

New York State Testing Program 2006: English Language Arts, 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> | 3 |
| <i>Proficiency Level Cut Scores 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..... | 5 |
| TEST DESCRIPTION | 5 |
| TEST CONFIGURATION..... | 5 |
| TEST BLUEPRINT | 6 |
| 2006 ITEM MAPPING BY NEW YORK STATE STANDARDS AND STRANDS..... | 14 |
| CONTENT RATIONALE | 14 |
| ITEM DEVELOPMENT | 15 |
| ITEM REVIEW | 16 |
| MATERIALS DEVELOPMENT | 16 |
| ITEM SELECTION AND TEST CREATION (CRITERIA AND PROCESS) | 17 |
| PROFICIENCY AND PERFORMANCE STANDARDS | 18 |
| SECTION III: VALIDITY | 19 |
| CONTENT VALIDITY | 19 |
| CONSTRUCT (INTERNAL STRUCTURE) VALIDITY | 20 |
| <i>Internal consistency</i> | 20 |
| <i>Unidimensionality</i> | 20 |
| <i>Minimization of Bias</i> | 22 |
| CONSEQUENTIAL VALIDITY..... | 23 |
| SECTION IV: TEST ADMINISTRATION AND SCORING | 25 |
| TEST ADMINISTRATION | 25 |
| SCORING PROCEDURES OF OPERATIONAL TESTS..... | 25 |
| SCORING MODELS | 25 |
| SCORING OF CONSTRUCTED RESPONSE ITEMS..... | 26 |
| SCORER QUALIFICATIONS AND TRAINING | 27 |
| QUALITY CONTROL PROCESS | 27 |
| SECTION V: OPERATIONAL TEST DATA COLLECTION AND CLASSICAL ANALYSIS..... | 28 |
| DATA COLLECTION | 28 |
| DATA PROCESSING | 28 |
| CLASSICAL ANALYSIS AND CALIBRATION SAMPLE CHARACTERISTICS..... | 30 |
| CLASSICAL DATA ANALYSIS | 34 |
| <i>Item Difficulty and Response Distribution</i> | 34 |
| <i>Point-Biserial Correlation Coefficients</i> | 40 |
| <i>Distracter Analysis</i> | 41 |
| <i>Test Statistics and Reliability Coefficients</i> | 41 |
| <i>Speededness</i> | 42 |

| | |
|--|------------|
| <i>Differential Item Functioning</i> | 42 |
| SECTION VI: IRT SCALING | 45 |
| IRT MODELS AND RATIONALE FOR USE | 45 |
| CALIBRATION SAMPLE | 46 |
| CALIBRATION PROCESS | 46 |
| ITEM-MODEL FIT | 47 |
| LOCAL INDEPENDENCE | 48 |
| SCALING | 49 |
| <i>Initial Scaling</i> | 49 |
| <i>Final Scaling</i> | 50 |
| ITEM PARAMETERS | 52 |
| TEST CHARACTERISTIC CURVES | 58 |
| EQUATING | 56 |
| SCORING PROCEDURE | 56 |
| <i>Weighting Constructed Response Items in Grades 4 and 8</i> | 57 |
| RAW SCORE TO SCALE SCORE AND SEM CONVERSION TABLES | 60 |
| STANDARD PERFORMANCE INDEX | 67 |
| IRT DIF STATISTICS | 69 |
| SECTION VII: STANDARD SETTING | 72 |
| DESCRIPTION OF STANDARD SETTING PROCESS | 72 |
| DESCRIPTION OF THE BOOKMARK METHOD | 75 |
| DESCRIPTION OF JUDGE/EXPERT PANELS | 76 |
| VERTICALLY MODERATED STANDARDS | 76 |
| DEFINITION OF PERFORMANCE LEVELS | 77 |
| FINAL CUT SCORES | 77 |
| SECTION VIII: RELIABILITY AND STANDARD ERROR OF MEASUREMENT | 78 |
| TEST RELIABILITY | 78 |
| <i>Reliability for Total Test</i> | 78 |
| <i>Reliability of MC items</i> | 79 |
| <i>Reliability of CR items</i> | 79 |
| <i>Test Reliability for NCLB reporting categories</i> | 79 |
| STANDARD ERROR OF MEASUREMENT | 85 |
| PERFORMANCE LEVEL CLASSIFICATION CONSISTENCY AND ACCURACY | 86 |
| <i>Consistency</i> | 87 |
| <i>Accuracy</i> | 87 |
| SECTION IX: SUMMARY OF OPERATIONAL TEST RESULTS | 89 |
| SCALE SCORE STATISTICS | 89 |
| <i>Grade 3</i> | 89 |
| <i>Grade 4</i> | 90 |
| <i>Grade 5</i> | 91 |
| <i>Grade 6</i> | 92 |
| <i>Grade 7</i> | 94 |
| <i>Grade 8</i> | 95 |
| PERFORMANCE LEVEL DISTRIBUTIONS | 96 |
| <i>Grade 3</i> | 97 |
| <i>Grade 4</i> | 98 |
| <i>Grade 5</i> | 99 |
| <i>Grade 6</i> | 100 |
| <i>Grade 7</i> | 102 |
| <i>Grade 8</i> | 103 |
| SECTION X: SPECIAL STUDIES | 105 |

| | |
|---|------------|
| LINKING ELA GRADES 4 AND 8 2006 TO 2005 ASSESSMENTS..... | 105 |
| SECTION XI: REFERENCES..... | 126 |
| APPENDICES: APPENDIX A – ELA PASSAGE SPECIFICATIONS..... | 129 |
| APPENDICES: APPENDIX B – CRITERIA FOR ITEM ACCEPTABILITY | 131 |
| APPENDICES: APPENDIX C – PSYCHOMETRIC GUIDELINES FOR OPERATIONAL ITEM SELECTION..... | 133 |
| APPENDICES: APPENDIX D – FACTOR ANALYSIS RESULTS..... | 134 |
| APPENDICES: APPENDIX E – ITEMS FLAGGED FOR DIF | 150 |
| APPENDICES: APPENDIX F – ITEM MODEL FIT STATISTICS..... | 152 |
| APPENDICES: APPENDIX G – DERIVATION OF THE GENERALIZED SPI PROCEDURE . | 158 |
| APPENDICES: APPENDIX H – DERIVATION OF CLASSIFICATION CONSISTENCY AND ACCURACY..... | 164 |
| APPENDICES: APPENDIX I – SCALE SCORE FREQUENCY DISTRIBUTIONS..... | 166 |

