	28712	0.94	4.03	0.95
SUA	All codes	39088	0.94	4.02	0.95	3.76
LEP	English	12512	0.94	4.01	0.95	3.76
	Chinese	246	0.92	3.43	0.92	3.25
	Haitian-Creole	47	0.90	4.13	0.91	3.88
	Korean	75	0.93	3.28	0.94	3.07
	Russian	50	0.96	3.90	0.97	3.53
	Spanish	2889	0.93	4.09	0.94	3.84
	All Translations	3307	0.94	4.05	0.95	3.79

Table 28c. Grade 5 Test Reliability by Subgroup

Group	Subgroup	N	Cronbach's Alpha	SEM of Cronbach's	Feldt-Raju Alpha	SEM of Feldt-Raju
State	All Students	209200	0.90	3.07	0.91	2.89
Gender	Female	102743	0.90	3.06	0.91	2.88
	Male	106457	0.91	3.08	0.92	2.88
Ethnicity	Asian	14543	0.89	2.71	0.91	2.53
	Black or African-American	40809	0.90	3.19	0.91	3.01
	Hispanic or Latino	42689	0.90	3.17	0.91	2.99
	American Indian or Alaska Native	1059	0.90	3.17	0.91	3.00
	Native Hawaiian/Other Pacific Islander	55	0.92	3.05	0.93	2.84
	White	110041	0.89	2.99	0.90	2.82
	NRC	New York City	74494	0.91	3.09	0.92
NRC	Four Big Cites	8421	0.90	3.27	0.91	3.07
	High Needs Urban-Suburban	16231	0.89	3.19	0.90	3.01
	High Needs Rural	12167	0.88	3.16	0.89	3.01
	Average Needs	63158	0.88	3.03	0.89	2.88
	Low Needs	31496	0.87	2.81	0.88	2.67
	Charter	2444	0.89	3.13	0.90	2.97
	SWD	All codes	30018	0.89	3.25	0.90
SUA	All codes	39987	0.89	3.25	0.90	3.07
LEP	English	10905	0.89	3.23	0.90	3.05
	Chinese	291	0.89	3.00	0.91	2.80
	Haitian-Creole	59	0.88	3.36	0.89	3.13
	Korean	79	0.85	2.58	0.87	2.39
	Russian	61	0.91	3.33	0.92	3.10
	Spanish	3023	0.88	3.26	0.89	3.07
	All Translations	3513	0.90	3.26	0.91	3.05

Table 28d. Grade 6 Test Reliability by Subgroup

Group	Subgroup	N	Cronbach's Alpha	SEM of Cronbach's	Feldt-Raju Alpha	SEM of Feldt-Raju
State	All Students	211376	0.91	3.36	0.93	3.10
Gender	Female	102985	0.91	3.35	0.92	3.11
	Male	108391	0.92	3.37	0.93	3.09
Ethnicity	Asian	14130	0.91	3.05	0.92	2.77
	Black or African-American	41827	0.90	3.42	0.91	3.20
	Hispanic or Latino	41960	0.91	3.42	0.92	3.18
	American Indian or Alaska Native	1138	0.91	3.41	0.92	3.18
	Native Hawaiian/Other Pacific Islander	44	0.90	3.32	0.92	3.02
	White	112276	0.90	3.29	0.91	3.05
	NRC	New York City	73988	0.92	3.39	0.93
Four Big Cites		8551	0.90	3.45	0.91	3.23
High Needs Urban-Suburban		16586	0.90	3.41	0.92	3.19
High Needs Rural		12922	0.89	3.38	0.91	3.18
Average Needs		65173	0.90	3.32	0.91	3.10
Low Needs		31608	0.89	3.12	0.90	2.89
Charter		1719	0.91	3.42	0.92	3.19
SWD	All codes	29778	0.89	3.31	0.91	3.12
SUA	All codes	36777	0.90	3.34	0.91	3.14
LEP	English	8746	0.90	3.35	0.92	3.14
	Chinese	288	0.90	3.26	0.91	2.96
	Haitian-Creole	62	0.90	3.20	0.91	3.01
	Korean	95	0.88	3.14	0.90	2.82
	Russian	48	0.88	3.31	0.90	3.11
	Spanish	2960	0.89	3.24	0.90	3.08
	All Translations	3453	0.91	3.29	0.92	3.09

Table 28e. Grade 7 Test Reliability by Subgroup

Group	Subgroup	N	Cronbach's Alpha	SEM of Cronbach's	Feldt-Raju Alpha	SEM of Feldt-Raju
State	All Students	217225	0.89	3.17	0.91	2.91
Gender	Female	105202	0.89	3.16	0.91	2.91
	Male	112022	0.90	3.17	0.91	2.90
Ethnicity	Asian	13911	0.89	2.94	0.91	2.68
	Black or African-American	43437	0.88	3.07	0.89	2.93
	Hispanic or Latino	42824	0.88	3.11	0.89	2.94
	American Indian or Alaska Native	1120	0.88	3.17	0.90	2.97
	Native Hawaiian/Other Pacific Islander	49	0.85	3.23	0.87	2.98
	White	115882	0.87	3.13	0.89	2.89
	NRC	New York City	75423	0.90	3.10	0.91
Four Big Cites		9446	0.86	3.06	0.88	2.93
High Needs Urban-Suburban		16824	0.88	3.17	0.89	2.99
High Needs Rural		13633	0.87	3.21	0.88	3.00
Average Needs		67836	0.87	3.16	0.89	2.93
Low Needs		31658	0.86	3.00	0.88	2.78
Charter		1336	0.88	3.06	0.89	2.89
SWD	All codes	29564	0.87	3.07	0.88	2.96
SUA	All codes	36685	0.88	3.09	0.89	2.97
LEP	English	9714	0.88	3.05	0.89	2.93
	Chinese	316	0.88	3.10	0.90	2.87
	Haitian-Creole	80	0.86	3.03	0.86	2.97
	Korean	105	0.89	2.93	0.91	2.68
	Russian	73	0.91	3.05	0.91	2.94
	Spanish	3266	0.85	3.05	0.86	2.96
	All Translations	3840	0.89	3.11	0.90	2.97

Table 28f. Grade 8 Test Reliability by Subgroup

Group	Subgroup	N	Cronbach's Alpha	SEM of Cronbach's	Feldt-Raju Alpha	SEM of Feldt-Raju
State	All Students	219294	0.95	4.00	0.96	3.62
Gender	Female	107180	0.95	3.98	0.96	3.62
	Male	112113	0.95	4.00	0.96	3.61
Ethnicity	Asian	13994	0.95	3.60	0.96	3.22
	Black or African-American	43407	0.93	3.97	0.94	3.72
	Hispanic or Latino	42116	0.93	4.01	0.94	3.74
	American Indian or Alaska Native	1073	0.94	4.04	0.95	3.72
	Native Hawaiian/Other Pacific Islander	33	0.95	3.99	0.96	3.56
	White	118668	0.94	3.87	0.95	3.54
	NRC	New York City	76162	0.95	3.98	0.96
NRC	Four Big Cites	9716	0.92	3.98	0.93	3.78
	High Needs Urban-Suburban	16576	0.94	4.04	0.94	3.74
	High Needs Rural	13591	0.93	4.00	0.94	3.72
	Average Needs	69213	0.93	3.89	0.94	3.59
	Low Needs	31713	0.93	3.56	0.94	3.29
	Charter	978	0.93	4.02	0.94	3.76
	SWD	All codes	29541	0.92	3.86	0.93
SUA	All codes	36773	0.93	3.91	0.94	3.67
LEP	English	9277	0.93	3.87	0.94	3.64
	Chinese	440	0.94	3.96	0.95	3.48
	Haitian-Creole	126	0.91	3.67	0.92	3.51
	Korean	126	0.93	3.84	0.95	3.38
	Russian	87	0.94	4.07	0.95	3.76
	Spanish	3452	0.91	3.83	0.92	3.65
	All Translations	4231	0.94	3.90	0.95	3.63

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 25. SEMs ranged from 2.35 to 4.00, which is reasonable and small given the maximum number of score points on Math tests. In other words, the error of measurement from the observed test score ranged from approximately +/-2 to 4 raw score points. SEMs are directly related to reliability: the

higher the reliability, the lower the standard error. As discussed, the reliability of these tests is relatively high, so it was expected that the SEMs would be very low.

The SEMs for subpopulations, as computed from Cronbach's Alpha and the Feldt-Raju reliability statistics, are presented in Tables 28a-28f. The SEMs associated with all reliability estimates for all subpopulations are in the range of 1.91-4.13, which is acceptably close to those for the entire population. This narrow range indicates that across the Grades 3-8 Math 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 Math 2006 Tests. In other words, this provides statistical information on the classification of students into the four performance categories (see Section VII for more detail on Standard Setting). 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 & 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 standard error of measurement 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 standard errors of measurement can be found in Section VI of this report 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 the tables below are case counts (N-count), classification consistency (Agreement), classification inconsistency (Inconsistency), and Cohen’s Kappa (Kappa). Consistency indicates the rate which 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 1-Agreement index. Cohen’s Kappa is a measure of agreement corrected for chance.

Table 29 (below) depicts the consistency study results based on the range of performance levels for all grades. Overall, between 73% 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, ranges from 0.59 to 0.72.

Table 29. Decision Consistency (All Cuts)

Grade	N	Agreement	Inconsistency	Kappa
3	201908	0.7411	0.2589	0.5887
4	202695	0.8008	0.1992	0.6913
5	209200	0.7410	0.2590	0.6170
6	211376	0.7614	0.2386	0.6511
7	217225	0.7339	0.2661	0.6171
8	219294	0.8079	0.1921	0.7195

Table 30 (below) depicts the consistency study results based on two performance levels (passing and not passing) as defined by the Level III cut. Overall, about 89% to 94% of the classifications of individual students are estimated to remain stable with a second administration. Kappa coefficients for classification consistency based on one cut range from 0.75 to 0.84.

Table 30. Decision Consistency (Level III Cut)

Grade	N	Agreement	Inconsistency	Kappa
3	201908	0.9209	0.0791	0.7546
4	202695	0.9370	0.0630	0.8196
5	209200	0.9023	0.0977	0.7751
6	211376	0.8986	0.1014	0.7883
7	217225	0.8864	0.1136	0.7701
8	219294	0.9224	0.0776	0.8436

Accuracy

The results of classification accuracy are presented in Table 31, below. Included in the table are case counts (N-count), classification accuracy (Accuracy) for all performance levels (All Cuts) and for the ‘passing’ cut score (Level III Cut) as well as ‘false positive’ and ‘false negative’ rates for both scenarios. It is always the case that the accuracy of the Level III cut score exceeds the accuracy referring to the entire set of cut scores, because there are only two categories for the true variable to be located in, instead of four. The accuracy rates indicate that the categorization of a student’s observed performance is in agreement with the location of their true ability approximately 80%-86% of the time across all performance levels, and approximately 93% to 96% of the time in regards to the Level III cut score.

Table 31. Decision Agreement (Accuracy)

Grade	N-count	Accuracy					
		All Cuts	False Positive (All Cuts)	False Negative (All Cuts)	Level III Cut	False Positive (Level III Cut)	False Negative (Level III Cut)
3	201908	0.8039	0.1357	0.0602	0.9439	0.0284	0.0277
4	202695	0.8552	0.0887	0.0560	0.9558	0.0201	0.0241
5	209200	0.8102	0.1201	0.0696	0.9273	0.0484	0.0242
6	211376	0.8275	0.1089	0.0636	0.9266	0.0457	0.0278
7	217225	0.8100	0.1053	0.0849	0.9196	0.0381	0.0423
8	219294	0.8627	0.0701	0.0671	0.9453	0.0229	0.0318

Section IX: Summary of Operational Test Results

This section summarizes the distribution of operational scale score results on the New York State 2006 Grades 3-8 Math Tests. These include the scale score means, standard deviations, percentiles and performance level distributions for each grade's population and specific subgroups. Gender, ethnic identification, Need/Resource Category (NRC), Limited English Proficiency (LEP), Disability, Accommodation and test language variables were used to calculate the results of subgroups required for federal reporting and test equity purposes. Because 2006 is the benchmark year, longitudinal comparisons are not yet available. Data include examinees with valid scores from all public and charter schools. Note that complete scale score frequency distribution tables can be found in Appendix I of this report.

Scale Score Summary

Scale score summary tables are presented and discussed, below. First, scale score statistics for total populations of students from public and charter schools are presented in Table 32. Next, scale score statistics are presented for selected sub-groups in each grade level. The statistics for groups with small N-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 Need and Average Need 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 Math test in English, Disabilities and/or Accommodations 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 can be found in Appendix H of this report.

Table 32. Math Grades 3-8 Scale Score Distribution Summary

Grade	N	SS Mean	SS Std Dev	10 th %tile	25 th %tile	50 th %tile	75 th %tile	90 th %tile
3	201908	677.49	37.75	634	653	676	704	717
4	202695	676.55	40.81	628	654	678	702	725
5	209200	665.59	39.85	615	642	667	689	715
6	211376	655.94	40.44	605	634	657	679	704
7	217225	651.08	40.55	602	628	653	678	696
8	219294	651.55	41.15	605	630	653	677	701

Grade 3

Scale score statistics and N-counts of demographic groups for grade 3 are presented in Table 33. The population scale score mean was 677.49. By gender subgroup, Females and Males performed very similarly, with a mean difference of less than one scale score point. Asian, Native Hawaiian/Other Pacific Islander, and White ethnic subgroups had scale score means that exceeded the State mean scale score on the exam, as did students from Low Need and Average Need NRC subgroups. The lowest performing NRC subgroup was the Big 4 Cities, with a mean of 650.74 and the lowest performing ethnic subgroups were Black or African-American (mean scale score of 662.00) and American Indian/Alaskan Native (664.28). Students with disabilities, testing accommodations, and LEP without testing in an alternate language scored consistently below the statewide percentile scale score rankings, and nearly one standard deviation below the mean scale score for the population. At the 50th percentile the scale scores on translated forms range from 639 (Haitian-Creole subgroup) to 688 (Korean subgroup), a difference that exceeds a standard deviation. The group of LEP students that took the Math test in English outperformed the total group of students that took translated forms in terms of test mean and reported percentile scores, except that both groups had the same 90th percentile scale score (695). The group of students that used the Haitian-Creole translation, which had a scale score mean 41 scale score units below the population mean (about one standard deviation), was the lowest performing group analyzed. At the 50th percentile, the following groups exceeded the population scale score of 676: Asian (695), White (682), Average Need (682), Low Need (688), and students who used the Chinese (682) and Korean (688) translations.

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

Demographic Category (Subgroup)		N	SS Mean	SS Std Dev	10 th %tile	25 th %tile	50 th %tile	75 th %tile	90 th %tile
State	All Students	201908	677.49	37.75	634	653	676	704	717
Gender	Female	98465	677.88	37.18	634	657	676	704	717
	Male	103443	677.11	38.28	631	653	676	695	717
Ethnicity	American Indian or Alaska Native	1016	664.28	34.55	624	644	664	682	704
	Asian	14367	699.73	37.82	653	676	695	717	740
	Black or African-American	39984	662.00	36.93	617	640	664	682	704
	Hispanic or Latino	42108	666.99	37.45	621	644	668	688	717
	Native Hawaiian/ Other Pacific Islander	40	678.90	38.97	633	657	668	700	740
	White	104380	684.72	34.65	644	664	682	704	740
	Need/ Resource Category	New York City	73451	673.87	41.00	624	650	672	695
Big 4 Cities		8221	650.74	36.93	609	628	650	672	695
High Need Urban/Suburban		16116	669.48	34.43	628	650	668	688	717
High Need Rural		11560	671.70	31.19	637	653	672	688	704
Average Need		59750	681.87	33.37	644	660	682	704	717
Low Need		30045	693.37	33.58	657	672	688	717	740
Charter		2050	664.68	31.31	628	647	664	682	704
Student With Disability	All Codes	26796	646.79	38.35	599	624	650	672	695
Accommodation	All Codes	38106	650.50	37.71	604	628	653	672	695
LEP	LEP = Y and Test language = English	16193	655.78	36.90	613	634	657	676	695
Test Language	Chinese	183	683.92	37.31	644	664	682	704	740
	Haitian-Creole	52	636.29	43.95	594	619	639	660	682
	Korean	68	688.53	39.35	644	664	688	704	740
	Russian	50	649.20	51.52	583	628	660	676	711
	Spanish	3353	645.16	36.68	599	624	647	668	688
	All Translations	3706	647.80	38.45	604	628	650	672	695

Grade 4

Scale score statistics and N-counts of demographic groups for grade 4 are presented in Table 34, below. The population scale score mean was 676.55. By gender subgroup, Females and Males performed very similarly, with a mean difference of less than two scale score points. Asian, Native Hawaiian/Other Pacific Islander, and White students' scale score means exceeded the State mean scale score on the exam. Asian students (the highest performing ethnic subgroup) exceeded the State mean by more than a half of a standard deviation. Black or African-American and Hispanic or Latino ethnic subgroups had mean scale scores almost one standard deviation below the Asian subgroup. Students from Low Need and Average Need districts outperformed the other NRC subgroups. The lowest performing NRC subgroup was the Big 4 Cities, with a mean of 653.13, slightly more than one-half a standard deviation below the State mean. Students with disabilities, testing accommodations, and LEP without testing in an alternate language scored consistently below the statewide percentile scale score rankings, and between 29 and 37 scale score points below the population mean. Haitian-Creole and Spanish translated forms had means over one standard deviation below the population. The group of LEP students that took the Math test in English outperformed the total group of students that took translated forms in terms of test mean and reported percentile scores. The group of students that used the Haitian-Creole translation, which had a scale score mean more than one standard deviation units below the population mean, was the lowest performing group analyzed. At the 50th percentile, the following groups exceeded the population scale score of 678: Asian (698), Native Hawaiian/Other Pacific Islander (695), White (682), Average Need (683), Low Need (695), and students who used the Chinese (685) and Korean (695) translations.