List of Tables

| | |
|--|----|
| TABLE 1. NYSTP ELA 2006 TEST CONFIGURATION..... | 5 |
| TABLE 2. NYSTP ELA 2006 CLUSTER ITEMS..... | 6 |
| TABLE 3. NYSTP ELA 2006 TEST BLUEPRINT | 7 |
| TABLE 4A. NYSTP ELA 2006 OPERATIONAL TEST MAP, GRADE 3 | 8 |
| TABLE 4B. NYSTP ELA 2006 OPERATIONAL TEST MAP, GRADE 4..... | 9 |
| TABLE 4C. NYSTP ELA 2006 OPERATIONAL TEST MAP, GRADE 5 | 10 |
| TABLE 4D. NYSTP ELA 2006 OPERATIONAL TEST MAP, GRADE 6 | 11 |
| TABLE 4E. NYSTP ELA 2006 OPERATIONAL TEST MAP, GRADE 7..... | 12 |
| TABLE 4F. NYSTP ELA 2006 OPERATIONAL TEST MAP, GRADE 8..... | 13 |
| TABLE 5. NYSTP ELA 2006 STANDARD COVERAGE | 14 |
| TABLE 6. FACTOR ANALYSIS RESULTS FOR ELA TESTS (TOTAL POPULATION)..... | 21 |
| TABLE 7A. NYSTP ELA DATA CLEANING..... | 29 |
| TABLE 7B. NYSTP ELA DATA CLEANING | 29 |
| TABLE 7C. NYSTP ELA DATA CLEANING..... | 29 |
| TABLE 7D. NYSTP ELA DATA CLEANING..... | 29 |
| TABLE 7E. NYSTP ELA DATA CLEANING | 30 |
| TABLE 7F. NYSTP ELA DATA CLEANING | 30 |
| TABLE 8A. GRADE 3 SAMPLE CHARACTERISTICS (N=179,552) | 31 |
| TABLE 8B. GRADE 4 SAMPLE CHARACTERISTICS (N=183,342)..... | 31 |
| TABLE 8C. GRADE 5 SAMPLE CHARACTERISTICS (N=191,161) | 32 |
| TABLE 8D. GRADE 6 SAMPLE CHARACTERISTICS (N=193,354) | 32 |
| TABLE 8E. GRADE 7 SAMPLE CHARACTERISTICS (N=200,208)..... | 33 |
| TABLE 8F. GRADE 8 SAMPLE CHARACTERISTICS (N=202,424)..... | 33 |
| TABLE 9A. P-VALUES, SCORED RESPONSE DISTRIBUTIONS, AND POINT BISERIALS, GRADE 3 | 35 |
| TABLE 9B. P-VALUES, SCORED RESPONSE DISTRIBUTIONS, AND POINT BISERIALS, GRADE 4 | 36 |
| TABLE 9C. P-VALUES, SCORED RESPONSE DISTRIBUTIONS, AND POINT BISERIALS, GRADE 5 | 37 |
| TABLE 9D. P-VALUES, SCORED RESPONSE DISTRIBUTIONS, AND POINT BISERIALS, GRADE 6 | 38 |
| TABLE 9E. P-VALUES, SCORED RESPONSE DISTRIBUTIONS, AND POINT BISERIALS, GRADE 7 | 39 |
| TABLE 9F. P-VALUES, SCORED RESPONSE DISTRIBUTIONS, AND POINT BISERIALS, GRADE 8 | 40 |
| TABLE 10. NYSTP ELA 2006 TEST FORM STATISTICS AND RELIABILITY..... | 42 |
| TABLE 11. NYSTP ELA 2006 CLASSICAL DIF SAMPLE N-COUNTS..... | 43 |

| | |
|--|-----------|
| TABLE 12. NUMBER OF ITEMS FLAGGED BY SMD AND MANTEL-HAENSZEL DIF METHODS..... | 44 |
| TABLE 13. NYSTP ELA 2006 CALIBRATION RESULTS..... | 47 |
| TABLE 14. NYSTP ELA 2006 INITIAL TRANSFORMATION CONSTANTS..... | 49 |
| TABLE 15. NYSTP ELA 2006 FINAL TRANSFORMATION CONSTANTS..... | 52 |
| TABLE 16A. GRADE 3 2006 OPERATIONAL ITEM PARAMETER ESTIMATES..... | 53 |
| TABLE 16B. GRADE 4 2006 OPERATIONAL ITEM PARAMETER ESTIMATES..... | 54 |
| TABLE 16C. GRADE 5 2006 OPERATIONAL ITEM PARAMETER ESTIMATES..... | 55 |
| TABLE 16D. GRADE 6 2006 OPERATIONAL ITEM PARAMETER ESTIMATES..... | 56 |
| TABLE 16E. GRADE 7 2006 OPERATIONAL ITEM PARAMETER ESTIMATES..... | 57 |
| TABLE 16F. GRADE 8 2006 OPERATIONAL ITEM PARAMETER ESTIMATES..... | 58 |
| TABLE 17. ELA GRADE 4 MC AND CR POINT DISTRIBUTION IN 2005 BY LEARNING STANDARDS..... | 58 |
| TABLE 18. ELA GRADE 4 MC AND CR POINT DISTRIBUTION IN 2006 BY LEARNING STANDARDS..... | 58 |
| TABLE 19. ELA GRADE 8 MC AND CR POINT DISTRIBUTION IN 2005 BY LEARNING STANDARDS..... | 58 |
| TABLE 20. ELA GRADE 8 MC AND CR POINT DISTRIBUTION IN 2006 BY LEARNING STANDARDS..... | 58 |
| TABLE 21. NEW YORK STATE ELA GRADES 4 AND 8 MC AND CR PROPORTIONS IN 2005 AND 2006..... | 59 |
| TABLE 22. NYSTP ELA 2006 MINIMUM AND MAXIMUM SCALE SCORES..... | 60 |
| TABLE 23A. GRADE 3 RAW SCORE TO SCALE SCORE (WITH STANDARD ERROR)..... | 61 |
| TABLE 23B. GRADE 4 RAW SCORE TO SCALE SCORE (WITH STANDARD ERROR)..... | 62 |
| TABLE 23C. GRADE 5 RAW SCORE TO SCALE SCORE (WITH STANDARD ERROR)..... | 63 |
| TABLE 23D. GRADE 6 RAW SCORE TO SCALE SCORE (WITH STANDARD ERROR)..... | 64 |
| TABLE 23E. GRADE 7 RAW SCORE TO SCALE SCORE (WITH STANDARD ERROR)..... | 65 |
| TABLE 23F. GRADE 8 RAW SCORE TO SCALE SCORE (WITH STANDARD ERROR)..... | 66 |
| TABLE 24. SPI TARGET RANGES..... | 68 |
| TABLE 25. NUMBER OF ITEMS FLAGGED FOR DIF BY THE LINN-HARNISCH METHOD..... | 71 |
| TABLE 26. MEASUREMENT REVIEW FORUM BASED RECOMMENDED IMPACT DATA..... | 72 |
| TABLE 27. PARTICIPANTS BASED CUT SCORES AND ASSOCIATED IMPACT DATA..... | 73 |
| TABLE 28. VERTICAL ARTICULATION PANEL BASED CUT SCORES AND ASSOCIATED IMPACT DATA (TABLE LEADER SMOOTHING)..... | 74 |
| TABLE 29. MEASUREMENT REVIEW FORUM (GROUP 1) BASED CUT SCORES AND ASSOCIATED IMPACT DATA..... | 74 |
| TABLE 30. MEASUREMENT REVIEW FORUM (GROUP 2) BASED CUT SCORES AND ASSOCIATED IMPACT DATA..... | 75 |
| TABLE 31. FINAL NYSED APPROVED CUT SCORES AND IMPACT DATA..... | 75 |
| TABLE 32. FINAL CUT SCORES NYSTP ELA..... | 77 |