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

Demographic Category (Subgroup)		N	SS Mean	SS Std Dev	10 th %tile	25 th %tile	50 th %tile	75 th %tile	90 th %tile
State	All Students	202695	676.55	40.81	628	654	678	702	725
Gender	Female	98692	675.70	39.36	628	652	676	698	725
	Male	104003	677.35	42.12	628	654	678	702	725
Ethnicity	American Indian or Alaska Native	971	664.55	39.95	618	643	667	688	712
	Asian	14482	699.94	39.98	654	676	698	725	747
	Black or African-American	38555	658.85	39.61	613	636	661	683	707
	Hispanic or Latino	41541	662.65	39.74	615	640	665	688	707
	Native Hawaiian/ Other Pacific Islander	50	687.12	47.65	625	676	695	712	734
	White	107092	685.25	37.47	641	663	685	707	734
Need/ Resource Category	New York City	72565	669.59	42.78	620	645	671	695	718
	Big 4 Cities	8022	653.13	41.73	604	630	656	680	702
	High Need Urban/Suburban	16090	668.33	39.52	622	645	669	691	712
	High Need Rural	11741	671.02	35.86	628	652	671	691	712
	Average Need	60888	682.14	36.38	640	661	683	702	725
	Low Need	31122	696.25	35.57	656	676	695	718	734
	Charter	1590	663.13	35.34	620	641	663	685	707
Student With Disability	All Codes	28712	639.85	44.78	586	615	645	669	688
Accommodation	All Codes	39088	644.15	43.38	590	622	649	671	691
LEP	LEP = Y and Test language = English	12512	647.79	41.6	598	626	650	673	695
Test Language	Chinese	246	685.68	33.38	643	665	685	707	725
	Haitian-Creole	47	630.83	37.16	586	620	634	650	669
	Korean	75	693.16	44.22	649	678	695	718	747
	Russian	50	648.24	49.98	590	618	656	683	707
	Spanish	2889	636.28	43.05	586	615	640	663	683
	All Translations	3307	641.35	45.09	590	618	645	671	691

Grade 5

Grade 5 demographic groups N-counts and scale score statistics are presented in Table 35, below. The population scale score mean was 665.59 with a standard deviation of 39.85. By gender subgroup, Females and Males performed very similarly, with a mean difference of less than one scale score point. Asian, Native Hawaiian/Other Pacific Islander, and White students' scale score means exceeded the State mean scale score on the exam. Asian students (the highest performing ethnic subgroup) exceeded the State mean by over 25 scale score points. Black or African-American and Hispanic or Latino ethnic subgroups had mean scale scores approximately one standard deviation below the Asian subgroup. Students from Low Need and Average Need districts outperformed the other NRC subgroups. The lowest performing NRC subgroup was the Big 4 Cities, with a mean of 635.82, one-half a standard deviation below the second lowest performing NRC subgroup (High Need Urban/Suburban, 655.79) and 50 scale score units below the Low Need subgroup mean. Students with disabilities, testing accommodations, and LEP without testing in an alternate language 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 mean. The Haitian-Creole translation subgroup, which had a scale score mean (619.36) more than 45 units below the population mean, was the lowest performing group analyzed. The Low Need subgroup was the highest performing group analyzed, with a scale score mean of 685.53, about one half of a standard deviation above the population mean. At the 50th percentile, the following groups exceeded the population scale score of 667: Asian (689), Native Hawaiian/Other Pacific Islander (677), White (674), Average Need (671), Low Need (685), and students who used the Chinese (677) and Korean (700) translations.

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

Demographic Category (Subgroup)		N	SS Mean	SS Std Dev	10 th %tile	25 th %tile	50 th %tile	75 th %tile	90 th %tile
State	All Students	209200	665.59	39.85	615	642	667	689	715
Gender	Female	102743	665.29	38.42	619	642	664	689	715
	Male	106457	665.89	41.19	615	642	667	689	715
Ethnicity	American Indian or Alaska Native	1059	654.39	38.34	607	629	656	681	700
	Asian	14543	691.42	40.53	644	667	689	715	750
	Black or African-American	40809	647.13	37.14	603	623	647	671	694
	Hispanic or Latino	42689	653.11	37.38	607	629	653	677	700
	Native Hawaiian/ Other Pacific Islander	55	670.73	42.52	619	644	677	700	728
	White	110041	673.98	37.25	629	653	674	694	715
Need/ Resource Category	New York City	74494	660.31	41.34	611	633	659	685	706
	Big 4 Cities	8421	635.82	40.05	586	611	636	661	685
	High Need Urban/Suburban	16231	655.79	37.13	607	633	656	677	700
	High Need Rural	12167	659.51	34.67	619	639	661	681	700
	Average Need	63158	670.63	35.71	629	650	671	689	715
	Low Need	31496	685.53	35.97	644	664	685	706	728
	Charter	2444	656.92	34.77	611	636	659	681	700
Student With Disability	All Codes	30018	629.89	38.18	586	607	633	656	677
Accommodation	All Codes	39987	634.42	38.66	586	611	636	659	681
LEP	LEP = Y and Test language = English	10905	637.73	37.51	592	615	639	661	681
Test Language	Chinese	291	676.12	40.38	629	656	677	700	715
	Haitian-Creole	59	619.36	40.48	578	592	626	644	667
	Korean	79	700.90	36.15	650	674	700	728	750
	Russian	61	637.97	43.57	578	607	653	671	681
	Spanish	3023	625.25	38.72	578	603	626	650	671
	All Translations	3513	631.29	42.7	578	607	633	659	681

Grade 6

Grade 6 scale score statistics and N-counts of demographic groups are presented in Table 36, below. The population scale score mean was 655.94 with a standard deviation of 40.44. By gender subgroup, Females and Males performed very similarly, with a mean difference of less than one scale score point. Asian, Native Hawaiian/Other Pacific Islander, and White students' scale score means exceeded the State mean scale score on the exam. American Indian or Alaska Native, Black or African-American and Hispanic or Latino ethnic subgroups had mean scale scores (642.58, 637.90 and 641.72, respectively) approximately one standard deviation below the Asian subgroup. Students from Low Need and Average Need districts outperformed the other NRC subgroups. The lowest performing NRC subgroup was the Big 4 Cities, with a mean of 629.48. New York City, High Need Urban/Suburban, High Need Rural, and Charter subgroups had similar scale score means (ranging from approximately 644-651) that were all slightly below the population mean. Students with disabilities, testing accommodations, and LEP without testing in an alternate language 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 (605.92) 50 units below the population mean, was the lowest performing group analyzed. Asian students (the highest performing subgroup with a mean of 682.64) exceeded the State mean by 26.5 scale score points. At the 50th percentile, the following groups exceeded the population scale score of 657: Asian (683), Native Hawaiian/Other Pacific Islander (664), White (665), Average Need (663), Low Need (676), and students who used the Chinese (674) and Korean (679) translations.

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

Demographic Category (Subgroup)		N	SS Mean	SS Std Dev	10 th %tile	25 th %tile	50 th %tile	75 th %tile	90 th %tile
State	All Students	211376	655.94	40.44	605	634	657	679	704
Gender	Female	102985	656.20	39.06	609	634	657	679	698
	Male	108391	655.69	41.71	605	632	657	683	704
Ethnicity	American Indian or Alaska Native	1138	642.58	39.78	596	623	647	668	686
	Asian	14130	682.64	40.56	634	660	683	704	737
	Black or African-American	41827	637.90	38.64	590	616	640	663	683
	Hispanic or Latino	41960	641.72	38.89	596	620	645	668	686
	Native Hawaiian/ Other Pacific Islander	44	664.70	36.50	626	647	664	690	704
	White	112276	664.74	36.99	623	645	665	686	704
	Need/ Resource Category	New York City	73988	649.84	42.72	601	626	650	676
Big 4 Cities		8551	629.48	40.40	583	605	632	655	676
High Need Urban/Suburban		16586	644.30	37.98	601	623	647	668	690
High Need Rural		12922	650.97	35.22	609	632	652	674	690
Average Need		65173	661.26	35.71	620	642	663	683	704
Low Need		31608	676.76	35.05	637	655	676	698	720
Charter		1719	645.70	37.49	601	623	647	671	690
Student With Disability	All Codes	29778	616.84	42.01	563	596	620	645	665
Accommodation	All Codes	36777	620.62	42.25	575	596	623	647	668
LEP	LEP = Y and Test language = English	8746	624.07	42.69	575	601	626	652	674
Test Language	Chinese	288	673.00	36.84	626	651	674	694	710
	Haitian-Creole	62	605.92	49.78	500	583	616	640	660
	Korean	95	680.26	33.36	634	660	679	698	720
	Russian	48	640.06	35.84	590	622	641	657	679
	Spanish	2960	611.50	40.95	563	590	613	637	660
	All Translations	3453	618.82	45.23	563	596	620	647	674

Grade 7

N-counts and score statistics of demographic groups for grade 7 are presented in Table 37, below. The population scale score mean was 651.08 with a standard deviation of 40.55. The gender subgroups, Female and Male, performed very similarly, with a mean difference of two scale score points. Asian, Native Hawaiian/Other Pacific Islander, and White subgroups' scale score means (674.47, 654.67, and 662.68, respectively) exceeded the State mean scale score on the exam, by not more than one-half a standard deviation. American Indian or Alaska Native, Black or African-American and Hispanic or Latino ethnic subgroups had mean scale scores (639.09, 629.34 and 634.45, respectively) approximately one quarter of a standard deviation below the population. NRC subgroup achievement, high to low, was as follows: Low Need, Average Need, High Need Rural, New York City, High Need Suburban, Charter, and Big 4 Cities. The lowest performing NRC subgroup, Big 4 Cities, had a scale score mean of 618.81, while the Low Need subgroup's scale score mean is 675.92. Students with disabilities, testing accommodations, and LEP without testing in an alternate language scored consistently below the statewide percentile scale score rankings and had means nearly a standard deviation below the population mean. Haitian-Creole, Russian, and Spanish translation subgroups had scale score means (604.60, 609.14, and 608.49 respectively) more than one standard deviation below the population. The Haitian-Creole translation was the lowest performing group analyzed, yet the Korean translation subgroup was the highest. At the 50th percentile, the following groups exceeded the population scale score of 653: Asian (678), White (665), Average Need (662), Low Need (675), and students who used the Chinese (665) and Korean (678) translations.

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

Demographic Category (Subgroup)		N	SS Mean	SS Std Dev	10 th %tile	25 th %tile	50 th %tile	75 th %tile	90 th %tile
State	All Students	217225	651.08	40.55	602	628	653	678	696
Gender	Female	105202	652.17	39.31	607	631	653	678	696
	Male	112022	650.05	41.65	597	628	653	678	696
Ethnicity	American Indian or Alaska Native	1120	639.09	38.73	590	616	644	662	682
	Asian	13911	674.47	39.70	628	653	678	696	719
	Black or African-American	43437	629.34	37.86	584	607	631	653	675
	Hispanic or Latino	42824	634.45	37.87	590	616	638	659	678
	Native Hawaiian/ Other Pacific Islander	49	654.67	31.61	607	635	653	671	691
	White	115882	662.68	36.47	620	644	665	686	702
	Need/ Resource Category	New York City	75423	640.99	40.53	590	616	644	668
Big 4 Cities		9446	618.81	38.01	567	597	620	644	662
High Need Urban/Suburban		16824	637.91	37.51	590	616	641	662	682
High Need Rural		13633	647.60	34.80	607	628	650	671	686
Average Need		67836	660.39	35.59	620	641	662	682	702
Low Need		31658	675.92	35.16	638	656	675	696	719
Charter		1336	636.43	35.89	590	616	638	662	678
Student With Disability	All Codes	29564	612.36	42.05	556	590	616	641	659
Accommodation	All Codes	36685	616.14	42.08	556	590	620	644	665
LEP	LEP = Y and Test language = English	9714	617.10	42.46	567	597	620	644	665
Test Language	Chinese	316	660.31	35.64	612	646	665	682	696
	Haitian-Creole	80	604.60	41.06	549	576	607	635	650
	Korean	105	677.27	37.94	628	659	678	696	719
	Russian	73	609.14	51.25	522	576	616	644	668
	Spanish	3266	608.49	40.95	556	584	612	638	656
	All Translations	3840	614.56	44.35	556	590	616	644	668

Grade 8

Grade 8 Scale score statistics and N-counts of demographic groups are presented in Table 38, below. The population scale score mean was 651.55 with a standard deviation of 41.15. By gender subgroup, Females and Males performed similarly, with a mean difference of three scale score points. Asian, Native Hawaiian/Other Pacific Islander, and White ethnic subgroups' scale score means (677.96, 651.73, and 663.21 respectively) exceeded the State mean scale score on the exam. The Black or African-American subgroup had the lowest performance by ethnic subgroup, with a scale score mean of 629.03, but Hispanic or Latino and American Indian or Alaska Native subgroups' scale score means (633.45 and 639.95) were also below the population. NRC subgroup achievement, high to low, was as follows: Low Need, Average Need, High Need Rural, High Need Suburban, New York City, Charter, and Big 4 Cities. The lowest performing NRC subgroup, Big 4 Cities, had a scale score mean of 623.48, while the Low Need subgroup's scale score mean was 676.93, which indicates a large performance discrepancy by school district NRC designation. Students with disabilities, testing accommodations, and LEP without testing in an alternate language scored consistently below the statewide percentile scale score rankings, and nearly one standard deviation below the mean scale score for the population. At the 50th percentile the scale scores on translated forms range from 608 (Haitian-Creole subgroup) to 672 (Korean subgroup). The total group of students that took translated forms met or exceeded the performance of the subgroup of LEP students that took the Math test in English in terms of test mean and reported percentile scores. The group of students that used the Haitian-Creole translation, which had a scale score mean about 46.5 scale score units below the population mean, was the lowest performing group analyzed. At the 50th percentile, the following groups exceeded the population scale score of 653: Female (654), Asian (679), White (663), Average Need (661), Low Need (675), and students who used the Chinese (671) and Korean (672) translations.

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

Demographic Category (Subgroup)		N	SS Mean	SS Std Dev	10 th %tile	25 th %tile	50 th %tile	75 th %tile	90 th %tile
State	All Students	219294	651.55	41.15	605	630	653	677	701
Gender	Female	107180	653.05	40.09	608	631	654	677	701
	Male	112113	650.12	42.10	601	628	651	675	697
Ethnicity	American Indian or Alaska Native	1073	639.95	39.00	596	619	643	663	684
	Asian	13994	677.96	42.59	628	651	679	707	725
	Black or African-American	43407	629.03	39.01	585	611	631	653	671
	Hispanic or Latino	42116	633.45	38.63	591	614	635	657	677
	Native Hawaiian/ Other Pacific Islander	33	651.73	46.57	617	628	653	681	701
	White	118668	663.21	35.97	624	643	663	684	707
	Need/ Resource Category	New York City	76162	639.52	43.11	591	617	640	664
Big 4 Cities		9716	623.48	38.58	585	605	628	646	664
High Need Urban/Suburban		16576	640.59	36.93	601	621	641	661	681
High Need Rural		13591	650.22	33.42	614	633	651	669	689
Average Need		69213	661.41	34.33	624	643	661	681	701
Low Need		31713	676.93	34.89	638	657	675	697	715
Charter		978	634.55	36.92	591	614	636	657	675
Student With Disability	All Codes	29541	613.38	43.66	567	596	621	641	658
Accommodation	All Codes	36773	618.23	43.45	567	601	624	646	664
LEP	LEP = Y and Test language = English	9277	618.93	44.08	567	601	624	646	666
Test Language	Chinese	440	670.72	40.50	626	647	671	697	725
	Haitian-Creole	126	604.03	46.14	552	585	608	631	647
	Korean	126	673.06	34.20	633	655	672	693	715
	Russian	87	642.24	36.15	596	624	646	666	684
	Spanish	3452	613.35	40.77	567	596	619	640	657
	All Translations	4231	621.4	45.3	567	601	626	647	671

Performance Level Summary

Tables 39-45 show the performance level summary for all examinees from public and charter school with valid scores. Table 39 presents performance level data for total populations of students in grades 3-8. Tables 40 to 45 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. 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 Need Urban/Suburban, and High Need 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 students with disabilities and students using testing accommodations subgroups were low, compared to the total population of examinees. Across grades, the following subgroups consistently performed above the population average: Asian, White, Average Need, Low Need, Chinese translation, Korean translation. Please note that the case counts for the Native Hawaiian/Other Pacific Islander, Haitian/Creole translation, and Russian translation subgroup are very low and are heavily influenced by very high and/or very low achieving individual students.

Table 39. Grades 3-8 Math Test Performance Level Distributions

Grade	N-count	Percent of Population in Performance Level				
		Level I	Level II	Level III	Level IV	Meets Standards
3	201908	6.35%	13.13%	55.42%	25.11%	80.52%
4	202695	7.41%	14.59%	52.12%	25.88%	78.00%
5	209200	10.29%	21.24%	49.31%	19.16%	68.47%
6	211376	13.32%	26.23%	47.26%	13.19%	60.45%
7	217225	13.19%	31.12%	43.52%	12.17%	55.69%
8	219294	14.98%	31.09%	43.74%	10.18%	53.93%

Grade 3

Performance level summaries and N-counts of demographic groups for grade 3 are presented in Table 40, below. Statewide, 81% of 3rd graders are Meeting Learning Standards or Meeting Learning Standards with Distinction (Levels III and IV). Over 10% of Black or African-American and Hispanic or Latino students are Not Meeting Standards (Level I), as compared to only 6% of the population. American Indian or Alaska Native, Black or African-American, Hispanic or Latino and Native Hawaiian/Other Pacific Islander ethnic subgroups had a lower percent of students meeting standards (Levels III and IV) than the population, but the percent of White and Asian subgroups meeting the standard (88% and 92%) exceeded the population. Student achievement varied widely by Need/Resource Category subgroup, as well. Over 93% of

students from Low Need districts are meeting or exceeding the Standards; whereas, about 47% Big 4 Cities students are in Level I or II (Partially Meeting Learning Standards). About half of students with disability status or testing accommodations or those who took translated test forms are meeting the Standards; however, the subgroups for Korean and Chinese translations had about 88% meeting standards. The following subgroups had meeting standards rates above the State average: Female, Asian, White, Average Need, Low Need, Chinese translation and Korean translation.

Table 40. Performance Level Summary, By Subgroup, Grade 3

Demographic Category (Subgroup)		N	Level I	Level II	Level III	Level IV	Levels III & IV
State	All Students	201908	6.35%	13.13%	55.42%	25.11%	80.52%
Gender	Female	98465	5.87%	13.29%	55.58%	25.26%	80.84%
	Male	103443	6.81%	12.97%	55.26%	24.96%	80.22%
Ethnicity	American Indian or Alaska Native	1016	9.94%	19.98%	56.69%	13.39%	70.08%
	Asian	14367	2.10%	5.41%	43.98%	48.51%	92.49%
	Black or African-American	39984	12.53%	20.99%	52.94%	13.54%	66.48%
	Hispanic or Latino	42108	10.02%	18.45%	54.71%	16.81%	71.53%
	Native Hawaiian/ Other Pacific Islander	40	2.50%	20.00%	52.50%	25.00%	77.50%
	White	104380	3.05%	8.96%	58.21%	29.78%	87.99%
Need/ Resource Category	New York City	73451	9.16%	15.58%	50.84%	24.43%	75.27%
	Big 4 Cities	8221	19.82%	27.43%	44.65%	8.10%	52.76%
	High Need Urban/Suburban	16116	7.59%	17.05%	58.64%	16.72%	75.36%
	High Need Rural	11560	4.70%	15.29%	64.20%	15.80%	80.01%
	Average Need	59750	3.13%	10.17%	60.41%	26.29%	86.70%
	Low Need	30045	1.39%	5.30%	54.72%	38.58%	93.30%
	Charter	2050	7.51%	21.37%	60.00%	11.12%	71.12%
Student With Disability	All Codes	26796	23.68%	26.25%	43.48%	6.59%	50.07%
Accommodation	All Codes	38106	19.97%	25.32%	47.19%	7.52%	54.71%
LEP	LEP = Y and Test language = English	16193	15.20%	24.16%	51.25%	9.38%	60.63%
Test Language	Chinese	183	2.19%	10.38%	57.92%	29.51%	87.43%
	Haitian-Creole	52	30.77%	30.77%	32.69%	5.77%	38.46%
	Korean	68	2.94%	8.82%	57.35%	30.88%	88.24%
	Russian	50	22.00%	16.00%	50.00%	12.00%	62.00%
	Spanish	3353	23.26%	28.81%	42.50%	5.43%	47.93%
	All Translations	3706	21.94%	27.39%	43.50%	7.18%	50.67%

Grade 4

Performance level summaries and N-counts of demographic groups for grade 4 are presented in Table 41, below. Statewide, 78% of the 4th grade population was placed in the Meeting Learning Standards or Meeting Learning Standards with Distinction (Levels III and IV). Over 10% of American Indian or Alaska Native, Black or African-American, Hispanic or Latino, and Native Hawaiian/Other Pacific Islander students are Not Meeting Standards (Level I), as compared to only 3% of Asian students and 4% of White students. American Indian or Alaska Native, Black or African-American, and Hispanic or Latino ethnic subgroups had percent of students meeting standards (Levels III and IV) ranging from 62-70%, but the percent of the White and Asian subgroups students meeting standards (86% and 92%) exceeded the population. Student achievement also varied widely by Need/Resource Category subgroup. Almost 93% of students from Low Need districts are meeting standards, but only about 56% Big 4 Cities students are. Less than half of students with disability status or testing accommodations or those who took translated test forms met or exceeded the Level III cut; however, the subgroups for Chinese and Korean translations had very high percent of students meeting standards (86% and 89%). The following subgroups had a higher percent of students meeting standards than the State population: Female, Asian, Native Hawaiian/Other Pacific Islander, White, Average Need, Low Need, Chinese translation and Korean translation.

Table 41. Performance Level Summary, By Subgroup, Grade 4

Demographic Category (Subgroup)		N	Level I	Level II	Level III	Level IV	Levels III & IV
State	All Students	202695	7.41%	14.59%	52.12%	25.88%	78.00%
Gender	Female	98692	7.01%	15.51%	52.94%	24.54%	77.48%
	Male	104003	7.79%	13.72%	51.34%	27.15%	78.49%
Ethnicity	American Indian or Alaska Native	971	11.84%	18.64%	54.38%	15.14%	69.52%
	Asian	14482	2.69%	5.59%	42.56%	49.16%	91.71%
	Black or African-American	38555	13.69%	23.82%	50.01%	12.49%	62.50%
	Hispanic or Latino	41541	12.09%	21.29%	51.84%	14.78%	66.62%
	Native Hawaiian/Other Pacific Islander	50	10.00%	8.00%	40.00%	42.00%	82.00%
	White	107092	3.92%	9.86%	54.27%	31.95%	86.22%

(Continued on next page)

Table 41. Performance Level Summary, By Subgroup, Grade 4 (cont.)

Demographic Category (Subgroup)		N	Level I	Level II	Level III	Level IV	Levels III & IV
Need/ Resource Category	New York City	72565	10.29%	18.85%	49.12%	21.74%	70.86%
	Big 4 Cities	8022	19.22%	24.46%	45.69%	10.63%	56.32%
	High Need Urban/Suburban	16090	9.89%	18.40%	53.30%	18.42%	71.72%
	High Need Rural	11741	7.00%	16.15%	59.02%	17.83%	76.85%
	Average Need	60888	4.17%	11.22%	56.54%	28.07%	84.61%
	Low Need	31122	1.75%	5.64%	49.44%	43.17%	92.61%
	Charter	1590	10.82%	22.33%	53.14%	13.71%	66.86%
Student With Disability	All Codes	28712	28.60%	26.52%	39.09%	5.79%	44.88%
Accommodation	All Codes	39088	24.67%	26.47%	42.15%	6.71%	48.87%
LEP	LEP = Y and Test language = English	12512	21.34%	27.49%	43.37%	7.80%	51.17%
Test Language	Chinese	246	2.44%	11.38%	56.91%	29.27%	86.18%
	Haitian-Creole	47	31.91%	42.55%	25.53%	0.00%	25.53%
	Korean	75	2.67%	8.00%	49.33%	40.00%	89.33%
	Russian	50	28.00%	16.00%	42.00%	14.00%	56.00%
	Spanish	2889	30.29%	29.84%	35.76%	4.12%	39.88%
	All Translations	3307	27.58%	27.94%	37.59%	6.89%	44.48%

Grade 5

Performance level summaries and N-counts of demographic groups for grade 5 are presented in Table 42. Statewide, 68% of the 5th grade population was placed in the Meeting Learning Standards or Meeting Learning Standards with Distinction (Levels III and IV), 21% in Partially Meeting Learning Standards (Level II), and 10% in Not Meeting Learning Standards (Level I). Eleven percent of Male students were in Level I, compared to 9% of Female students, but overall there was little performance differentiation by gender subgroup. However, across ethnic and test translation subgroups, there were marked differences. Over 15% of Black or African-American and Hispanic or Latino students were in Level I, as compared to less than 4% of Asian students and 6% of White students. American Indian or Alaska Native, Black or African-American, and Hispanic or Latino ethnic subgroups had relatively few students meeting standards (Levels III and IV ranged from 49-56%), as compared to the percent of White and Asian students meeting standards (88% and 78%). Nearly 90% of students from Low Need districts were in Levels III or IV, but only slightly more than 36% of the Big 4 Cities students were. Only 3-4% of students with disability status or testing accommodations were placed in Level IV, compared to the population's 19.16% in Level IV. Less than 7% of students that took translated test forms or that reported LEP with English language test forms were placed in Level IV, except for the Chinese and Korean translation subgroups which had very high percents of students in Level IV (28.52% and

50.63%). The following subgroups had a higher percent of students meeting standards than the State population: Male, Asian, Native Hawaiian/Other Pacific Islander, White, Average Need, Low Need, Chinese translation and Korean translation.

Table 42. Performance Level Summary, By Subgroup, Grade 5

Demographic Category (Subgroup)		N	Level I	Level II	Level III	Level IV	Levels III & IV
State	All Students	209200	10.29%	21.24%	49.31%	19.16%	68.47%
Gender	Female	102743	9.44%	22.39%	49.92%	18.25%	68.17%
	Male	106457	11.11%	20.13%	48.72%	20.03%	68.76%
Ethnicity	American Indian or Alaska Native	1059	14.92%	28.99%	44.95%	11.14%	56.09%
	Asian	14543	3.52%	8.84%	45.15%	42.49%	87.64%
	Black or African-American	40809	19.63%	30.98%	41.76%	7.63%	49.39%
	Hispanic or Latino	42689	15.13%	28.79%	45.73%	10.36%	56.09%
	Native Hawaiian/Other Pacific Islander	55	9.09%	18.18%	45.45%	27.27%	72.73%
	White	110041	5.80%	16.27%	54.09%	23.83%	77.92%
Need/Resource Category	New York City	74494	13.29%	25.40%	44.40%	16.91%	61.31%
	Big 4 Cities	8421	32.03%	31.94%	30.15%	5.88%	36.03%
	High Need Urban/Suburban	16231	14.02%	26.46%	47.89%	11.63%	59.52%
	High Need Rural	12167	9.85%	25.59%	52.61%	11.95%	64.56%
	Average Need	63158	6.12%	18.09%	55.60%	20.19%	75.79%
	Low Need	31496	2.63%	10.01%	53.65%	33.71%	87.36%
	Charter	2444	11.82%	27.21%	50.29%	10.68%	60.97%
Student With Disability	All Codes	30018	35.68%	32.68%	28.55%	3.09%	31.64%
Accommodation	All Codes	39987	31.08%	32.67%	32.09%	4.16%	36.25%
LEP	LEP = Y and Test language = English	10905	27.09%	34.31%	34.03%	4.58%	38.61%
Test Language	Chinese	291	6.53%	14.09%	50.86%	28.52%	79.38%
	Haitian-Creole	59	47.46%	30.51%	18.64%	3.39%	22.03%
	Korean	79	0.00%	8.86%	40.51%	50.63%	91.14%
	Russian	61	29.51%	16.39%	47.54%	6.56%	54.10%
	Spanish	3023	39.27%	33.61%	24.88%	2.25%	27.13%
	All Translations	3513	35.64%	31.08%	27.67%	5.61%	33.28%

Grade 6

Performance level summaries and N-counts of demographic groups for grade 6 are presented in Table 43. Statewide, 60.45% of the 6th grade population was placed in the Meeting Learning Standards or Meeting Learning Standards with Distinction (Levels III and IV), 26.23% in Partially Meeting Learning Standards (Level II), and 13.32% in Not Meeting Learning Standards (Level I). Fourteen percent of Male students were in Level I, compared to 12% of Female students, but overall there was little performance differentiation by gender subgroup. However, across ethnic and test translation subgroups, there were marked differences. Over 20% of Black or African-American and Hispanic or Latino students were in Level I, as compared to 4% of Asian students and 7% of White students. Black or African-American and Hispanic or Latino ethnic subgroups had meeting standards rates (Levels III and IV) of 40.55% and 45.18%, with less than 6% of those students in Level IV, whereas 83% of Asian students were meeting standards and almost 35% were in Level IV. Over 80% of students from Low Need districts were in Levels III or IV, but only slightly more than 31% of the Big 4 Cities students were. Approximately 2% of students with disability status or testing accommodations were placed in Level IV, but over 40% were in Level I. Less than 5% of students that took translated test forms or that reported LEP with English language test forms were placed in Level IV, except for the Chinese and Korean translation subgroups which had very high rates (23.26% and 26.32%). The following subgroups had a higher percent of students meeting standards than the State population: Female, Asian, Native Hawaiian/Other Pacific Islander, White, Average Need, Low Need, Chinese translation and Korean translation.

Table 43. Performance Level Summary, By Subgroup, Grade 6

Demographic Category (Subgroup)		N	Level I	Level II	Level III	Level IV	Levels III & IV
State	All Students	211376	13.32%	26.23%	47.26%	13.19%	60.45%
Gender	Female	102985	12.26%	27.18%	48.05%	12.50%	60.56%
	Male	108391	14.32%	25.33%	46.51%	13.85%	60.35%
Ethnicity	American Indian or Alaska Native	1138	19.68%	32.95%	41.30%	6.06%	47.36%
	Asian	14130	4.32%	12.24%	48.89%	34.56%	83.45%
	Black or African-American	41827	23.97%	35.48%	35.91%	4.63%	40.55%
	Hispanic or Latino	41960	21.38%	33.44%	39.39%	5.79%	45.18%
	Native Hawaiian/Other Pacific Islander	44	6.82%	22.73%	50.00%	20.45%	70.45%
	White	112276	7.40%	21.78%	54.28%	16.53%	70.81%

(Continued on next page)

Table 43. Performance Level Summary, By Subgroup, Grade 6 (cont.)

Demographic Category (Subgroup)		N	Level I	Level II	Level III	Level IV	Levels III & IV
Need/ Resource Category	New York City	73988	17.99%	29.34%	40.99%	11.68%	52.66%
	Big 4 Cities	8551	32.78%	35.53%	28.30%	3.39%	31.69%
	High Need Urban/Suburban	16586	18.91%	33.52%	41.12%	6.45%	47.57%
	High Need Rural	12922	12.28%	32.08%	48.17%	7.46%	55.63%
	Average Need	65173	8.19%	24.11%	54.36%	13.34%	67.70%
	Low Need	31608	3.44%	14.41%	56.55%	25.60%	82.15%
	Charter	1719	18.91%	32.46%	41.36%	7.27%	48.63%
Student With Disability	All Codes	29778	44.21%	34.11%	20.17%	1.51%	21.68%
Accommodation	All Codes	36777	40.50%	34.70%	22.70%	2.11%	24.81%
LEP	LEP = Y and Test language = English	8746	37.69%	34.84%	24.42%	3.05%	27.48%
Test Language	Chinese	288	5.21%	18.06%	53.47%	23.26%	76.74%
	Haitian-Creole	62	46.77%	35.48%	14.52%	3.23%	17.74%
	Korean	95	1.05%	15.79%	56.84%	26.32%	83.16%
	Russian	48	14.58%	47.92%	33.33%	4.17%	37.50%
	Spanish	2960	50.54%	32.64%	15.71%	1.11%	16.82%
	All Translations	3453	44.83%	31.22%	20.21%	3.74%	23.95%

Grade 7

Performance level summaries and N-counts of demographic groups for grade 7 are presented in Table 44. Statewide, 55.69% of the 7th grade population was placed in the Meeting Learning Standards or Meeting Learning Standards with Distinction (Levels III and IV), 31.12% in Partially Meeting Learning Standards (Level II), and 13.19% in Not Meeting Learning Standards (Level I). Fourteen percent of Male students were in Level I, compared to 12% of Female students, but overall there was only slight performance differentiation by gender subgroup. However, across ethnic and test translation subgroups, there were marked differences. Over 25.74% of Black or African-American and 21.35% of Hispanic or Latino students were in Level I, as compared to 5.02% of Asian students and 6.4% of White students. Black or African-American and Hispanic or Latino ethnic subgroups had of 31.13% and 37.19% of students meeting standards (Levels III and IV), with less than 4% of those students in Level IV, whereas over 78% of Asian students were meeting standards and almost 30% were in Level IV. About 20% of Big 4 Cities students were meeting standards, with less than 2% in Level IV, yet over 82% of students from Low Need districts were meeting standards with about 28% in Level IV. Less than 2% of students with disability status or testing accommodations were placed in Level IV, but nearly 40% were in Level I. Less than 3% of students that took translated test forms or that reported LEP with English language test forms were placed in Level IV, except for the Chinese and Korean translation subgroups which had

very high rates (14.24% and 34.29%). Across all subgroups, the Haitian-Creole translation subgroup had the largest percent of its students placed in Level I (51.29%), and the Korean translation subgroup had the largest percent its of students placed in Level IV. The following subgroups had percent of students meeting standards above the population: Female, Asian, Native Hawaiian/Other Pacific Islander, White, Average Need, Low Need, Chinese translation and Korean translation.