| | |
|---|------------|
| TABLE 33. ELA 3-8 TESTS RELIABILITY AND STANDARD ERROR OF MEASUREMENT ... | 78 |
| TABLE 34. RELIABILITY AND STANDARD ERROR OF MEASUREMENT – MC ITEMS ONLY | 79 |
| TABLE 35. RELIABILITY AND STANDARD ERROR OF MEASUREMENT - CR ITEMS ONLY | 79 |
| TABLE 36A. GRADE 3 TEST RELIABILITY BY SUBGROUP | 80 |
| TABLE 36B. GRADE 4 TEST RELIABILITY BY SUBGROUP | 81 |
| TABLE 36C. GRADE 5 TEST RELIABILITY BY SUBGROUP | 82 |
| TABLE 36D. GRADE 6 TEST RELIABILITY BY SUBGROUP | 83 |
| TABLE 36E. GRADE 7 TEST RELIABILITY BY SUBGROUP | 84 |
| TABLE 36F. GRADE 8 TEST RELIABILITY BY SUBGROUP | 85 |
| TABLE 37. DECISION CONSISTENCY (ALL CUTS) | 87 |
| TABLE 38. DECISION CONSISTENCY (LEVEL III CUT) | 87 |
| TABLE 39. DECISION AGREEMENT (ACCURACY) | 88 |
| TABLE 40. ELA GRADES 3-8 SCALE SCORE DISTRIBUTION SUMMARY | 89 |
| TABLE 41. SCALE SCORE DISTRIBUTION SUMMARY, BY SUBGROUP, GRADE 3 | 90 |
| TABLE 42. SCALE SCORE DISTRIBUTION SUMMARY, BY SUBGROUP, GRADE 4 | 91 |
| TABLE 43. SCALE SCORE DISTRIBUTION SUMMARY, BY SUBGROUP, GRADE 5 | 92 |
| TABLE 44. SCALE SCORE DISTRIBUTION SUMMARY, BY SUBGROUP, GRADE 6 | 93 |
| TABLE 45. SCALE SCORE DISTRIBUTION SUMMARY, BY SUBGROUP, GRADE 7 | 94 |
| TABLE 46. SCALE SCORE DISTRIBUTION SUMMARY, BY SUBGROUP, GRADE 8 | 95 |
| TABLE 48. PERFORMANCE LEVEL DISTRIBUTIONS, BY SUBGROUP, GRADE 3 | 98 |
| TABLE 49. PERFORMANCE LEVEL DISTRIBUTIONS, BY SUBGROUP, GRADE 4 | 99 |
| TABLE 50. PERFORMANCE LEVEL DISTRIBUTIONS, BY SUBGROUP, GRADE 5 | 100 |
| TABLE 51. PERFORMANCE LEVEL DISTRIBUTIONS, BY SUBGROUP, GRADE 6 | 101 |
| TABLE 52. PERFORMANCE LEVEL DISTRIBUTIONS, BY SUBGROUP, GRADE 7 | 102 |
| TABLE 53. PERFORMANCE LEVEL DISTRIBUTIONS, BY SUBGROUP, GRADE 8 | 103 |
| TABLE 54. GRADE 4 ELA LINKING OF 2006 TO 2005 SCALE SCORES | 107 |
| TABLE 55. GRADE 8 ELA LINKING OF 2006 TO 2005 SCALE SCORES | 116 |
| TABLE 56. ELA GRADE 4 EQUIPERCENTILE LINKING SUMMARY | 125 |
| TABLE 57. ELA GRADE 8 EQUIPERCENTILE LINKING SUMMARY | 125 |

Section I: Introduction and Overview

Introduction

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

Test Purpose

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

Target Population

Students in New York State public school grades 3, 4, 5, 6, 7, and 8 (and ungraded students of equivalent age) are the target population for the Grades 3-8 testing program. Nonpublic schools may participate in the testing program but the participation is not mandatory for them. In 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, the New York State Education Department (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://emsc33.nysed.gov/3-8/sam/ela06p.pdf>).

Test Use and Decisions Based on Assessment

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

Scale Scores

The scale score is a quantification of the ability measured by the Grades 3-8 ELA Tests at each grade level. The scale scores are comparable within each grade level but not across grades because the Grades 3-8 ELA Tests are not on a vertical scale. The test scores are reported at the individual level and can also be aggregated. Detailed information on derivation and properties of scale scores is provided in Section VI (IRT Scaling) of this report. Uses of Grades 3-8 ELA Tests scores include: determining student progress within schools and districts, supporting registration of schools and districts, determining eligibility of students for additional instruction time, and providing teachers with indicators of a student's need, or lack of need, for remediation in specific subject area knowledge.

Proficiency Level Cut Scores 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 ELA proficiency cut scores reflect the abilities intended by the New York State Education Department. Performance of students on the Grades 3-8 ELA Tests in relation to proficiency level cut scores is reported in a form of Performance Level classification. 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 English Language Arts* and the *New York State ELA 2006 Measurement Review Technical Report 2006 for English Language Arts*.

Standard Performance Index Scores

Standard Performance Index (SPI) scores are obtained from the Grades 3-8 ELA Tests. The SPI score is an indicator of student ability, knowledge and skills in specific learning standards and 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 assessments of student proficiency in English Language Arts, the Grades 3-8 ELA tests are not translated into any other 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 (February 2005)
 - c. Ranging and scoring of field test data (March-April 2005)
 - d. Analyze data from field test (April-May 2005)
3. Operational test construction (June 2005)
4. Test administration (January 2006)
5. Scoring and data retrieval (January-April 2006)
6. Data analysis (April-July 2006)
7. Standard setting (June 2006)
8. Score reporting (September 2006)

Section II: Test Design and Development

Test Description

The Grades 3-8 ELA Tests are New York State standards based criterion-referenced exams composed of multiple-choice (MC) and constructed-response (CR) items. The tests were administered in New York classrooms during January 2006 over a two day (Grades 3, 5, 7, and 8) or three day (Grades 4 and 6) period. The tests were printed in black and white and incorporated the concepts of universal design. Copies of the operational tests are available online (<http://www.nysedregents.org/testing/elaei/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. All students are administered a Reading section (Book 1 and Book 3), a Listening section (Book 2), and grades 3, 5, and 7 also complete an Editing Paragraph (also in Book 2). The *Teacher's Directions* (<http://www.nysedregents.org/testing/elaei/06exams/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 ELA 2006 Test Configuration

| Grade | Day | Book | Number of Items | | | Allotted Time (minutes) | |
|-------|--------|------|-----------------|-----|--------|--------------------------|------|
| | | | MC | CR* | Total* | Testing | Prep |
| 3 | 1 | 1 | 20 | 1 | 21 | 40 | 10 |
| | 2 | 2 | 4 | 3 | 7 | 35 | 15 |
| | Totals | | 24 | 4 | 28 | 75 | 25 |
| 4 | 1 | 1 | 28 | 0 | 28 | 45 | 10 |
| | 2 | 2 | 0 | 3 | 3 | 45 | 15 |
| | 3 | 3 | 0 | 4 | 4 | 60 | 10 |
| | Totals | | 28 | 7 | 35 | 150 | 35 |
| 5 | 1 | 1 | 20 | 1 | 21 | 40 | 10 |
| | 2 | 2 | 4 | 2 | 6 | 30 | 15 |
| | Totals | | 24 | 3 | 27 | 70 | 25 |
| 6 | 1 | 1 | 26 | 0 | 26 | 55 | 10 |
| | 2 | 2 | 0 | 4 | 4 | 45 | 15 |
| | 3 | 3 | 0 | 4 | 4 | 60 | 10 |
| | Totals | | 26 | 8 | 34 | 160 | 35 |

(Continued on next page)

Table 1. NYSTP ELA 2006 Test Configuration (cont.)

| Grade | Day | Book | Number of Items | | | Allotted Time (minutes) | |
|-------|--------|------|-----------------|-----|--------|--------------------------|------|
| | | | MC | CR* | Total* | Testing | Prep |
| 7 | 1 | 1 | 26 | 2 | 28 | 50 | 10 |
| | 2 | 2 | 4 | 3 | 7 | 30 | 15 |
| | Totals | | 30 | 5 | 35 | 80 | 25 |
| 8 | 1 | 1 | 26 | 0 | 26 | 45 | 10 |
| | 1 | 2 | 0 | 4 | 4 | 45 | 15 |
| | 2 | 3 | 0 | 4 | 4 | 60 | 10 |
| | Totals | | 26 | 8 | 34 | 150 | 35 |

*Reflects actual items in the test books; does not reflect cluster-scoring

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

Table 2. NYSTP ELA 2006 Cluster Items

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

Test Blueprint

The NYSTP Grades 3-8 ELA Tests assess student performance on three learning standards (S1 – Information and Understanding, S2 – Literary Response and Expression, and S3 – Critical Analysis and Evaluation). The test items assess a variety of reading, writing, and listening performance indicators in each of the three learning standards. Standard 1 is assessed primarily through test items associated with informational passages; Standard 2 is assessed primarily through test items associated with literary passages; Standard 3 is assessed through test items associated with a combination of genres. In addition, students are also tested in writing mechanics, which is assessed independent of alignment to the learning standards since writing mechanics is associated with all three learning standards. The distribution of score points across the learning standards was determined during blueprint specifications meetings held with panels of

New York State educators at the start of the testing program, prior to item development. The distribution in each grade reflects the number of assessable performance indicators in each standard at that grade and the emphasis placed on those performance indicators by the blueprint specifications panel members. Table 3 shows the Grades 3-8 ELA Tests blueprint and actual number of score points in 2006 operational tests.

Table 3. NYSTP ELA 2006 Test Blueprint

| Grade | Total Points | Writing Mechanics Points | Standard | Target Reading and Listening # Points | Selected Reading and Listening # Points | Target % of test (excluding Writing) | Selected % of test (excluding Writing) |
|-------|--------------|--------------------------|----------|---------------------------------------|---|--------------------------------------|--|
| 3 | 33 | 3 | S1 | 10 | 9 | 33% | 30% |
| | | | S2 | 14 | 15 | 47% | 50% |
| | | | S3 | 6 | 6 | 20% | 20% |
| 4 | 39 | 3 | S1 | 13 | 13 | 36% | 36% |
| | | | S2 | 16 | 15 | 44.5% | 42% |
| | | | S3 | 7 | 8 | 19.5% | 22% |
| 5 | 31 | 3 | S1 | 12 | 13 | 43% | 46% |
| | | | S2 | 10 | 9 | 36% | 32% |
| | | | S3 | 6 | 6 | 21% | 21% |
| 6 | 39 | 3 | S1 | 13 | 12 | 36% | 33.5% |
| | | | S2 | 16 | 16 | 44.5% | 44.5% |
| | | | S3 | 7 | 8 | 19.5% | 22% |
| 7 | 41 | 3 | S1 | 15 | 16 | 39% | 42% |
| | | | S2 | 15 | 15 | 39% | 40% |
| | | | S3 | 8 | 7 | 22% | 18% |
| 8 | 39 | 3 | S1 | 14 | 15 | 39% | 42% |
| | | | S2 | 14 | 13 | 39% | 36% |
| | | | S3 | 8 | 8 | 22% | 22% |

Tables 4a to 4f present Grades 3-8 ELA Tests items maps with item type indicator, answer key, maximum number of points obtainable from each item and learning standard measured by each item.

Table 4a. NYSTP ELA 2006 Operational Test Map, Grade 3

| Test | Item# | Item Type | Answer Key | Max Points | Standard |
|-------|-------|-----------|------------|------------|----------|
| 3 ELA | 1 | MC | B | 1 | S1 |
| 3 ELA | 2 | MC | H | 1 | S1 |
| 3 ELA | 3 | MC | A | 1 | S1 |
| 3 ELA | 4 | MC | G | 1 | S1 |
| 3 ELA | 5 | MC | D | 1 | S1 |
| 3 ELA | 6 | MC | J | 1 | S2 |
| 3 ELA | 7 | MC | C | 1 | S2 |
| 3 ELA | 8 | MC | F | 1 | S2 |
| 3 ELA | 9 | MC | B | 1 | S3 |
| 3 ELA | 10 | MC | F | 1 | S3 |
| 3 ELA | 11 | MC | C | 1 | S1 |
| 3 ELA | 12 | MC | J | 1 | S1 |
| 3 ELA | 13 | MC | A | 1 | S1 |
| 3 ELA | 14 | MC | F | 1 | S3 |
| 3 ELA | 15 | MC | B | 1 | S1 |
| 3 ELA | 16 | MC | H | 1 | S3 |
| 3 ELA | 17 | MC | A | 1 | S2 |
| 3 ELA | 18 | CR | n/a | 2 | S2 |
| 3 ELA | 19 | MC | D | 1 | S2 |
| 3 ELA | 20 | MC | H | 1 | S2 |
| 3 ELA | 21 | MC | C | 1 | S3 |
| 3 ELA | 22 | MC | F | 1 | S2 |
| 3 ELA | 23 | MC | D | 1 | S2 |
| 3 ELA | 24 | MC | G | 1 | S2 |
| 3 ELA | 25 | CR | n/a | 2 | S2 |
| 3 ELA | 26 | CR | n/a | 2 | S2 |
| 3 ELA | 27 | MC | C | 1 | S3 |
| 3 ELA | 28 | CR | n/a | 3 | n/a |