Table 44. Performance Level Summary, By Subgroup, Grade 7

Demographic Category (Subgroup)		N	Level I	Level II	Level III	Level IV	Levels III & IV
State	All Students	217225	13.19%	31.12%	43.52%	12.17%	55.69%
Gender	Female	105202	12.06%	31.32%	44.43%	12.19%	56.62%
	Male	112022	14.25%	30.93%	42.67%	12.15%	54.82%
Ethnicity	American Indian or Alaska Native	1120	18.84%	37.41%	38.48%	5.27%	43.75%
	Asian	13911	5.02%	16.40%	49.01%	29.56%	78.57%
	Black or African-American	43437	25.74%	43.14%	28.40%	2.73%	31.13%
	Hispanic or Latino	42824	21.35%	41.47%	33.64%	3.55%	37.19%
	Native Hawaiian/Other Pacific Islander	49	10.20%	24.49%	57.14%	8.16%	65.31%
	White	115882	6.40%	24.50%	52.23%	16.88%	69.10%
Need/Resource Category	New York City	75423	18.75%	37.37%	36.07%	7.81%	43.88%
	Big 4 Cities	9446	36.35%	43.57%	18.51%	1.57%	20.07%
	High Need Urban/Suburban	16824	18.30%	41.53%	35.49%	4.68%	40.17%
	High Need Rural	13633	11.68%	36.26%	45.34%	6.73%	52.06%
	Average Need	67836	6.72%	26.16%	52.50%	14.62%	67.13%
	Low Need	31658	2.93%	15.05%	54.49%	27.53%	82.02%
	Charter	1336	20.21%	40.72%	35.48%	3.59%	39.07%
Student With Disability	All Codes	29564	41.74%	40.04%	17.08%	1.13%	18.21%
Accommodation	All Codes	36685	38.13%	40.67%	19.63%	1.58%	21.20%
LEP	LEP = Y and Test language = English	9714	37.28%	41.46%	19.18%	2.09%	21.27%
Test Language	Chinese	316	9.18%	18.99%	57.59%	14.24%	71.84%
	Haitian-Creole	80	51.25%	38.75%	8.75%	1.25%	10.00%
	Korean	105	5.71%	15.24%	44.76%	34.29%	79.05%
	Russian	73	41.10%	38.36%	19.18%	1.37%	20.55%
	Spanish	3266	45.44%	39.90%	13.99%	0.67%	14.67%
	All Translations	3840	41.41%	37.45%	18.41%	2.73%	21.15%

Grade 8

Performance level summaries and N-counts of demographic groups for grade 8 are presented in Table 45. Statewide, 53.93% of the 8th grade population was placed in the Meeting Learning Standards or Meeting Learning Standards with Distinction (Levels III and IV), 31.09% in Partially Meeting Learning Standards (Level II), and 14.98% in Not Meeting Learning Standards (Level I). Sixteen percent of Male students were in Level I, compared to 14% of Female students, but overall there was little performance differentiation by gender subgroup. Across ethnic and test translation subgroups, there were marked differences in performance. Almost 28% of Black or African-American and 33% of Hispanic or Latino students were in Level I, compared to less than 7% of Asian and White students. Black or African-American and Hispanic or Latino ethnic subgroups had 28.34% and 32.99% of students meeting standards (Levels III and IV), with less than 3% of those students in Level IV, whereas 77% of Asian students were meeting standards and almost 30% were in Level IV. About 21% of Big 4 Cities students were in Levels III and IV, with less than 2% in Level IV, yet over 82% of students from Low Need districts passed with about 23% in Level IV. Approximately 44% of students with disability status and 40% with testing accommodations were placed in Level I, 39% in Level II, and less than 2% in Level IV. Less than 4% of students that took translated test forms or that reported LEP with English language test forms were placed in Level IV, except for the Chinese and Korean translation subgroups which had very high percent of students in Level IV (22.5% and 19.05%). Across all subgroups, the Haitian-Creole translation subgroup had the largest percent of its students placed in Level I (3.97%), and the Asian subgroup had the largest percent its of students placed in Level IV (29.81%). The following subgroups had percent of students meeting standards above the 8th grade population: Female, Asian, Native Hawaiian/Other Pacific Islander, White, Average Need, Low Need, Chinese translation and Korean translation.

Table 45. Performance Level Summary, By Subgroup, Grade 8

Demographic Category (Subgroup)		N	Level I	Level II	Level III	Level IV	Levels III & IV
State	All Students	219294	14.98%	31.09%	43.74%	10.18%	53.93%
Gender	Female	107180	13.67%	31.11%	44.81%	10.41%	55.22%
	Male	112113	16.24%	31.07%	42.72%	9.97%	52.69%
Ethnicity	American Indian or Alaska Native	1073	21.16%	37.37%	36.91%	4.57%	41.47%
	Asian	13994	5.79%	17.22%	47.18%	29.81%	76.99%
	Black or African-American	43407	29.88%	41.78%	26.03%	2.31%	28.34%
	Hispanic or Latino	42116	25.83%	41.18%	30.02%	2.97%	32.99%
	Native Hawaiian/Other Pacific Islander	33	9.09%	33.33%	42.42%	15.15%	57.58%
	White	118668	6.71%	25.18%	54.75%	13.36%	68.11%

(Continued on next page)

Table 45. Performance Level Summary, By Subgroup, Grade 8 (cont.)

Demographic Category (Subgroup)		N	Level I	Level II	Level III	Level IV	Levels III & IV
Need/ Resource Category	New York City	76162	24.14%	36.93%	31.56%	7.36%	38.92%
	Big 4 Cities	9716	34.45%	44.67%	19.43%	1.45%	20.88%
	High Need Urban/Suburban	16576	19.05%	40.82%	35.83%	4.30%	40.13%
	High Need Rural	13591	10.71%	36.42%	47.63%	5.24%	52.87%
	Average Need	69213	6.39%	26.88%	55.33%	11.41%	66.73%
	Low Need	31713	2.91%	14.67%	59.59%	22.83%	82.41%
	Charter	978	25.15%	38.85%	33.54%	2.45%	35.99%
Student With Disability	All Codes	29541	44.03%	38.75%	16.69%	0.53%	17.22%
Accommodation	All Codes	36773	39.89%	38.86%	20.04%	1.22%	21.25%
LEP	LEP = Y and Test language = English	9277	40.26%	37.87%	20.09%	1.78%	21.87%
Test Language	Chinese	440	6.36%	22.50%	48.64%	22.50%	71.14%
	Haitian-Creole	126	53.97%	36.51%	7.94%	1.59%	9.52%
	Korean	126	6.35%	13.49%	61.11%	19.05%	80.16%
	Russian	87	18.39%	34.48%	43.68%	3.45%	47.13%
	Spanish	3452	44.55%	40.24%	14.92%	0.29%	15.21%
	All Translations	4231	39.19%	37.37%	20.18%	3.26%	23.45%

Section X: References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: 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., & Aitkin, M. (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, Version]. Unpublished.
- Burket, G. R. (2002). *PARDUX* [Computer program, Version 1.26]. Unpublished.
- 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., Schmitt, A. P., & Bleistein, C. A. (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*. Unpublished manuscript
- Fitzpatrick, A. R. (1994). *Two studies comparing parameter estimates produced by PARDUX and BIGSTEPS*. Unpublished manuscript.
- Fitzpatrick, A. R. & Julian, M. W. (1996). *Two studies comparing the parameter estimates produced by PARDUX and PARSCLE*. Monterey, CA: CTB/McGraw-Hill.
- Fitzpatrick, A. R., Link, V., Yen, W. M., Burket, G., Ito, K., & Sykes, R. (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., Yen, W. M., & Burket, G. R. (1989). Experiences in the application of item response theory in test construction. *Applied Measurement in Education*, 2, 297-312.
- Hambleton, R. K. Clouser, B. E. Mazor, K. M. & Jones, R. W. (1993) Advances in the detection of differentially functioning test items. *European Journal of Psychological Assessment*, Vol 9, Issue 1 pp. 1-18.
- Huynh, H. & Schneider, C. (2004). Vertically Moderated Standards as an Alternative to VerticalScaling: Assumptions, Practices, and an Odyssey through NAEP. Paper

- presented at the National Conference on Large-Scale Assessment, June 21, 2004, Boston, MA.
- Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.
- Johnson, N. L. & Kotz, S. (1970). *Distributions in statistics: Continuous univariate distributions, Vol. 2*. New York: John Wiley.
- Karkee T., Lewis, D., Barton, K., & Haug, C. (2002). The effect of including or excluding students with testing accommodations on IRT calibrations. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Kim, D. (2004). WLCLASS [Computer program]. Unpublished
- Kolen, M. J. & Brennan R. L. (1995). *Test equating. Methods and practices*. New York, NY: Springer-Verlag.
- Lee, W., Hanson, B. A., & Brennan R. L. (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 Harnisch, D. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement, 18*, pp. 109–118.
- Livingston, S. A., & Lewis, C. (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., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Menlo Park, CA: Addison-Wesley.
- Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and Evaluation in Education and Psychology, 3rd ed.* New York: Holt, Rinehart, and Winston.
- Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The Bookmark procedure: Psychological perspectives. In Cizek, G. J. (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249-281). Mahwah, NJ: Lawrence Erlbaum Associates.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159–176.
- Muraki, E., & Bock, R. D. (1991). *PARSCALE: Parameter Scaling of Rating Data* [Computer program]. Chicago, IL: Scientific Software, Inc.
- Novick, M. R. & Jackson, P. H. (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. Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, 1979, 4, 207-230.
- Sandoval, J. H., & Mille, M. P. (1979, August). *Accuracy of judgments of WISC-R item difficulty for minority groups*. Paper presented at the annual meeting of the American Psychological Association, New York.
- Stocking, M. L., & Lord, F. M. (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., Kolen, M. J., & Harris, D. J. (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., & Linacre, J. M. (1992). *BIGSTEPS Rasch Analysis* [Computer program]. Chicago, IL: MESA Press.
- Yen, W. M. (1997). The technical quality of performance assessments: Standard errors of percents of students reaching standards. *Educational Measurement: Issues and Practice*, 5-15.
- Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30, 187-213.
- Yen, W. M. (1984). Obtaining maximum likelihood trait estimates from number correct scores for the three-parameter logistic model. *Journal of Educational Measurement*, 21, 93-111.
- Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, 5, 245-262.
- Yen, W. M. Sykes, R. C., Ito, K., & Julian, M. (1997, March). 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.
- Yen, W., Sykes, R. C., Ito, K., and Julian M. (1997). A Bayesian/IRT Index of Objective Performance for Tests with Mixed-Item Types. Paper presented at the National Council on Measurement in Education in Chicago, IL, 1997
- Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, 36, 233-25.

Appendices: 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 distracters
- 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 distracters
- 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 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

Appendices: Appendix B – Psychometric Guidelines for Operational Item Selection

It is primarily up to the Content Development to select items for the 2006 Operational Test. Research will provide support, as necessary, and will review the final item selection. Research will provide DAT files with parameters for all FT items eligible for item pool. The pools of items eligible for 2006 item selection will include 2005 FT items for grades 3, 5, 6 and 7 and 2003 and 2005 FT items for grades 4 and 8. All items for each grade will be on the same (grade specific) scale.

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% of the %s of score points and items per objective.
- Avoid selecting poor-fitting items, items with too high/low p-values, items with flagged point biserials; Research will provide a list of such items.
- Minimize the number of items flagged for DIF (gender, ethnic, and high/low needs schools). Flagged items should be reviewed for content again. It needs to be remembered that some items maybe flagged for DIF by chance only and their content may not necessary 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’s 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 operational forms (e.g., the first item in a FT form should also be the first item in an operational form). When that is impossible, please ensure that they are in the same one-third section of the forms.
- The target is the OP test blueprint.
- Research will provide a comprehensive summary of item flagged for different reasons (difficulty, DIF, misfit, calibration problems etc), along with recommendation as to which items should be avoided when selecting OP test forms
- After selecting OP forms, please submit the following to Research for our review:
 - List and order of items on the OP form (item parameters, items IDs)
 - Content coverage sheet
 - Plot of TCCs
 - Plot of SEM (include SEM for total item pool)
 - Item #s and the percent of proposed items and score points flagged for gender and ethnic DIF
 - Item #s and the percent of proposed items and score points that have poor model-to-data fit
 - .SUM files from the proposed selections

Appendices: Appendix C – Factor Analysis Results

As described in Section III (Validity) a Principal Component factor analysis was conducted on the Grades 3-8 Math 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). This Appendix contains figures of scree plots obtained from the analysis of the total population and subpopulation data Math data and a table of eigenvalues and proportion of variance accounted for by extracted factors for subgroups.

Figure C1. Grade 3 Scree Plot (Total Population)

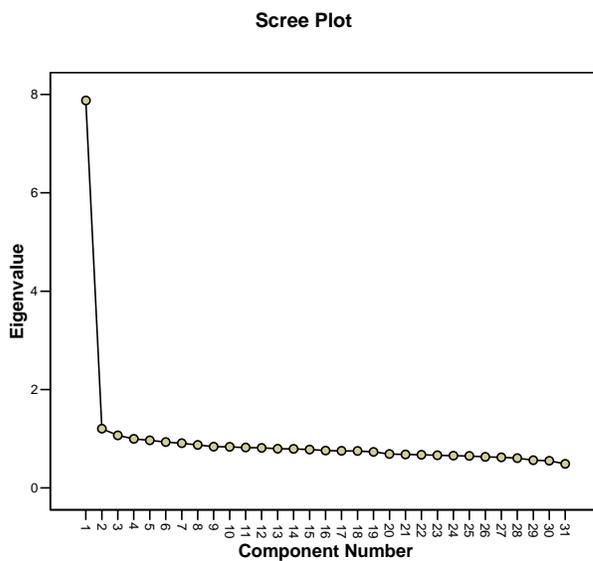


Figure C2. Grade 3 Scree Plot (LEP Students)

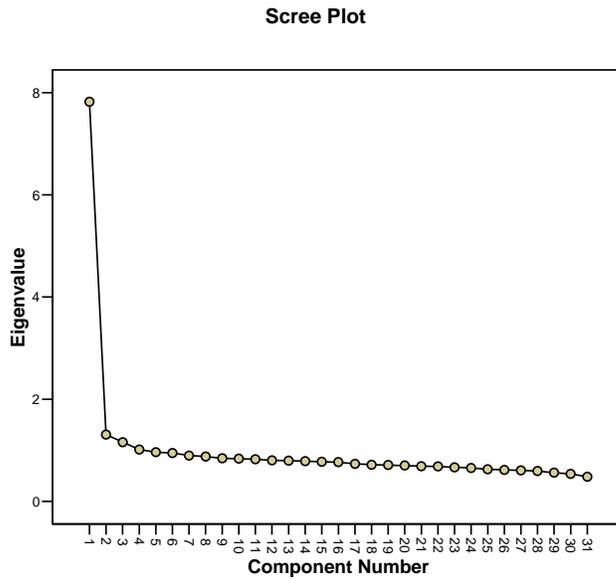


Figure C3. Grade 3 Scree Plot (Students with Disabilities)

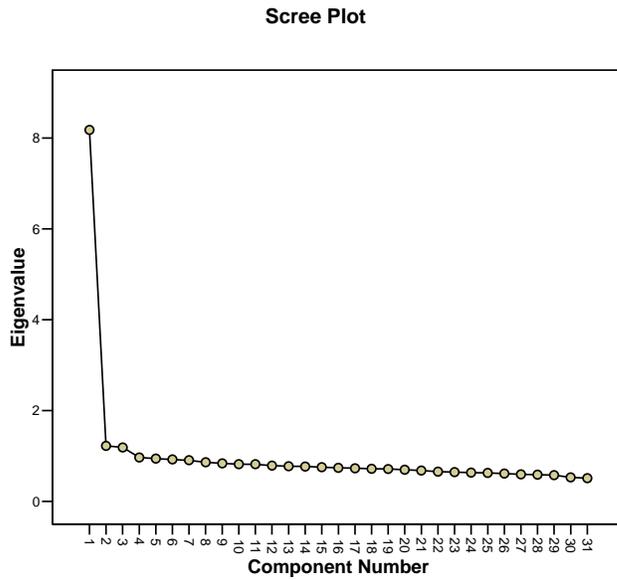


Figure C4. Grade 3 Scree Plot (Students with Accommodations)

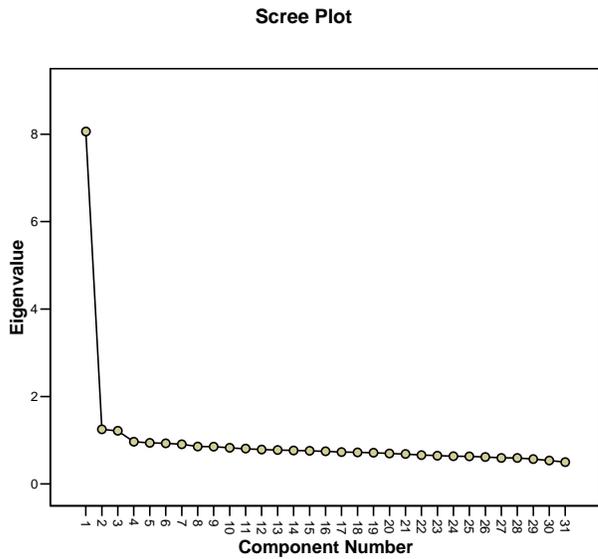


Figure C5. Grade 4 Scree Plot (Total Population)

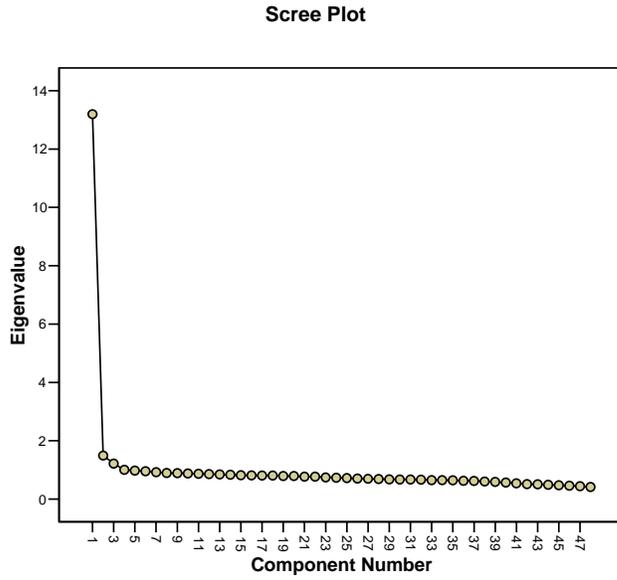


Figure C6. Grade 4 Scree Plot (LEP Students)

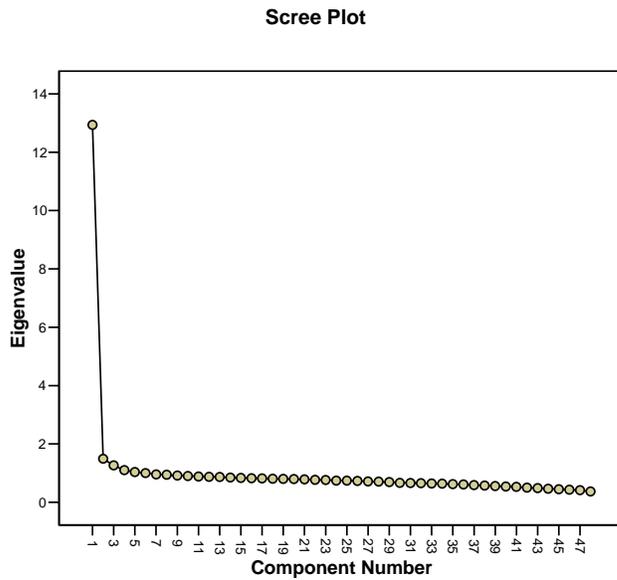


Figure C7. Grade 4 Scree Plot (Students with Disabilities)

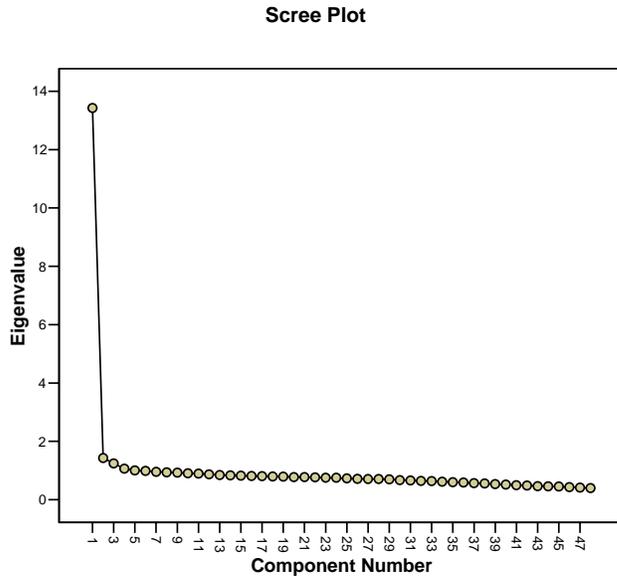


Figure C8. Grade 4 Scree Plot (Students with Accommodations)

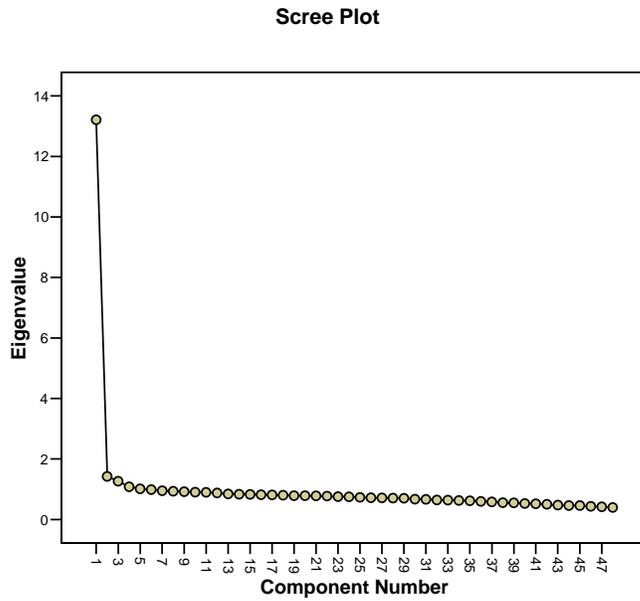


Figure C9. Grade 5 Scree Plot (Total Population)

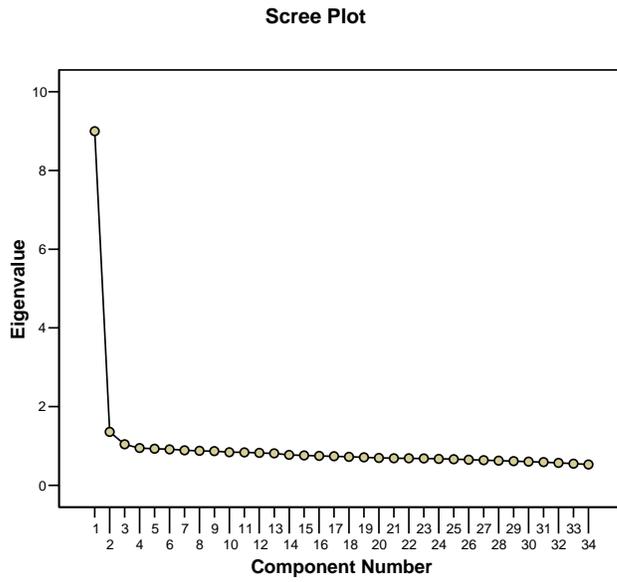


Figure C10. Grade 5 Scree Plot (LEP Students)

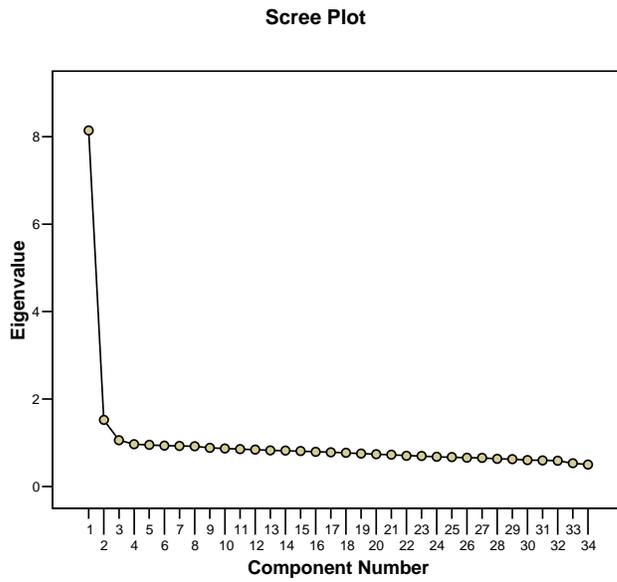


Figure C11. Grade 5 Scree Plot (Students with Disabilities)

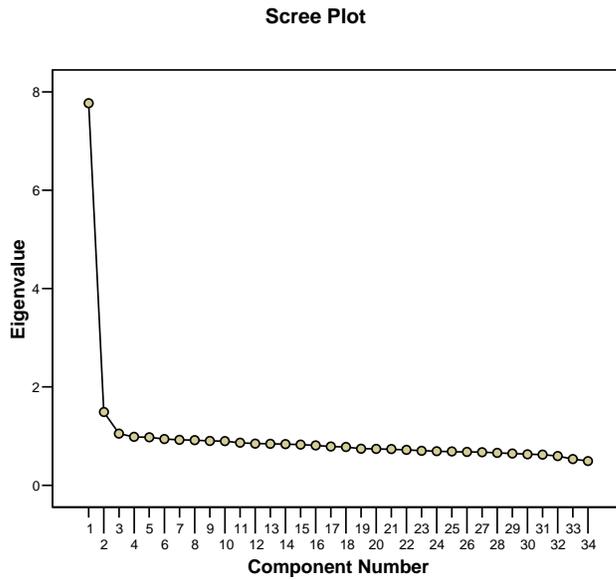


Figure C12. Grade 5 Scree Plot (Students with Accommodations)

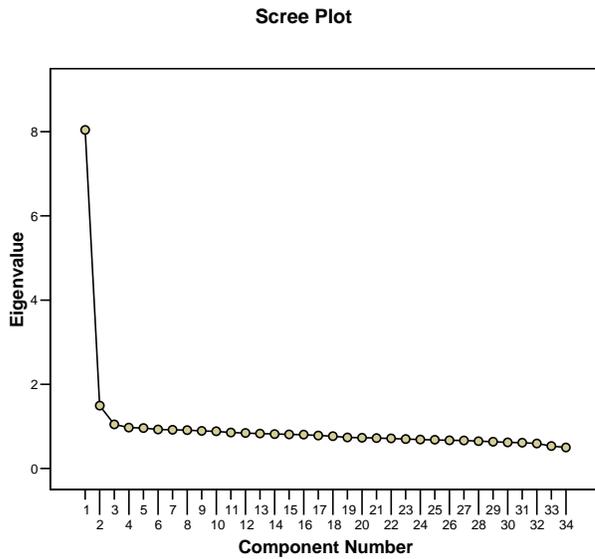


Figure C13. Grade 6 Scree Plot (Total Population)

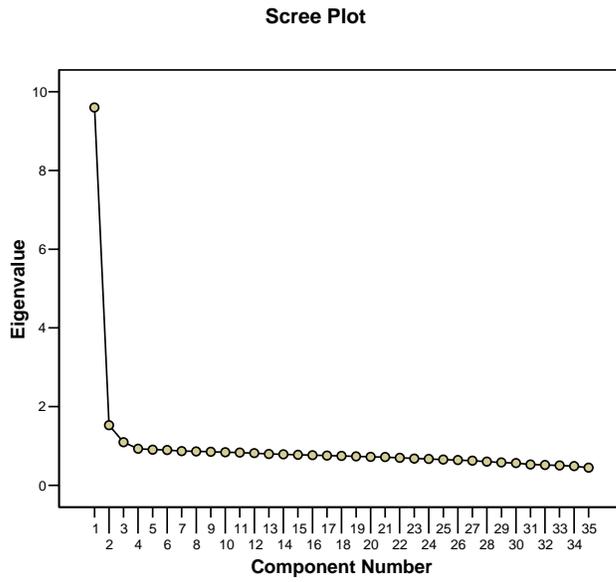


Figure C14. Grade 6 Scree Plot (LEP Students)

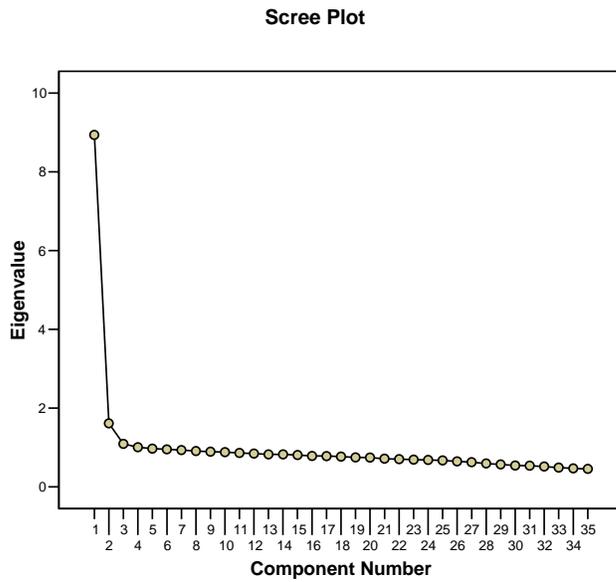


Figure C15. Grade 6 Scree Plot (Students with Disabilities)

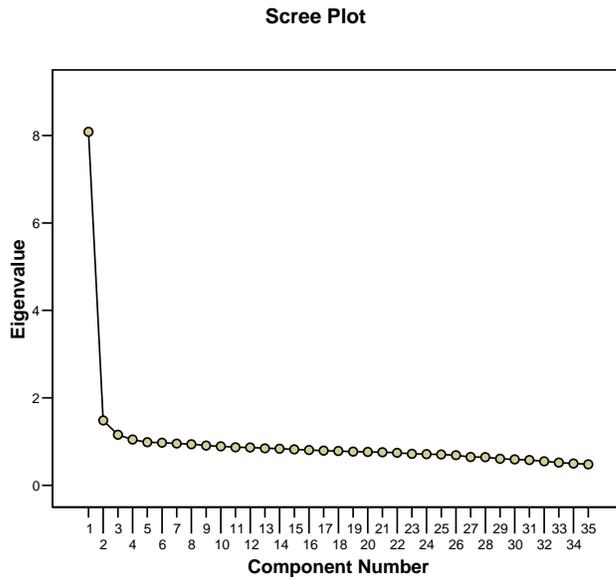


Figure C16. Grade 6 Scree Plot (Students with Accommodations)

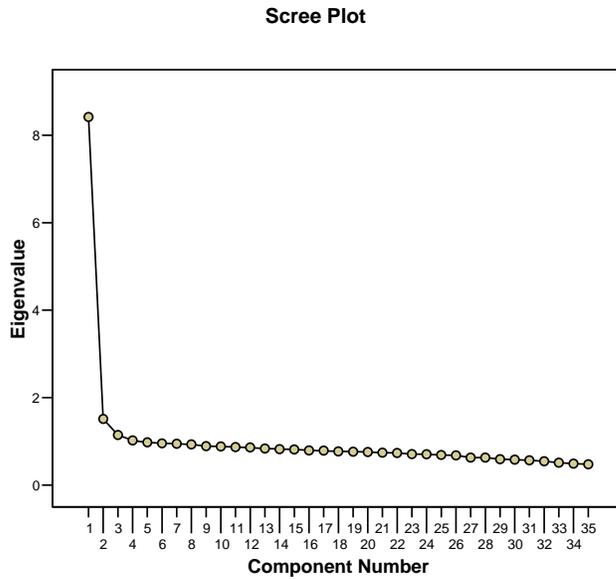


Figure C17. Grade 7 Scree Plot (Total Population)

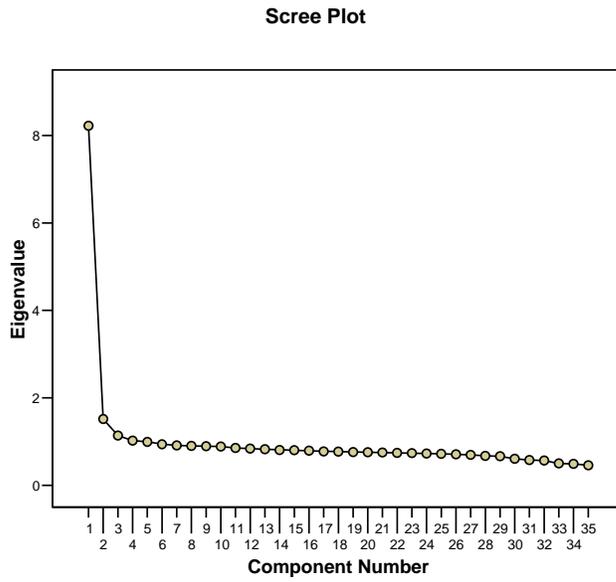


Figure C18. Grade 7 Scree Plot (LEP Students)

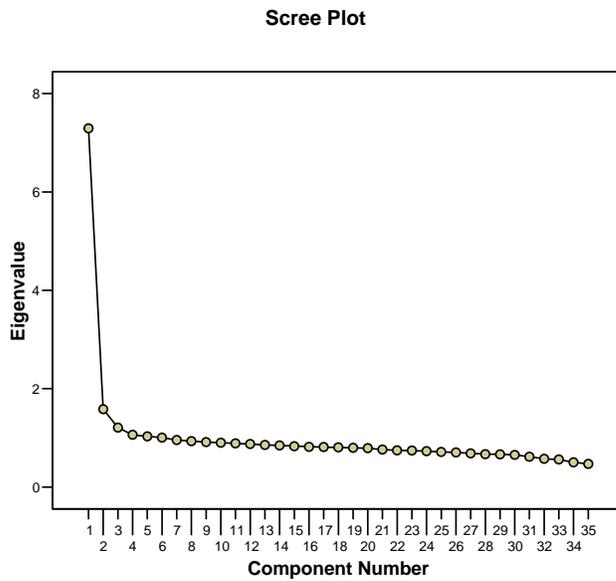


Figure C19. Grade 7 Scree Plot (Students with Disabilities)

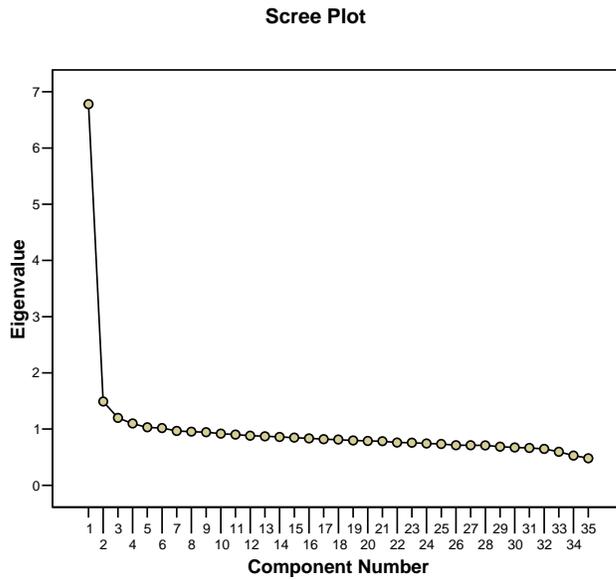


Figure C20. Grade 7 Scree Plot (Students with Accommodations)

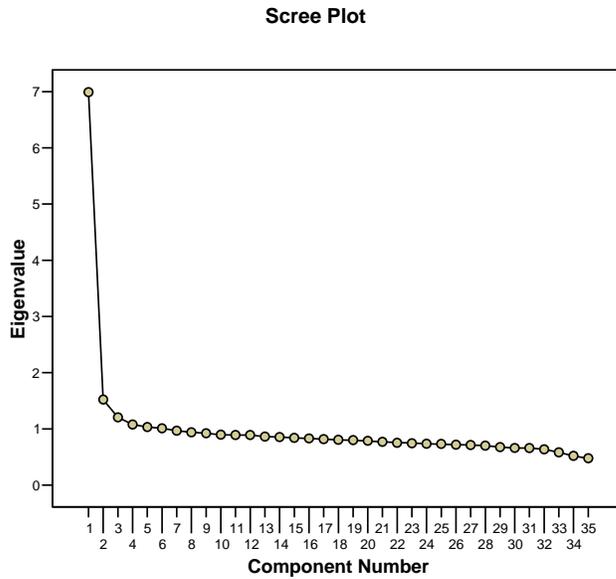


Figure C21. Grade 8 Scree Plot (Total Population)