Table 4b. NYSTP ELA 2006 Operational Test Map, Grade 4

| Test | Item# | Item Type | Answer Key | Max Points | Standard |
|---------------------------------|-------|-----------|------------|------------|----------|
| 4 ELA | 1 | MC | D | 1 | S1 |
| 4 ELA | 2 | MC | H | 1 | S1 |
| 4 ELA | 3 | MC | C | 1 | S1 |
| 4 ELA | 4 | MC | G | 1 | S3 |
| 4 ELA | 5 | MC | D | 1 | S2 |
| 4 ELA | 6 | MC | F | 1 | S2 |
| 4 ELA | 7 | MC | B | 1 | S2 |
| 4 ELA | 8 | MC | H | 1 | S3 |
| 4 ELA | 9 | MC | A | 1 | S2 |
| 4 ELA | 10 | MC | G | 1 | S2 |
| 4 ELA | 11 | MC | B | 1 | S1 |
| 4 ELA | 12 | MC | J | 1 | S1 |
| 4 ELA | 13 | MC | C | 1 | S1 |
| 4 ELA | 14 | MC | G | 1 | S1 |
| 4 ELA | 15 | MC | A | 1 | S1 |
| 4 ELA | 16 | MC | J | 1 | S1 |
| 4 ELA | 17 | MC | C | 1 | S3 |
| 4 ELA | 18 | MC | G | 1 | S2 |
| 4 ELA | 19 | MC | A | 1 | S2 |
| 4 ELA | 20 | MC | H | 1 | S2 |
| 4 ELA | 21 | MC | D | 1 | S2 |
| 4 ELA | 22 | MC | F | 1 | S2 |
| 4 ELA | 23 | MC | C | 1 | S2 |
| 4 ELA | 24 | MC | G | 1 | S1 |
| 4 ELA | 25 | MC | D | 1 | S3 |
| 4 ELA | 26 | MC | F | 1 | S1 |
| 4 ELA | 27 | MC | C | 1 | S1 |
| 4 ELA | 28 | MC | H | 1 | S1 |
| 4 ELA (Listening) | 29 | CR | n/a | 4 | S2 |
| 4 ELA (Writing Mechanics) | 30 | CR | n/a | 3 | n/a |
| 4 ELA (Reading) | 31 | CR | n/a | 4 | S3 |

Table 4c. NYSTP ELA 2006 Operational Test Map, Grade 5

| Test | Item# | Item Type | Answer Key | Max Points | Standard |
|-------|-------|-----------|------------|------------|----------|
| 5 ELA | 1 | MC | B | 1 | S2 |
| 5 ELA | 2 | MC | F | 1 | S2 |
| 5 ELA | 3 | MC | B | 1 | S2 |
| 5 ELA | 4 | MC | H | 1 | S2 |
| 5 ELA | 5 | MC | B | 1 | S3 |
| 5 ELA | 6 | MC | J | 1 | S2 |
| 5 ELA | 7 | MC | A | 1 | S3 |
| 5 ELA | 8 | MC | H | 1 | S1 |
| 5 ELA | 9 | MC | B | 1 | S1 |
| 5 ELA | 10 | MC | G | 1 | S1 |
| 5 ELA | 11 | MC | A | 1 | S1 |
| 5 ELA | 12 | CR | n/a | 2 | S1 |
| 5 ELA | 13 | MC | A | 1 | S2 |
| 5 ELA | 14 | MC | J | 1 | S2 |
| 5 ELA | 15 | MC | B | 1 | S2 |
| 5 ELA | 16 | MC | H | 1 | S2 |
| 5 ELA | 17 | MC | C | 1 | S3 |
| 5 ELA | 18 | MC | G | 1 | S1 |
| 5 ELA | 19 | MC | C | 1 | S1 |
| 5 ELA | 20 | MC | J | 1 | S3 |
| 5 ELA | 21 | MC | C | 1 | S1 |
| 5 ELA | 22 | MC | G | 1 | S1 |
| 5 ELA | 23 | MC | B | 1 | S1 |
| 5 ELA | 24 | MC | F | 1 | S1 |
| 5 ELA | 25 | MC | D | 1 | S1 |
| 5 ELA | 26 | CR | n/a | 2 | S3 |
| 5 ELA | 27 | CR | n/a | 3 | n/a |

Table 4d. NYSTP ELA 2006 Operational Test Map, Grade 6

| Test | Item# | Item Type | Answer Key | Max Points | Standard |
|---------------------------------|-------|-----------|------------|------------|----------|
| 6 ELA | 1 | MC | D | 1 | S3 |
| 6 ELA | 2 | MC | G | 1 | S2 |
| 6 ELA | 3 | MC | A | 1 | S3 |
| 6 ELA | 4 | MC | J | 1 | S2 |
| 6 ELA | 5 | MC | D | 1 | S2 |
| 6 ELA | 6 | MC | F | 1 | S1 |
| 6 ELA | 7 | MC | A | 1 | S1 |
| 6 ELA | 8 | MC | J | 1 | S1 |
| 6 ELA | 9 | MC | A | 1 | S1 |
| 6 ELA | 10 | MC | G | 1 | S1 |
| 6 ELA | 11 | MC | D | 1 | S2 |
| 6 ELA | 12 | MC | G | 1 | S2 |
| 6 ELA | 13 | MC | A | 1 | S2 |
| 6 ELA | 14 | MC | G | 1 | S3 |
| 6 ELA | 15 | MC | C | 1 | S2 |
| 6 ELA | 16 | MC | J | 1 | S1 |
| 6 ELA | 17 | MC | B | 1 | S1 |
| 6 ELA | 18 | MC | H | 1 | S1 |
| 6 ELA | 19 | MC | A | 1 | S1 |
| 6 ELA | 20 | MC | J | 1 | S1 |
| 6 ELA | 21 | MC | B | 1 | S1 |
| 6 ELA | 22 | MC | G | 1 | S1 |
| 6 ELA | 23 | MC | B | 1 | S2 |
| 6 ELA | 24 | MC | F | 1 | S2 |
| 6 ELA | 25 | MC | B | 1 | S2 |
| 6 ELA | 26 | MC | G | 1 | S2 |
| 6 ELA (Listening) | 27 | CR | n/a | 5 | S2 |
| 6 ELA (Writing Mechanics) | 28 | CR | n/a | 3 | n/a |
| 6 ELA (Reading) | 29 | CR | n/a | 5 | S3 |