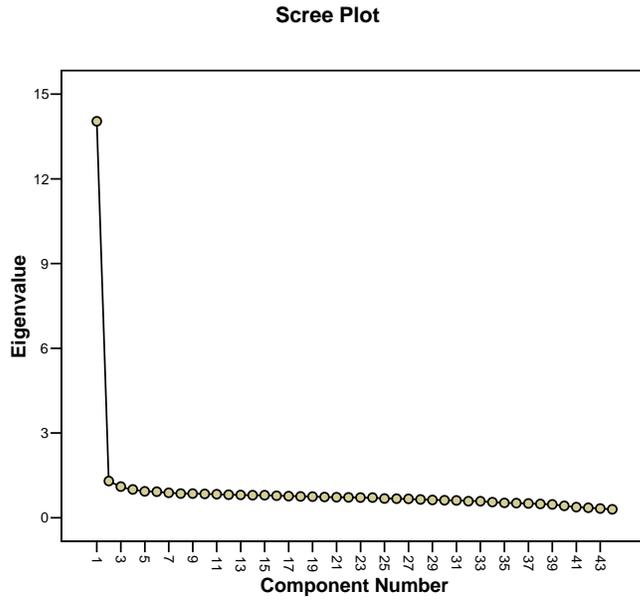


Figure C22. Grade 8 Scree Plot (LEP Students)

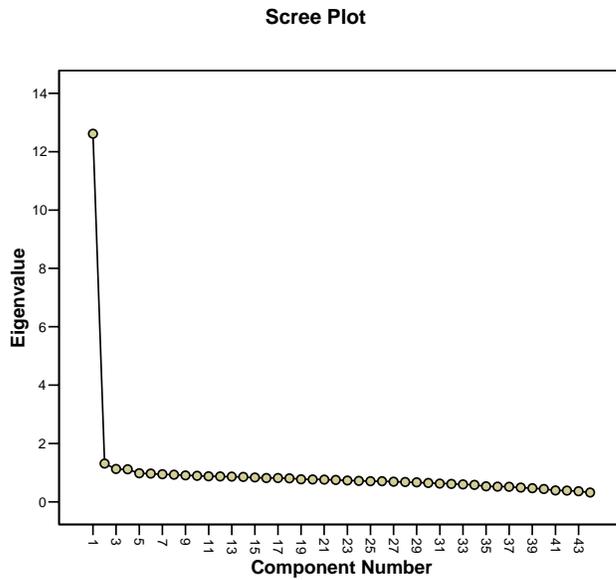


Figure C23. Grade 8 Scree Plot (Students with Disabilities)

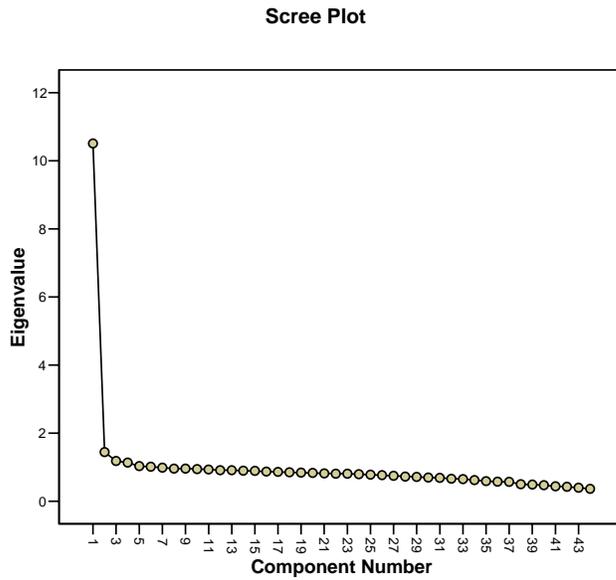


Figure C24. Grade 8 Scree Plot (Students with Accommodations)

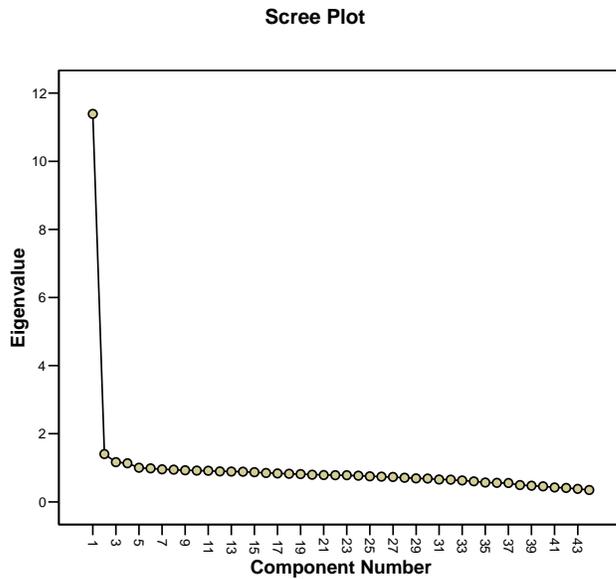


Table C1. Factor Analysis Results for MA tests (Selected Sub-Populations)

Grade	Subgroup	Initial Eigenvalues			
		Component	Total	% of Variance	Cumulative %
3	LEP	1	7.82	25.24	25.24
		2	1.31	4.22	29.45
		3	1.16	3.74	33.19
		4	1.01	3.27	36.47
	SWD	1	8.18	26.39	26.39
		2	1.22	3.94	30.33
		3	1.19	3.84	34.16
	SUA	1	8.06	26.01	26.01
		2	1.25	4.03	30.04
3		1.22	3.92	33.96	
4	LEP	1	12.93	26.94	26.94
		2	1.49	3.11	30.05
		3	1.27	2.64	32.69
		4	1.10	2.30	34.99
		5	1.04	2.17	37.15
		6	1.00	2.09	39.24
	SWD	1	13.43	27.98	27.98
		2	1.43	2.98	30.96
		3	1.24	2.59	33.55
		4	1.06	2.22	35.77
		5	1.01	2.10	37.87
	SUA	1	13.21	27.52	27.52
		2	1.42	2.97	30.49
		3	1.27	2.64	33.13
		4	1.08	2.25	35.38
5	LEP	1	8.14	23.95	23.95
		2	1.52	4.47	28.42
		3	1.06	3.11	31.53
	SWD	1	7.77	22.85	22.85
		2	1.49	4.39	27.24
		3	1.05	3.09	30.33
	SUA	1	8.04	23.65	23.65
		2	1.49	4.39	28.04
		3	1.05	3.08	31.12

(Continued on next page)

Table C1. Factor Analysis Results for MA tests (Selected Sub-Populations) (cont.)

Grade	Subgroup	Initial Eigenvalues			
		Component	Total	% of Variance	Cumulative %
6	LEP	1	8.94	25.54	25.54
		2	1.61	4.60	30.14
		3	1.09	3.12	33.26
		4	1.01	2.87	36.13
	SWD	1	8.09	23.10	23.10
		2	1.48	4.24	27.34
		3	1.16	3.30	30.65
		4	1.04	2.98	33.63
	SUA	1	8.42	24.05	24.05
		2	1.51	4.32	28.38
		3	1.14	3.27	31.65
		4	1.02	2.92	34.56
7	LEP	1	7.29	20.84	20.84
		2	1.58	4.53	25.37
		3	1.21	3.45	28.82
		4	1.06	3.03	31.85
		5	1.03	2.95	34.79
		6	1.01	2.87	37.67
	SWD	1	6.78	19.38	19.38
		2	1.49	4.26	23.64
		3	1.20	3.42	27.06
		4	1.10	3.14	30.20
		5	1.03	2.95	33.15
		6	1.02	2.91	36.05
	SUA	1	6.99	19.97	19.97
		2	1.52	4.35	24.32
		3	1.21	3.44	27.76
		4	1.08	3.08	30.84
		5	1.03	2.95	33.79
		6	1.01	2.89	36.68

(Continued on next page)

Table C1. Factor Analysis Results for MA tests (Selected Sub-Populations) (cont.)

Grade	Subgroup	Initial Eigenvalues			
		Component	Total	% of Variance	Cumulative %
8	LEP	1	12.62	28.67	28.67
		2	1.31	2.98	31.66
		3	1.13	2.56	34.22
		4	1.11	2.53	36.75
	SWD	1	10.51	23.88	23.88
		2	1.44	3.28	27.15
		3	1.18	2.68	29.84
		4	1.13	2.57	32.41
		5	1.03	2.34	34.75
		6	1.01	2.30	37.05
	SUA	1	11.39	25.88	25.88
		2	1.40	3.19	29.07
		3	1.17	2.65	31.72
		4	1.13	2.57	34.30
		5	1.00	2.27	36.57

Note: LEP=Limited English Proficiency, SWD=Students with Disabilities, and SUA=Students using Accommodations

Appendices: Appendix D – DIF Statistics

These tables support the DIF information in Section V (Operational Test Data Collection and Classical Analyses) and Section VI (IRT Scaling). They include item numbers, focal group, direction of DIF and DIF statistics. Table E1 shows items flagged by SMD and Mantel- Haenszel methods and Table E2 presents items flagged by Linn-Harnisch method. Note that positive values of SMD and Delta in Table D1 indicate differential item functioning in favor of a focal group and negative values of SMD and Delta indicate differential item functioning against a focal group.

Table D1. NYSTP Math 2006 Classical DIF Item Flags

Grade	Item #	Subgroup	DIF	SMD	Mantel	Delta
3	12	Spanish	In Favor	0.11	No flag	No flag
3	28	Spanish	In Favor	0.12	n/a	n/a
3	29	Spanish	Against	-0.11	n/a	n/a
3	30	Black or African-American	Against	-0.19	n/a	n/a
3	30	Hispanic or Latino	Against	-0.15	n/a	n/a
3	30	High NRC	Against	-0.16	n/a	n/a
3	30	Spanish	Against	-0.13	n/a	n/a
4	1	Spanish	Against	-0.16	66.70	-2.25
4	4	Asian	In Favor	No flag	9.12	1.84
4	6	Asian	In Favor	No flag	21.08	1.53
4	14	Spanish	Against	-0.11	31.99	-1.59
4	19	Female	Against	-0.12	146.70	-1.63
4	22	Spanish	Against	-0.15	44.67	-1.69
4	30	Asian	In Favor	No flag	14.14	1.58
4	31	Spanish	In Favor	0.11	n/a	n/a
4	34	Spanish	In Favor	0.10	n/a	n/a
4	35	Female	In Favor	0.12	n/a	n/a
4	38	Black or African-American	Against	-0.13	n/a	n/a
4	38	Spanish	Against	-0.13	n/a	n/a
4	39	Asian	Against	-0.11	n/a	n/a
4	39	Spanish	Against	-0.13	n/a	n/a
4	41	Spanish	Against	-0.21	n/a	n/a
4	45	Spanish	In Favor	0.11	n/a	n/a
5	3	Spanish	Against	-0.12	61.78	-2.68
5	5	Asian	Against	-0.10	No flag	No flag
5	11	Spanish	In Favor	0.12	30.84	1.60

(Continued on next page)

Table D1. NYSTP Math 2006 Classical DIF Item Flags (cont.)

Grade	Item #	Subgroup	DIF	SMD	Mantel	Delta
5	16	Asian	In Favor	No flag	7.20	1.59
5	19	Black or African-American	Against	-0.11	No flag	No flag
5	19	Asian	Against	No flag	27.76	-1.68
5	22	Asian	In Favor	No flag	17.82	1.54
5	23	Spanish	In Favor	0.10	No flag	No flag
5	27	Spanish	In Favor	0.19	n/a	n/a
5	28	Black or African-American	In Favor	0.10	n/a	n/a
5	28	Spanish	Against	-0.20	n/a	n/a
5	29	Female	In Favor	0.11	n/a	n/a
5	29	Spanish	In Favor	0.18	n/a	n/a
5	30	Spanish	Against	-0.11	n/a	n/a
5	33	Black or African-American	Against	-0.18	n/a	n/a
5	33	Female	In Favor	0.12	n/a	n/a
5	33	High NRC	Against	-0.10	n/a	n/a
6	1	Black or African-American	In Favor	0.10	61.98	1.92
6	1	Asian	In Favor	No flag	8.56	1.89
6	1	High NRC	In Favor	No flag	64.04	1.57
6	2	Spanish	Against	-0.11	No flag	No flag
6	4	Asian	Against	No flag	27.81	-1.99
6	4	Spanish	Against	-0.14	45.05	-1.69
6	24	Asian	In Favor	0.12	30.85	1.56
6	26	Spanish	In Favor	0.12	n/a	n/a
6	28	Asian	Against	-0.18	n/a	n/a
6	29	Hispanic	Against	-0.15	n/a	n/a
6	29	High NRC	Against	-0.14	n/a	n/a
6	31	Asian	Against	-0.15	n/a	n/a
6	32	Spanish	In Favor	0.13	n/a	n/a
6	33	Spanish	In Favor	0.13	n/a	n/a
6	34	Spanish	Against	-0.12	n/a	n/a
7	2	Asian	Against	No flag	14.79	-2.27
7	5	Spanish	Against	-0.12	No flag	No flag
7	7	Spanish	Against	No flag	20.05	-1.54
7	12	Spanish	Against	-0.20	117.97	-3.1
7	13	Spanish	In Favor	0.13	35.59	1.55

(Continued on next page)

Table D1. NYSTP Math 2006 Classical DIF Item Flags (cont.)

Grade	Item #	Subgroup	DIF	SMD	Mantel	Delta
7	18	Spanish	In Favor	0.11	No flag	No flag
7	19	Asian	In Favor	0.10	No flag	No flag
7	20	Spanish	In Favor	0.18	61.37	2.28
7	26	Spanish	Against	-0.10	No flag	No flag
7	28	Spanish	In Favor	0.11	No flag	No flag
7	30	Spanish	Against	-0.10	No flag	No flag
7	31	Spanish	Against	-0.12	n/a	n/a
7	32	Black or African-American	Against	-0.11	n/a	n/a
7	32	Hispanic or Latino	Against	-0.11	n/a	n/a
7	32	High NRC	Against	-0.12	n/a	n/a
7	34	Spanish	In Favor	0.15	n/a	n/a
7	35	Spanish	In Favor	0.10	n/a	n/a
7	36	Spanish	In Favor	0.11	n/a	n/a
7	37	Asian	Against	-0.18	n/a	n/a
7	37	High NRC	Against	-0.14	n/a	n/a
7	38	Black or African-American	Against	-0.12	n/a	n/a
7	38	Hispanic or Latino	Against	-0.20	n/a	n/a
7	38	Asian	Against	-0.12	n/a	n/a
7	38	Female	Against	-0.19	n/a	n/a
7	38	High NRC	Against	-0.10	n/a	n/a
7	38	Spanish	Against	-0.19	n/a	n/a
8	5	Spanish	In Favor	0.14	47.10	1.76
8	7	Asian	In Favor	0.10	No flag	No flag
8	19	Spanish	Against	-0.10	No flag	No flag
8	28	Black or African-American	Against	-0.10	n/a	n/a
8	30	Black or African-American	Against	-0.13	n/a	n/a
8	30	Asian	Against	-0.10	n/a	n/a
8	30	Female	Against	-0.10	n/a	n/a
8	30	Spanish	Against	-0.18	n/a	n/a
8	31	High NRC	Against	-0.14	n/a	n/a

(Continued on next page)

Table D1. NYSTP Math 2006 Classical DIF Item Flags (cont.)

Grade	Item #	Subgroup	DIF	SMD	Mantel	Delta
8	33	Black or African-American	In Favor	0.14	n/a	n/a
8	33	Hispanic or Latino	In Favor	0.16	n/a	n/a
8	33	Asian	In Favor	0.13	n/a	n/a
8	33	High NRC	In Favor	0.13	n/a	n/a
8	34	Spanish	Against	-0.10	n/a	n/a
8	36	High NRC	Against	-0.13	n/a	n/a
8	37	Black or African-American	Against	-0.13	n/a	n/a
8	37	Hispanic or Latino	Against	-0.12	n/a	n/a
8	37	Asian	Against	-0.10	n/a	n/a
8	37	High NRC	Against	-0.11	n/a	n/a
8	38	High NRC	Against	-0.10	n/a	n/a
8	42	Black or African-American	In Favor	0.21	n/a	n/a
8	42	Hispanic or Latino	In Favor	0.13	n/a	n/a
8	42	Asian	In Favor	0.14	n/a	n/a
8	42	High NRC	In Favor	0.16	n/a	n/a
8	42	Spanish	In Favor	0.12	n/a	n/a
8	43	Black or African-American	Against	-0.11	n/a	n/a
8	43	Asian	Against	-0.11	n/a	n/a
8	44	Spanish	Against	-0.12	n/a	n/a
8	45	Black or African-American	Against	-0.13	n/a	n/a
8	45	Hispanic or Latino	Against	-0.11	n/a	n/a
8	45	Female	Against	-0.11	n/a	n/a
8	45	Spanish	Against	-0.22	n/a	n/a

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

Grade	Item	Focal Group	Direction	Magnitude
3	7	Chinese	Against	-0.120
3	8	Chinese	Against	-0.243
3	11	Chinese	In Favor	0.136
3	12	Chinese	In Favor	0.142
3	27	Chinese	Against	-0.197
3	29	Chinese	Against	-0.187
3	29	Spanish	Against	-0.160
3	30	Spanish	Against	-0.158
4	2	Chinese	In Favor	0.199
4	5	Chinese	Against	-0.137
4	8	Spanish	In Favor	0.106
4	14	Spanish	Against	-0.105
4	17	Chinese	In Favor	0.102
4	22	Chinese	Against	-0.112
4	22	Spanish	Against	-0.146
4	31	Spanish	In Favor	0.106
4	32	Chinese	Against	-0.122
4	33	Chinese	Against	-0.112
4	34	Spanish	In Favor	0.105
4	37	Spanish	In Favor	0.117
4	38	Chinese	Against	-0.248
4	38	Spanish	Against	-0.118
4	39	Chinese	Against	-0.176
4	39	Spanish	Against	-0.124
4	41	Chinese	Against	-0.262
4	41	Spanish	Against	-0.173
4	42	Spanish	Against	-0.117
4	43	Chinese	In Favor	0.164
4	45	Chinese	Against	-0.177
4	46	Chinese	Against	-0.234
4	47	Chinese	Against	-0.177
5	8	Chinese	In Favor	0.186
5	20	Chinese	Against	-0.180
5	27	Chinese	In Favor	0.167
5	27	Spanish	In Favor	0.162
5	28	Chinese	Against	-0.353
5	28	Spanish	Against	-0.167
5	29	Spanish	In Favor	0.118
5	33	Black or African-American	Against	-0.118

(Continued on next page)

Table D2. Items Flagged for DIF by the Linn-Harnisch Method (cont.)

Grade	Item	Focal Group	Direction	Magnitude
5	33	Chinese	Against	-0.263
6	4	Chinese	Against	-0.155
6	4	Spanish	Against	-0.102
6	7	Chinese	In Favor	0.109
6	26	Spanish	In Favor	0.125
6	28	Chinese	Against	-0.145
6	29	Chinese	Against	-0.234
6	31	Chinese	In Favor	0.199
6	32	Spanish	In Favor	0.113
6	34	Chinese	Against	-0.597
6	34	Spanish	Against	-0.195
6	35	Chinese	Against	-0.183
7	5	Chinese	Against	-0.134
7	5	Spanish	Against	-0.107
7	12	Chinese	Against	-0.128
7	12	Spanish	Against	-0.175
7	14	Chinese	In Favor	0.225
7	20	Spanish	In Favor	0.161
7	24	Chinese	In Favor	0.129
7	30	Spanish	Against	-0.111
7	31	Chinese	Against	-0.516
7	31	Spanish	Against	-0.105
7	32	Chinese	Against	-0.216
7	33	Chinese	In Favor	0.313
7	34	Spanish	In Favor	0.119
7	35	Chinese	In Favor	0.178
7	35	Spanish	In Favor	0.126
7	36	Spanish	In Favor	0.170
7	37	Asian	Against	-0.152
7	37	Chinese	Against	-0.211
7	37	Spanish	Against	-0.108
7	38	Hispanic or Latino	Against	-0.102
7	38	Spanish	Against	-0.187
8	5	Spanish	In Favor	0.123
8	7	Chinese	In Favor	0.142
8	8	Chinese	Against	-0.119
8	14	Chinese	Against	-0.13
8	19	Chinese	In Favor	0.116
8	19	Spanish	Against	-0.109
8	24	Chinese	In Favor	0.106

(Continued on next page)

Table D2. Items Flagged for DIF by the Linn-Harnisch Method (cont.)

Grade	Item	Focal Group	Direction	Magnitude
8	29	Chinese	Against	-0.18
8	30	Chinese	Against	-0.332
8	30	Spanish	Against	-0.178
8	33	Chinese	In Favor	0.142
8	34	Spanish	Against	-0.123
8	35	Chinese	Against	-0.186
8	36	Chinese	In Favor	0.180
8	37	Chinese	Against	-0.175
8	39	Chinese	Against	-0.101
8	41	Chinese	In Favor	0.135
8	41	Spanish	In Favor	0.107
8	42	Asian	In Favor	0.117
8	44	Spanish	Against	-0.163
8	45	Spanish	Against	-0.213

Appendices: Appendix E – Item Model Fit Statistics

These tables support the item-model fit information in Section VI (IRT Scaling). The item number, calibration model, chi-square, degrees of freedom, N-count, Z (observed) fit statistic, and Z crit (critical fit) statistic are presented for each item. Fit for most items in the Grades 3-8 Math Tests was ok ($Z_{crit} > Z$).

Table E1. Q1 Fit Statistics, Grade 3

Item number	IRT Model	Chi Square	DF	Total N	Z	Z_crit	Fit OK
1	3PL	80.61	7	177539	19.67	473.437	YES
2	3PL	125.26	7	177539	31.61	473.437	YES
3	3PL	29.62	7	177539	6.05	473.437	YES
4	3PL	53.93	7	177539	12.54	473.437	YES
5	3PL	198.86	7	177539	51.28	473.437	YES
6	3PL	124.24	7	177539	31.33	473.437	YES
7	3PL	264.90	7	177539	68.93	473.437	YES
8	3PL	145.68	7	177539	37.06	473.437	YES
9	3PL	57.89	7	177539	13.60	473.437	YES
10	3PL	600.34	7	177539	158.58	473.437	YES
11	3PL	280.13	7	177539	73.00	473.437	YES
12	3PL	1122.58	7	177539	298.15	473.437	YES
13	3PL	56.09	7	177539	13.12	473.437	YES
14	3PL	80.64	7	177539	19.68	473.437	YES
15	3PL	431.82	7	177539	113.54	473.437	YES
16	3PL	593.31	7	177539	156.70	473.437	YES
17	3PL	68.69	7	177539	16.49	473.437	YES
18	3PL	46.93	7	177539	10.67	473.437	YES
19	3PL	176.75	7	177539	45.37	473.437	YES
20	3PL	126.41	7	177539	31.91	473.437	YES
21	3PL	67.07	7	177539	16.05	473.437	YES
22	3PL	67.62	7	177539	16.20	473.437	YES
23	3PL	283.73	7	177539	73.96	473.437	YES
24	3PL	217.03	7	177539	56.13	473.437	YES
25	3PL	164.49	7	177539	42.09	473.437	YES
26	2PPC	396.96	17	177446	65.16	473.189	YES
27	2PPC	1762.14	17	177327	299.29	472.872	YES
28	2PPC	486.16	26	177436	63.81	473.163	YES
29	2PPC	542.43	26	177411	71.62	473.096	YES

(Continued on next page)

Table E1. Q1 Fit Statistics, Grade 3 (cont.)

Item number	IRT Model	Chi Square	DF	Total N	Z	Z_crit	Fit OK
30	2PPC	1120.62	17	177230	189.27	472.613	YES
31	2PPC	404.44	17	177182	66.44	472.485	YES

Table E2. Q1 Fit Statistics, Grade 4

Item number	IRT Model	Chi Square	DF	Total N	Z	Z_crit	Fit OK
1	3PL	21.61	7	186530	3.90	497.413	YES
2	3PL	45.16	7	186530	10.20	497.413	YES
3	3PL	33.65	7	186530	7.12	497.413	YES
4	3PL	54.45	7	186530	12.68	497.413	YES
5	3PL	25.83	7	186530	5.03	497.413	YES
6	3PL	241.06	7	186530	62.55	497.413	YES
7	3PL	64.96	7	186530	15.49	497.413	YES
8	3PL	16.17	7	186530	2.45	497.413	YES
9	3PL	36.41	7	186530	7.86	497.413	YES
10	3PL	49.95	7	186530	11.48	497.413	YES
11	3PL	106.94	7	186530	26.71	497.413	YES
12	3PL	136.05	7	186530	34.49	497.413	YES
13	3PL	118.60	7	186530	29.83	497.413	YES
14	3PL	158.07	7	186530	40.38	497.413	YES
15	3PL	107.65	7	186530	26.9	497.413	YES
16	3PL	105.7	7	186530	26.38	497.413	YES
17	3PL	30.45	7	186530	6.27	497.413	YES
18	3PL	24.64	7	186530	4.71	497.413	YES
19	3PL	70.08	7	186530	16.86	497.413	YES
20	3PL	120.98	7	186530	30.46	497.413	YES
21	3PL	199.16	7	186530	51.36	497.413	YES
22	3PL	13.91	7	186530	1.85	497.413	YES
23	3PL	49.80	7	186530	11.44	497.413	YES
24	3PL	93.60	7	186530	23.14	497.413	YES
25	3PL	187.10	7	186530	48.14	497.413	YES
26	3PL	132.26	7	186530	33.48	497.413	YES
27	3PL	206.89	7	186530	53.42	497.413	YES
28	3PL	31.72	7	186530	6.61	497.413	YES
29	3PL	129.49	7	186530	32.74	497.413	YES

(Continued on next page)

Table E2. Q1 Fit Statistics, Grade 4 (cont.)

Item number	IRT Model	Chi Square	DF	Total N	Z	Z_crit	Fit OK
30	3PL	172.53	7	186530	44.24	497.413	YES
31	2PPC	961.47	17	186447	161.98	497.192	YES
32	2PPC	998.59	26	186341	134.87	496.909	YES
33	2PPC	1068.86	17	186095	180.39	496.253	YES
34	2PPC	1480.45	17	186392	250.98	497.045	YES
35	2PPC	410.01	17	186311	67.40	496.829	YES
36	2PPC	531.99	17	186223	88.32	496.595	YES
37	2PPC	3768.26	17	186036	643.34	496.096	NO
38	2PPC	702.34	26	186340	93.79	496.907	YES
39	2PPC	1246.64	17	184674	210.88	492.464	YES
40	2PPC	382.36	17	186456	62.66	497.216	YES
41	2PPC	915.17	17	186409	154.03	497.091	YES
42	2PPC	1107.67	17	186370	187.05	496.987	YES
43	2PPC	1022.05	17	186229	172.36	496.611	YES
44	2PPC	542.99	17	186369	90.21	496.984	YES
45	2PPC	684.80	26	186336	91.36	496.896	YES
46	2PPC	397.48	17	186271	65.25	496.723	YES
47	2PPC	1976.70	26	186308	270.51	496.821	YES
48	2PPC	1496.90	17	185918	253.80	495.781	YES

Table E3. Q1 Fit Statistics, Grade 5

Item number	IRT Model	Chi Square	DF	Total N	Z	Z_crit	Fit OK
1	3PL	73.35	7	192893	17.73	514.381	YES
2	3PL	15.20	7	192893	2.19	514.381	YES
3	3PL	175.76	7	192893	45.10	514.381	YES
4	3PL	155.63	7	192893	39.72	514.381	YES
5	3PL	472.22	7	192893	124.33	514.381	YES
6	3PL	119.81	7	192893	30.15	514.381	YES
7	3PL	122.09	7	192893	30.76	514.381	YES
8	3PL	237.55	7	192893	61.62	514.381	YES
9	3PL	120.29	7	192893	30.28	514.381	YES
10	3PL	117.64	7	192893	29.57	514.381	YES
11	3PL	282.66	7	192893	73.67	514.381	YES
12	3PL	207.03	7	192893	53.46	514.381	YES

(Continued on next page)

Table E3. Q1 Fit Statistics, Grade 5 (cont.)

Item number	IRT Model	Chi Square	DF	Total N	Z	Z_crit	Fit OK
13	3PL	45.46	7	192893	10.28	514.381	YES
14	3PL	157.91	7	192893	40.33	514.381	YES
15	3PL	83.90	7	192893	20.55	514.381	YES
16	3PL	107.09	7	192893	26.75	514.381	YES
17	3PL	268.71	7	192893	69.95	514.381	YES
18	3PL	75.70	7	192893	18.36	514.381	YES
19	3PL	18.04	7	192893	2.95	514.381	YES
20	3PL	297.93	7	192893	77.76	514.381	YES
21	3PL	86.41	7	192893	21.22	514.381	YES
22	3PL	113.57	7	192893	28.48	514.381	YES
23	3PL	27.24	7	192893	5.41	514.381	YES
24	3PL	92.63	7	192893	22.89	514.381	YES
25	3PL	77.82	7	192893	18.93	514.381	YES
26	3PL	115.99	7	192893	29.13	514.381	YES
27	2PPC	421.63	17	192669	69.39	513.784	YES
28	2PPC	1834.44	26	192636	250.79	513.696	YES
29	2PPC	1256.90	26	192525	170.69	513.4	YES
30	2PPC	1019.30	17	192604	171.89	513.611	YES
31	2PPC	1757.53	17	192734	298.50	513.957	YES
32	2PPC	1760.16	17	192305	298.95	512.813	YES
33	2PPC	973.29	26	192424	131.37	513.131	YES
34	2PPC	3209.81	26	192533	441.52	513.421	YES

Table E4. Q1 Fit Statistics, Grade 6

Item number	IRT Model	Chi Square	DF	Total N	Z	Z_crit	Fit OK
1	3PL	101.04	7	197962	25.13	527.899	YES
2	3PL	65.77	7	197962	15.71	527.899	YES
3	3PL	88.20	7	197962	21.70	527.899	YES
4	3PL	97.48	7	197962	24.18	527.899	YES
5	3PL	409.71	7	197962	107.63	527.899	YES
6	3PL	116.47	7	197962	29.26	527.899	YES
7	3PL	83.32	7	197962	20.40	527.899	YES
8	3PL	558.92	7	197962	147.51	527.899	YES
9	3PL	41.19	7	197962	9.14	527.899	YES

(Continued on next page)

Table E4. Q1 Fit Statistics, Grade 6 (cont.)

Item number	IRT Model	Chi Square	DF	Total N	Z	Z_crit	Fit OK
10	3PL	53.01	7	197962	12.30	527.899	YES
11	3PL	53.18	7	197962	12.34	527.899	YES
12	3PL	33.06	7	197962	6.97	527.899	YES
13	3PL	76.89	7	197962	18.68	527.899	YES
14	3PL	315.62	7	197962	82.48	527.899	YES
15	3PL	798.29	7	197962	211.48	527.899	YES
16	3PL	187.30	7	197962	48.19	527.899	YES
17	3PL	111.57	7	197962	27.95	527.899	YES
18	3PL	41.89	7	197962	9.32	527.899	YES
19	3PL	1163.01	7	197962	308.96	527.899	YES
20	3PL	67.69	7	197962	16.22	527.899	YES
21	3PL	52.12	7	197962	12.06	527.899	YES
22	3PL	2145.19	7	197962	571.46	527.899	NO
23	3PL	123.06	7	197962	31.02	527.899	YES
24	3PL	186.56	7	197962	47.99	527.899	YES
25	3PL	156.81	7	197962	40.04	527.899	YES
26	2PPC	671.50	17	197716	112.25	527.243	YES
27	2PPC	620.15	17	197417	103.44	526.445	YES
28	2PPC	217.65	17	196962	34.41	525.232	YES
29	2PPC	925.84	26	197433	124.79	526.488	YES
30	2PPC	4507.52	17	197458	770.12	526.555	NO
31	2PPC	343.70	17	196067	56.03	522.845	YES
32	2PPC	726.87	17	197353	121.74	526.275	YES
33	2PPC	945.60	26	196838	127.53	524.901	YES
34	2PPC	1856.15	26	197196	253.8	525.856	YES
35	2PPC	997.64	26	196970	134.74	525.253	YES

Table E5. Q1 Fit Statistics, Grade 7

Item number	IRT Model	Chi Square	DF	Total N	Z	Z_crit	Fit OK
1	3PL	187.84	7	204959	48.33	546.557	YES
2	3PL	234.17	7	204959	60.71	546.557	YES
3	3PL	1230.94	7	204959	327.11	546.557	YES
5	3PL	60.74	7	204959	14.36	546.557	YES
6	3PL	167.88	7	204959	43.00	546.557	YES

(Continued on next page)

Table E5. Q1 Fit Statistics, Grade 7 (cont.)

Item number	IRT Model	Chi Square	DF	Total N	Z	Z_crit	Fit OK
7	3PL	495.49	7	204959	130.55	546.557	YES
8	3PL	107.39	7	204959	26.83	546.557	YES
9	3PL	667.34	7	204959	176.48	546.557	YES
10	3PL	42.48	7	204959	9.48	546.557	YES
12	3PL	187.48	7	204959	48.24	546.557	YES
13	3PL	137.00	7	204959	34.74	546.557	YES
14	3PL	164.76	7	204959	42.16	546.557	YES
16	3PL	147.17	7	204959	37.46	546.557	YES
17	3PL	102.72	7	204959	25.58	546.557	YES
18	3PL	36.31	7	204959	7.83	546.557	YES
19	3PL	555.15	7	204959	146.50	546.557	YES
20	3PL	44.65	7	204959	10.06	546.557	YES
21	3PL	518.05	7	204959	136.58	546.557	YES
22	3PL	98.18	7	204959	24.37	546.557	YES
23	3PL	23.61	7	204959	4.44	546.557	YES
24	3PL	197.52	7	204959	50.92	546.557	YES
25	3PL	81.12	7	204959	19.81	546.557	YES
26	3PL	637.99	7	204959	168.64	546.557	YES
27	3PL	309.10	7	204959	80.74	546.557	YES
28	3PL	150.08	7	204959	38.24	546.557	YES
29	3PL	1071.02	7	204959	284.37	546.557	YES
30	3PL	41.49	7	204959	9.22	546.557	YES
31	2PPC	1253.57	17	204309	212.07	544.824	YES
32	2PPC	1986.88	17	203780	337.83	543.413	YES
33	2PPC	1444.89	26	203318	196.77	542.181	YES
34	2PPC	657.43	26	203687	87.56	543.165	YES
35	2PPC	4063.63	17	202154	693.99	539.077	NO
36	2PPC	206.15	17	203841	32.44	543.576	YES
37	2PPC	1098.38	26	204163	148.71	544.435	YES
38	2PPC	771.17	26	204030	103.34	544.08	YES

Table E6. Q1 Fit Statistics, Grade 8

Item number	IRT Model	Chi Square	DF	Total N	Z	Z_crit	Fit OK
1	3PL	36.82	7	206684	7.97	551.157	YES
2	3PL	307.91	7	206684	80.42	551.157	YES
3	3PL	328.23	7	206684	85.85	551.157	YES
4	3PL	20.64	7	206684	3.65	551.157	YES
5	3PL	66.48	7	206684	15.90	551.157	YES
6	3PL	239.71	7	206684	62.19	551.157	YES
7	3PL	139.05	7	206684	35.29	551.157	YES
8	3PL	116.58	7	206684	29.29	551.157	YES
9	3PL	88.34	7	206684	21.74	551.157	YES
10	3PL	245.23	7	206684	63.67	551.157	YES
11	3PL	107.56	7	206684	26.87	551.157	YES
12	3PL	134.39	7	206684	34.05	551.157	YES
13	3PL	39.03	7	206684	8.56	551.157	YES
14	3PL	85.25	7	206684	20.91	551.157	YES
15	3PL	153.86	7	206684	39.25	551.157	YES
16	3PL	70.30	7	206684	16.92	551.157	YES
18	3PL	71.05	7	206684	17.12	551.157	YES
19	3PL	69.36	7	206684	16.67	551.157	YES
20	3PL	45.85	7	206684	10.38	551.157	YES
21	3PL	13.78	7	206684	1.81	551.157	YES
22	3PL	33.96	7	206684	7.20	551.157	YES
23	3PL	32.57	7	206684	6.83	551.157	YES
24	3PL	63.94	7	206684	15.22	551.157	YES
25	3PL	38.85	7	206684	8.51	551.157	YES
26	3PL	102.18	7	206684	25.44	551.157	YES
27	3PL	108.18	7	206684	27.04	551.157	YES
28	2PPC	497.98	17	205916	82.49	549.109	YES
29	2PPC	1784.96	26	205406	243.92	547.749	YES
30	2PPC	970.93	17	206012	163.6	549.365	YES
31	2PPC	485.78	26	200460	63.76	534.56	YES
32	2PPC	1274.75	17	203995	215.7	543.987	YES
33	2PPC	7732.41	17	204924	1323.18	546.464	NO
34	2PPC	360.83	17	204380	58.97	545.013	YES
35	2PPC	228.41	17	203244	36.26	541.984	YES
36	2PPC	1059.60	26	202598	143.33	540.261	YES
37	2PPC	996.80	17	205265	168.03	547.373	YES

(Continued on next page)

Table E6. Q1 Fit Statistics, Grade 8 (cont.)