Table 4e. NYSTP ELA 2006 Operational Test Map, Grade 7

| Test | Item# | Item Type | Answer Key | Max Points | Standard |
|-------|-------|-----------|------------|------------|----------|
| 7 ELA | 1 | MC | D | 1 | S2 |
| 7 ELA | 2 | MC | H | 1 | S2 |
| 7 ELA | 3 | MC | C | 1 | S2 |
| 7 ELA | 4 | MC | F | 1 | S1 |
| 7 ELA | 5 | MC | A | 1 | S3 |
| 7 ELA | 6 | MC | J | 1 | S1 |
| 7 ELA | 7 | MC | B | 1 | S2 |
| 7 ELA | 8 | MC | F | 1 | S1 |
| 7 ELA | 9 | MC | B | 1 | S1 |
| 7 ELA | 10 | MC | G | 1 | S3 |
| 7 ELA | 11 | MC | A | 1 | S1 |
| 7 ELA | 12 | MC | F | 1 | S1 |
| 7 ELA | 13 | MC | C | 1 | S1 |
| 7 ELA | 14 | MC | G | 1 | S1 |
| 7 ELA | 15 | MC | C | 1 | S1 |
| 7 ELA | 16 | MC | F | 1 | S1 |
| 7 ELA | 17 | CR | n/a | 2 | S3 |
| 7 ELA | 18 | MC | F | 1 | S2 |
| 7 ELA | 19 | MC | C | 1 | S2 |
| 7 ELA | 20 | MC | H | 1 | S3 |
| 7 ELA | 21 | MC | B | 1 | S2 |
| 7 ELA | 22 | MC | G | 1 | S2 |
| 7 ELA | 23 | CR | n/a | 2 | S2 |
| 7 ELA | 24 | MC | J | 1 | S2 |
| 7 ELA | 25 | MC | C | 1 | S2 |
| 7 ELA | 26 | MC | H | 1 | S2 |
| 7 ELA | 27 | MC | D | 1 | S2 |
| 7 ELA | 28 | MC | G | 1 | S2 |
| 7 ELA | 29 | MC | B | 1 | S1 |
| 7 ELA | 30 | MC | F | 1 | S1 |
| 7 ELA | 31 | MC | C | 1 | S1 |
| 7 ELA | 32 | CR | n/a | 2 | S3 |
| 7 ELA | 33 | CR | n/a | 2 | S1 |
| 7 ELA | 34 | MC | G | 1 | S1 |
| 7 ELA | 35 | CR | n/a | 3 | n/a |

Table 4f. NYSTP ELA 2006 Operational Test Map, Grade 8

| Test | Item# | Item Type | Answer Key | Max Points | Standard |
|---------------------------------|-------|-----------|------------|------------|----------|
| 8 ELA | 1 | MC | A | 1 | S2 |
| 8 ELA | 2 | MC | G | 1 | S2 |
| 8 ELA | 3 | MC | C | 1 | S2 |
| 8 ELA | 4 | MC | H | 1 | S2 |
| 8 ELA | 5 | MC | B | 1 | S2 |
| 8 ELA | 6 | MC | J | 1 | S1 |
| 8 ELA | 7 | MC | C | 1 | S1 |
| 8 ELA | 8 | MC | H | 1 | S1 |
| 8 ELA | 9 | MC | A | 1 | S1 |
| 8 ELA | 10 | MC | H | 1 | S1 |
| 8 ELA | 11 | MC | B | 1 | S1 |
| 8 ELA | 12 | MC | H | 1 | S2 |
| 8 ELA | 13 | MC | A | 1 | S2 |
| 8 ELA | 14 | MC | F | 1 | S3 |
| 8 ELA | 15 | MC | A | 1 | S2 |
| 8 ELA | 16 | MC | H | 1 | S2 |
| 8 ELA | 17 | MC | C | 1 | S2 |
| 8 ELA | 18 | MC | J | 1 | S2 |
| 8 ELA | 19 | MC | C | 1 | S2 |
| 8 ELA | 20 | MC | G | 1 | S3 |
| 8 ELA | 21 | MC | A | 1 | S2 |
| 8 ELA | 22 | MC | F | 1 | S1 |
| 8 ELA | 23 | MC | A | 1 | S3 |
| 8 ELA | 24 | MC | G | 1 | S1 |
| 8 ELA | 25 | MC | D | 1 | S1 |
| 8 ELA | 26 | MC | G | 1 | S1 |
| 8 ELA (Listening) | 27 | CR | n/a | 5 | S1 |
| 8 ELA (Writing Mechanics) | 28 | CR | n/a | 3 | n/a |
| 8 ELA (Reading) | 29 | CR | n/a | 5 | S3 |

2006 Item Mapping by New York State Standards and Strands

Table 5. NYSTP ELA 2006 Standard Coverage

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

Content Rationale

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

- which performance indicators were to be assessed
- which item types were to be used for the assessable performance indicators
- 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 were comprised of teachers selected from around the state for their grade-level expertise. The committees 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.

Item Development

The first step in the process of item development for the Grades 3-8 ELA Tests was selection of passages. The CTB/McGraw-Hill passage selectors were provided with specifications based on the test design (see Appendix A). After an internal CTB/McGraw-Hill editorial and supervisory review, the passages were submitted to NYSED for their approval and then brought to a formal Passage Review meeting in Albany, New York in June 2004. The purpose of the meeting was for committees of New York educators to review and decide whether to approve the passages. CTB/McGraw-Hill and NYSED staff were both present, with CTB/McGraw-Hill staff facilitating. After the committees completed their reviews, NYSED reviewed and approved the committees' decisions regarding the passages.

The lead content editors at CTB/McGraw-Hill then selected from the approved passages those passages that would best elicit the types of items outlined during the specifications meetings and distributed writing assignments to experienced item writers. The writers' assignments outlined the number and type of items (including depth of knowledge or thinking skill level) to write for each passage. Writers were trained in the New York State Testing Program and in the test specifications. This training entails specific assignments that spell out the performance indicators and depth-of-knowledge levels to assess for each passage. In addition, item writers are trained in the New York State Standards and specifications (which provide information such as limitations and examples for assessing performance indicators), and are provided with item writing guidelines (see Appendix B), sample New York State items, and the New York State Style Guide.

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

Item Review

As was done for the Specifications and Passage Review meetings, 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 reading load of the items.
- the existence of any ethnic, gender, regional, or other possible bias evident in the items.

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

Materials Development

Following Item Review, CTB/McGraw-Hill staff assembled the approved passages and items into field test forms and submitted the field test forms to NYSED for their review and approval. The Grades 3-8 ELA Field Tests were administered to students across New York State during the week of February 7, 2005, using the State Matrix to ensure appropriate sampling of students. In addition, CTB/McGraw-Hill, in conjunction with NYSED test specialists, developed a field test *Teacher's Directions and School Administrator's Manual* to help insure that the field tests were administered in a uniform manner to all participating students. Field test forms were assigned to participants at the school (grade) level while balancing the demographic statistics across forms, in order to proactively sample the students.

After administration of the field tests, Rangefinding Meetings were conducted in March 2005 in New York State to examine a sampling of the short and extended student responses. Committees of New York State educators with content and grade-level expertise were again assembled. CTB/McGraw-Hill staff facilitated the meetings, and NYSED staff reviewed the decisions made by the committees and verified that the decisions made were consistent across grades. The committees' charge was to select student responses that exemplified each score point of each constructed-response item.

These responses, in conjunction with the rubrics, were then used by CTB/McGraw-Hill scoring staff to score the constructed-response field test items.

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 to: <http://emsc33.nysed.gov/3-8/intro.pdf>) as well as a sample test (which mirrored the operational test and also used field-tested items), a *Teacher's Directions*, and a *Scoring Guide* for each grade (posted to: <http://www.emsc.nysed.gov/3-8/ela-sample/home.htm>). This *Guide* was also printed and delivered to schools.

Item Selection and Test Creation (criteria and process)

The first operational NYSTP Grades 3-8 ELA Tests were administered in January 2006. The test items were selected from the pool of items field-tested in 2005, using the data from those field tests. CTB/McGraw-Hill made preliminary selections for each grade. The selections were reviewed for alignment with the test design, blueprint, and the Research guidelines for item selection (Appendix C). Item selection for the NYSTP Grades 3-8 ELA Tests was based on the classical and 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 C).

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 June 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 January 2006.

In addition to the test books, CTB/McGraw-Hill and NYSED produced two *School Administrator's Manuals* (one for public schools (see: <http://www.emsc.nysed.gov/3-8/sam/ela06p.pdf>) and one for nonpublic schools (see: <http://www.emsc.nysed.gov/3-8/sam/ela06np.pdf>) and *Teacher's Directions* (see: <http://www.emsc.nysed.gov/3-8/directions/home.htm>) for each grade 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 ELA Standard Setting held in Albany, in June 2006. The results were reviewed by a Measurement Review committee, and were approved in August 2006. For each grade level, there are four proficiency levels. Three cut points demarcate the performance standards needed to demonstrate each ascending level of proficiency. 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 English Language Arts* and *New York State ELA 2006 Measurement Review Technical Report 2006 for English Language Arts*.

Section III: Validity

Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests. Test validation is an 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 ELA as defined by the New York State ELA Learning Standards. To allow test score interpretations appropriate for this purpose, the content of the test must be carefully matched to the specified standards. The 1999 AERA/APA/NCME Standards state that content-related evidence of validity is a central concern during test development. Expert professional judgment should play an integral part in developing the definition of what is to be measured, such as describing the universe of the content, generating or selecting the content sample, and specifying the item format and scoring system.

Logical 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 3 to 5 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 ELA Tests was conducted using Norman Webb's method. The results of the study provided additional evidence of test content validity (refer to *An External Alignment Study for New York State's Assessment Program*, April 2006, Educational Testing Services)

Construct (Internal Structure) Validity

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

Internal consistency

Empirical studies of the internal structure of the test provide one type of evidence of construct validity. For example, high internal consistency constitutes evidence of validity. This is because high coefficients imply that the test questions are measuring the same domain of skill, 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 0.82 to 0.89 and for most subgroups the reliability coefficient was greater than 0.80 (the exception was for grade 5 students from districts classified as Low Need). Overall, high internal consistency of New York State ELA 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 on the 2006 Grades 3-8 ELA 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 ELA 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., English Language Arts ability.

To demonstrate the common factor (ability) underlying student responses ELA test items, a principal component factor analysis was conducted on a correlation matrix of individual items for each test. Factoring a correlation matrix rather than actual item response data is preferable when dichotomous variables are in the analyzed data set. Because the New York State 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 D 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 4 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 ELA Tests exhibited first principle components accounting for more than 10 percent of the test variance. The results of factor analysis including eigenvalues greater than 1.0 and proportion of variance explained by extracted factors are presented in Table 6, below.

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

| Grade | Initial Eigenvalues | | | |
|-------|---------------------|-------------|---------------|--------------|
| | Component | Total | % of Variance | Cumulative % |
| 3 | 1 | 6.02 | 21.49 | 21.49 |
| | 2 | 1.28 | 4.56 | 26.05 |
| | 3 | 1.03 | 3.68 | 29.73 |
| 4 | 1 | 7.24 | 23.35 | 23.35 |
| | 2 | 1.20 | 3.86 | 27.21 |
| | 3 | 1.08 | 3.49 | 30.70 |
| 5 | 1 | 5.21 | 19.28 | 19.28 |
| | 2 | 1.16 | 4.30 | 23.58 |
| | 3 | 1.13 | 4.19 | 27.77 |
| 6 | 1 | 6.06 | 20.89 | 20.89 |
| | 2 | 1.22 | 4.19 | 25.09 |
| | 3 | 1.05 | 3.60 | 28.69 |

(Continued on next page)

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

| Grade | Initial Eigenvalues | | | |
|-------|---------------------|-------------|---------------|--------------|
| | Component | Total | % of Variance | Cumulative % |
| 7 | 1 | 6.91 | 19.75 | 19.75 |
| | 2 | 1.22 | 3.48 | 23.23 |
| | 3 | 1.13 | 3.22 | 26.44 |
| | 4 | 1.07 | 3.05 | 29.49 |
| 8 | 1 | 6.17 | 21.28 | 21.28 |
| | 2 | 1.20 | 4.15 | 25.43 |
| | 3 | 1.04 | 3.58 | 29.01 |
| | 4 | 1.00 | 3.45 | 32.46 |

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

As an additional evidence for construct validity, the same factor analysis procedure was employed to assess dimensionality of ELA construct for selected subgroups of students in each grade: Limited English Proficiency (LEP) students, students with disabilities (SWD), and students using test accommodations (SUA). The results were comparable to the results obtained from the total population data. Evaluation of eigenvalue magnitude and proportions of variance explained by the main and secondary factors provide evidence of essential unidimensionality of the construct measured by the ELA tests for the analyzed subgroups. Factor analysis results for LEP, SWD and students using accommodations are provided in Table D1 of Appendix D 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, 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 ELA 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 filed 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). A few items in each grade were flagged for DIF and typically the amount of DIF present was not large. Very few items were flagged by multiple methods. Items that were flagged for statistically significant DIF were carefully reviewed by multiple reviewers during the operational test item selection. Only 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 (for details, please see Section X: ELA Linking Study). An equipercentile linking that was conducted to provide a cross-walk between the two years of testing relies on an assumption of no year-to-year growth and as such will hinder growth assessments between 2005 and 2006.

Given the limitations of the first year test data, it is advisable 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 ELA Tests were administered at the classroom level, during January, 2006. The testing window for grades 3, 4, and 5 was January 9th through 13th, 2006. The testing window for grades 6, 7, and 8 was January 17th through 20th. The make-up test administration window was January 23rd through 27th, which 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 ELA Leader” trained and monitored sessions, and a “School ELA 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 ELA Tests. Schools were able to score these tests regionally, district-wide, or individually. Schools were required to enter one of the following “scoring model codes” on student answer sheets:

1. Regional scoring – The 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 Spring 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 SED. 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, ELA FAQs (at: <http://www.emsc.nysed.gov/3-8/faq/ela-scoring06.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 ELA 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, ELA 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 Warehouse, 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 April 2006.

Not all test data were uploaded to the 100% Data files. For example, only public schools data were submitted to the Data Warehouse. 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 Warehouse 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 ELA 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 7a-7f, below.