Item number	IRT Model	Chi Square	DF	Total N	Z	Z_crit	Fit OK
38	2PPC	722.06	17	202609	120.92	540.291	YES
39	2PPC	2363.52	26	204844	324.16	546.251	YES
40	2PPC	237.57	17	201359	37.83	536.957	YES
41	2PPC	1347.93	17	201804	228.25	538.144	YES
42	2PPC	14134.56	26	203677	1956.51	543.139	NO
43	2PPC	1065.32	26	204528	144.13	545.408	YES
44	2PPC	1185.17	17	202510	200.34	540.027	YES
45	2PPC	398.81	17	202438	65.48	539.835	YES

Appendices: Appendix F – Derivation of the Generalized SPI Procedure

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

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

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

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

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 :

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

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

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

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

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 selected-response 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 model (2PPC) 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 1_i score levels, integer scores are assigned that ranged from 0 to $1_i - 1$:

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

where

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

and $\gamma_{i0} = 0$. Alpha (α_i) is the item discrimination and gamma (γ_{ih}) is related to the difficulty of the item levels: the trace lines for adjacent score levels intersect at γ_{ih}/α_i .

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

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

where 1_i is the number of score levels in item i , including 0. T_j , the expected proportion of maximum score for standard j , is

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

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

The theoretical random variation in item response vectors and resulting ($\hat{\theta}$) values for a given examinee produces the distribution $g(\hat{T}_j | \hat{\theta})$ 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 & 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 θ (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 p. 85) produces

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

and

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

Thus,

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

and the parameters of the prior beta distribution for T_j can be expressed in terms of the parameters of the three-parameter IRT and two-parameter partial credit models. Furthermore, the parameters of the posterior distribution of T_j also can be expressed in terms of the IRT parameters:

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

and

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

The OPI is

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

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

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

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

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

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

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

Check on Consistency and Adjustment of Weight Given to Prior

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

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

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

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

Possible Violations of the Assumptions

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

If the items in the standard do not permit guessing, it is reasonable to assume \hat{T}_j , the expected proportion correct of maximum score for a standard, will be greater or equal to zero. If correct guessing is possible, as it is with selected-response 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-scoring examinees. Yen (1987), working with tests containing exclusively selected-response items, found that there does not appear to be a practical importance to this underestimation. The impact of any such effect would be reduced as the proportion of constructed-response items in the test increases. The size of this effect, nonetheless, was evaluated using simulations (Yen, Sykes, Ito, and Julian 1997).

Third, 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 constructed-response 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.

Appendices: Appendix G – Derivation of Classification Consistency and Accuracy

Classification Consistency

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

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

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

In this report, the marginal distribution $P(X = x)$ is denoted as $f(x)$, and the conditional error distribution $P(X = x | \Phi = \theta)$ 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^{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 operational 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 θ , X_1 and X_2 are independent and identically distributed. Consequently, the conditional bivariate distribution of X_1 and X_2 is

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

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 Γ_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 $\theta \in \Gamma_w$.

Appendices: Appendix H – Scale Score Frequency Distributions

Tables H1 to H6 depict the scale score 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 MA 2006 Scale Score FD, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
470	106	0.05	106	0.05
503	129	0.06	235	0.12
536	250	0.12	485	0.24
556	346	0.17	831	0.41
569	494	0.24	1325	0.66
579	670	0.33	1995	0.99
587	829	0.41	2824	1.40
594	953	0.47	3777	1.87
599	1089	0.54	4866	2.41
604	1283	0.64	6149	3.05
609	1369	0.68	7518	3.72
613	1630	0.81	9148	4.53
617	1742	0.86	10890	5.39
621	1930	0.96	12820	6.35
624	2156	1.07	14976	7.42
628	2481	1.23	17457	8.65
631	2681	1.33	20138	9.97
634	3101	1.54	23239	11.51
637	3413	1.69	26652	13.20
640	3820	1.89	30472	15.09
644	4157	2.06	34629	17.15
647	4695	2.33	39324	19.48
650	5436	2.69	44760	22.17
653	6020	2.98	50780	25.15
657	6860	3.40	57640	28.55
660	7547	3.74	65187	32.29
664	8586	4.25	73773	36.54

(Continued on next page)

Table H1. Grade 3 MA 2006 Scale Score FD, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
668	9602	4.76	83375	41.29
672	10828	5.36	94203	46.66
676	12129	6.01	106332	52.66
682	13786	6.83	120118	59.49
688	14923	7.39	135041	66.88
695	16172	8.01	151213	74.89
704	16568	8.21	167781	83.10
717	15601	7.73	183382	90.82
740	12256	6.07	195638	96.89
770	6270	3.11	201908	100.00

Table H2. Grade 4 MA 2006 Scale Score FD, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
485	618	0.30	618	0.30
528	429	0.21	1047	0.52
547	434	0.21	1481	0.73
559	510	0.25	1991	0.98
568	532	0.26	2523	1.24
575	594	0.29	3117	1.54
581	619	0.31	3736	1.84
586	664	0.33	4400	2.17
590	698	0.34	5098	2.52
594	758	0.37	5856	2.89
598	755	0.37	6611	3.26
601	877	0.43	7488	3.69
604	869	0.43	8357	4.12
607	896	0.44	9253	4.56
610	1053	0.52	10306	5.08
613	1103	0.54	11409	5.63
615	1121	0.55	12530	6.18
618	1172	0.58	13702	6.76
620	1311	0.65	15013	7.41

(Continued on next page)

Table H2. Grade 4 MA 2006 Scale Score FD, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
622	1333	0.66	16346	8.06
624	1424	0.70	17770	8.77
626	1491	0.74	19261	9.50
628	1627	0.80	20888	10.31
630	1690	0.83	22578	11.14
632	1758	0.87	24336	12.01
634	1842	0.91	26178	12.91
636	1994	0.98	28172	13.90
638	2067	1.02	30239	14.92
640	2138	1.05	32377	15.97
641	2232	1.10	34609	17.07
643	2339	1.15	36948	18.23
645	2414	1.19	39362	19.42
647	2477	1.22	41839	20.64
649	2753	1.36	44592	22.00
650	2837	1.40	47429	23.40
652	3035	1.50	50464	24.90
654	3107	1.53	53571	26.43
656	3086	1.52	56657	27.95
658	3398	1.68	60055	29.63
659	3455	1.70	63510	31.33
661	3721	1.84	67231	33.17
663	3923	1.94	71154	35.10
665	4154	2.05	75308	37.15
667	4314	2.13	79622	39.28
669	4536	2.24	84158	41.52
671	4825	2.38	88983	43.90
673	5054	2.49	94037	46.39
676	5111	2.52	99148	48.91
678	5628	2.78	104776	51.69
680	5736	2.83	110512	54.52
683	6007	2.96	116519	57.48
685	6231	3.07	122750	60.56

(Continued on next page)

Table H2. Grade 4 MA 2006 Scale Score FD, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
688	6553	3.23	129303	63.79
691	6842	3.38	136145	67.17
695	6888	3.40	143033	70.57
698	7202	3.55	150235	74.12
702	7299	3.60	157534	77.72
707	7386	3.64	164920	81.36
712	7531	3.72	172451	85.08
718	7143	3.52	179594	88.60
725	6841	3.38	186435	91.98
734	6079	3.00	192514	94.98
747	5050	2.49	197564	97.47
769	3545	1.75	201109	99.22
800	1586	0.78	202695	100.00

Table H3. Grade 5 MA 2006 Scale Score FD, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
495	258	0.12	258	0.12
523	313	0.15	571	0.27
546	532	0.25	1103	0.53
560	868	0.41	1971	0.94
570	1067	0.51	3038	1.45
578	1386	0.66	4424	2.11
586	1655	0.79	6079	2.91
592	1890	0.90	7969	3.81
597	2138	1.02	10107	4.83
603	2431	1.16	12538	5.99
607	2748	1.31	15286	7.31
611	3031	1.45	18317	8.76
615	3211	1.53	21528	10.29
619	3504	1.67	25032	11.97
623	3632	1.74	28664	13.70

(Continued on next page)

Table H3. Grade 5 MA 2006 Scale Score FD, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
626	3807	1.82	32471	15.52
629	4145	1.98	36616	17.50
633	4269	2.04	40885	19.54
636	4574	2.19	45459	21.73
639	4896	2.34	50355	24.07
642	5022	2.40	55377	26.47
644	5165	2.47	60542	28.94
647	5426	2.59	65968	31.53
650	5757	2.75	71725	34.29
653	6023	2.88	77748	37.16
656	6216	2.97	83964	40.14
659	6536	3.12	90500	43.26
661	6749	3.23	97249	46.49
664	7090	3.39	104339	49.88
667	7375	3.53	111714	53.40
671	7462	3.57	119176	56.97
674	7911	3.78	127087	60.75
677	8168	3.90	135255	64.65
681	8239	3.94	143494	68.59
685	8473	4.05	151967	72.64
689	8538	4.08	160505	76.72
694	8619	4.12	169124	80.84
700	8687	4.15	177811	85.00
706	8185	3.91	185996	88.91
715	7909	3.78	193905	92.69
728	6967	3.33	200872	96.02
750	5388	2.58	206260	98.59
780	2940	1.41	209200	100.00

Table H4. Grade 6 MA 2006 Scale Score FD, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
500	1747	0.83	1747	0.83
543	1381	0.65	3128	1.48
563	1767	0.84	4895	2.32
575	2131	1.01	7026	3.32
583	2457	1.16	9483	4.49
590	2725	1.29	12208	5.78
596	2888	1.37	15096	7.14
601	2993	1.42	18089	8.56
605	3247	1.54	21336	10.09
609	3355	1.59	24691	11.68
613	3459	1.64	28150	13.32
616	3590	1.70	31740	15.02
620	3845	1.82	35585	16.83
623	3947	1.87	39532	18.70
626	4078	1.93	43610	20.63
629	4408	2.09	48018	22.72
632	4574	2.16	52592	24.88
634	4769	2.26	57361	27.14
637	4903	2.32	62264	29.46
640	5113	2.42	67377	31.88
642	5402	2.56	72779	34.43
645	5354	2.53	78133	36.96
647	5462	2.58	83595	39.55
650	5754	2.72	89349	42.27
652	5872	2.78	95221	45.05
655	6102	2.89	101323	47.93
657	6247	2.96	107570	50.89
660	6231	2.95	113801	53.84
663	6281	2.97	120082	56.81
665	6441	3.05	126523	59.86
668	6437	3.05	132960	62.90
671	6342	3.00	139302	65.90
674	6438	3.05	145740	68.95
676	6398	3.03	152138	71.98
679	6414	3.03	158552	75.01

(Continued on next page)

Table H4. Grade 6 MA 2006 Scale Score FD, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
683	6357	3.01	164909	78.02
686	6250	2.96	171159	80.97
690	6200	2.93	177359	83.91
694	6132	2.90	183491	86.81
698	5969	2.82	189460	89.63
704	5811	2.75	195271	92.38
710	5327	2.52	200598	94.90
720	4679	2.21	205277	97.11
737	3725	1.76	209002	98.88
780	2374	1.12	211376	100.00

Table H5. Grade 7 MA 2006 Scale Score FD, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
500	1451	0.67	1451	0.67
522	1014	0.47	2465	1.13
542	1335	0.61	3800	1.75
556	1696	0.78	5496	2.53
567	2123	0.98	7619	3.51
576	2420	1.11	10039	4.62
584	2894	1.33	12933	5.95
590	3282	1.51	16215	7.46
597	3730	1.72	19945	9.18
602	4134	1.90	24079	11.08
607	4570	2.10	28649	13.19
612	4985	2.29	33634	15.48
616	5260	2.42	38894	17.90
620	5524	2.54	44418	20.45
624	5863	2.70	50281	23.15
628	5985	2.76	56266	25.90
631	6339	2.92	62605	28.82
635	6459	2.97	69064	31.79

(Continued on next page)

Table H5. Grade 7 MA 2006 Scale Score FD, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
638	6513	3.00	75577	34.79
641	6594	3.04	82171	37.83
644	6995	3.22	89166	41.05
647	7079	3.26	96245	44.31
650	7316	3.37	103561	47.67
653	7209	3.32	110770	50.99
656	7270	3.35	118040	54.34
659	7306	3.36	125346	57.70
662	7328	3.37	132674	61.08
665	7415	3.41	140089	64.49
668	7249	3.34	147338	67.83
671	7324	3.37	154662	71.20
675	7422	3.42	162084	74.62
678	7568	3.48	169652	78.10
682	7312	3.37	176964	81.47
686	6968	3.21	183932	84.67
691	6856	3.16	190788	87.83
696	6451	2.97	197239	90.80
702	5990	2.76	203229	93.56
710	4997	2.30	208226	95.86
719	3963	1.82	212189	97.68
733	2794	1.29	214983	98.97
756	1604	0.74	216587	99.71
800	638	0.29	217225	100.00

Table H6. Grade 8 MA 2006 Scale Score FD, State

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
480	1764	0.80	1764	0.80
517	1245	0.57	3009	1.37
552	1567	0.71	4576	2.09
567	1901	0.87	6477	2.95
578	2266	1.03	8743	3.99

(Continued on next page)

Table H6. Grade 8 MA 2006 Scale Score FD, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
585	2444	1.11	11187	5.10
591	2649	1.21	13836	6.31
596	2826	1.29	16662	7.60
601	3136	1.43	19798	9.03
605	3088	1.41	22886	10.44
608	3313	1.51	26199	11.95
611	3327	1.52	29526	13.46
614	3333	1.52	32859	14.98
617	3588	1.64	36447	16.62
619	3615	1.65	40062	18.27
621	3488	1.59	43550	19.86
624	3561	1.62	47111	21.48
626	3560	1.62	50671	23.11
628	3647	1.66	54318	24.77
630	3686	1.68	58004	26.45
631	3604	1.64	61608	28.09
633	3641	1.66	65249	29.75
635	3639	1.66	68888	31.41
636	3540	1.61	72428	33.03
638	3716	1.69	76144	34.72
640	3593	1.64	79737	36.36
641	3589	1.64	83326	38.00
643	3464	1.58	86790	39.58
644	3605	1.64	90395	41.22
646	3523	1.61	93918	42.83
647	3540	1.61	97458	44.44
648	3577	1.63	101035	46.07
650	3541	1.61	104576	47.69
651	3848	1.75	108424	49.44
653	3742	1.71	112166	51.15
654	3705	1.69	115871	52.84
655	3619	1.65	119490	54.49
657	3620	1.65	123110	56.14
658	3654	1.67	126764	57.81

(Continued on next page)

Table H6. Grade 8 MA 2006 Scale Score FD, State (cont.)

SS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
660	3562	1.62	130326	59.43
661	3678	1.68	134004	61.11
663	3628	1.65	137632	62.76
664	3622	1.65	141254	64.41
666	3669	1.67	144923	66.09
668	3716	1.69	148639	67.78
669	3782	1.72	152421	69.51
671	3680	1.68	156101	71.18
673	3696	1.69	159797	72.87
675	3838	1.75	163635	74.62
677	3860	1.76	167495	76.38
679	3787	1.73	171282	78.11
681	4036	1.84	175318	79.95
684	4074	1.86	179392	81.80
686	4381	2.00	183773	83.80
689	4354	1.99	188127	85.79
693	4297	1.96	192424	87.75
697	4535	2.07	196959	89.82
701	4556	2.08	201515	91.89
707	4597	2.10	206112	93.99
715	4406	2.01	210518	96.00
725	3892	1.77	214410	97.77
744	3230	1.47	217640	99.25
775	1654	0.75	219294	100.00