Table 7a. NYSTP ELA Data Cleaning

| Dataset | Exclusion Rule | N. Deleted | N. Cases Remain |
|--------------|-----------------------|------------|-----------------|
| Grade 3 100% | | | 184,868 |
| Grade 3 100% | Out of grade | 1,319 | 183,549 |
| Grade 3 100% | Duplicate student ID | 0 | 183,549 |
| Grade 3 100% | Duplicate string | 0 | 183,549 |
| Grade 3 100% | LEP = Yes | 3,802 | 179,747 |
| Grade 3 100% | Out-of-range response | 0 | 179,747 |
| Grade 3 100% | Invalid Score | 195 | 179,552 |

Table 7b. NYSTP ELA Data Cleaning

| Dataset | Exclusion Rule | N. Deleted | N. Cases Remain |
|--------------|-----------------------|------------|-----------------|
| Grade 4 100% | | | 189,213 |
| Grade 4 100% | Out of grade | 1,147 | 188,066 |
| Grade 4 100% | Duplicate student ID | 0 | 188,066 |
| Grade 4 100% | Duplicate string | 4 | 188,062 |
| Grade 4 100% | LEP = Yes | 4,366 | 183,696 |
| Grade 4 100% | Out-of-range response | 0 | 183,696 |
| Grade 4 100% | Invalid Score | 354 | 183,342 |

Table 7c. NYSTP ELA Data Cleaning

| Dataset | Exclusion Rule | N. Deleted | N. Cases Remain |
|--------------|-----------------------|------------|-----------------|
| Grade 5 100% | | | 199,245 |
| Grade 5 100% | Out of grade | 1,386 | 197,859 |
| Grade 5 100% | Duplicate student ID | 0 | 197,859 |
| Grade 5 100% | Duplicate string | 4 | 197,855 |
| Grade 5 100% | LEP = Yes | 6,434 | 191,421 |
| Grade 5 100% | Out-of-range response | 0 | 191,421 |
| Grade 5 100% | Invalid Score | 260 | 191,161 |

Table 7d. NYSTP ELA Data Cleaning

| Dataset | Exclusion Rule | N. Deleted | N. Cases Remain |
|--------------|-----------------------|------------|-----------------|
| Grade 6 100% | | | 200,750 |
| Grade 6 100% | Out of grade | 1,357 | 199,393 |
| Grade 6 100% | Duplicate student ID | 0 | 199,393 |
| Grade 6 100% | Duplicate string | 2 | 199,391 |
| Grade 6 100% | LEP = Yes | 5,425 | 193,966 |
| Grade 6 100% | Out-of-range response | 0 | 193,966 |
| Grade 6 100% | Invalid Score | 612 | 193,354 |

Table 7e. NYSTP ELA Data Cleaning

| Dataset | Exclusion Rule | N. Deleted | N. Cases Remain |
|--------------|-----------------------|------------|-----------------|
| Grade 7 100% | | | 207,622 |
| Grade 7 100% | Out of grade | 812 | 206,810 |
| Grade 7 100% | Duplicate student ID | 2 | 206,808 |
| Grade 7 100% | Duplicate string | 2 | 206,806 |
| Grade 7 100% | LEP = Yes | 6,099 | 200,707 |
| Grade 7 100% | Out-of-range response | 0 | 200,707 |
| Grade 7 100% | Invalid Score | 499 | 200,208 |

Table 7f. NYSTP ELA Data Cleaning

| Dataset | Exclusion Rule | N. Deleted | N. Cases Remain |
|--------------|-----------------------|------------|-----------------|
| Grade 8 100% | | | 209,458 |
| Grade 8 100% | Out of grade | 577 | 208,881 |
| Grade 8 100% | Duplicate student ID | 4 | 208,877 |
| Grade 8 100% | Duplicate string | 2 | 208,875 |
| Grade 8 100% | LEP = Yes | 5,849 | 203,026 |
| Grade 8 100% | Out-of-range response | 0 | 203,026 |
| Grade 8 100% | Invalid Score | 602 | 202,424 |

Classical Analysis and Calibration Sample Characteristics

The demographic characteristics of students in the cleaned calibration and equating datasets are presented in the proceeding tables (see next page). The clean data sets included over 90% of New York State students and were used for classical analyses presented in this section and calibrations. The Needs Resource Code (NRC) is assigned at district-level and is an indicator of district and school socioeconomic status. The ethnicity and gender designations are assigned at the student level. Please note that the tables do not include data for gender variable as it was found that the New York State population is fairly evenly split by gender categories (Males and Females).

Table 8a. Grade 3 Sample Characteristics (N=179,552)

| | Demographic Category | N-count | Percent of total N |
|-----------|---|---------|--------------------|
| NRC | New York City | 58727 | 32.71 |
| | Big 4 Cities | 5548 | 3.09 |
| | Urban-suburban | 13397 | 7.46 |
| | Rural | 11403 | 6.35 |
| | Average Needs | 57678 | 32.12 |
| | Low Needs | 29273 | 16.30 |
| | Charter | 2199 | 1.22 |
| | (unassigned) | 1327 | 0.74 |
| Ethnicity | Asian | 11219 | 6.25 |
| | Black or African American | 37646 | 20.97 |
| | Hispanic or Latino | 26895 | 14.98 |
| | American Indian or Alaska Native | 970 | 0.54 |
| | Native Hawaiian /Other Pacific Islander | 39 | 0.02 |
| | White | 102770 | 57.24 |
| | Blank (No Response) | 13 | 0.01 |

Table 8b. Grade 4 Sample Characteristics (N=183,342)

| | Demographic Category | N-count | Percent of total N |
|-----------|---|---------|--------------------|
| NRC | New York City | 61164 | 33.36 |
| | Big 4 Cities | 5268 | 2.87 |
| | Urban-suburban | 13668 | 7.45 |
| | Rural | 11590 | 6.32 |
| | Average Needs | 59231 | 32.31 |
| | Low Needs | 30328 | 16.54 |
| | Charter | 1399 | 0.76 |
| | (unassigned) | 694 | 0.38 |
| Ethnicity | Asian | 12046 | 6.57 |
| | Black or African American | 36099 | 19.69 |
| | Hispanic or Latino | 29114 | 15.88 |
| | American Indian or Alaska Native | 942 | 0.51 |
| | Native Hawaiian /Other Pacific Islander | 37 | 0.02 |
| | White | 105099 | 57.32 |
| | Blank (No Response) | 5 | 0 |

Table 8c. Grade 5 Sample Characteristics (N=191,161)

| Demographic Category | | N-count | Percent of total N |
|----------------------|---|---------|--------------------|
| NRC | New York City | 63640 | 33.29 |
| | Big 4 Cities | 5670 | 2.97 |
| | Urban-suburban | 14049 | 7.35 |
| | Rural | 11938 | 6.24 |
| | Average Needs | 61398 | 32.12 |
| | Low Needs | 30832 | 16.13 |
| | Charter | 2142 | 1.12 |
| | (unassigned) | 1492 | 0.78 |
| Ethnicity | Asian | 12462 | 6.52 |
| | Black or African American | 38139 | 19.95 |
| | Hispanic or Latino | 30996 | 16.21 |
| | American Indian or Alaska Native | 1012 | 0.53 |
| | Native Hawaiian /Other Pacific Islander | 49 | 0.03 |
| | White | 108499 | 56.76 |
| | Blank (No Response) | 4 | 0 |

Table 8d. Grade 6 Sample Characteristics (N=193,354)

| Demographic Category | | N-count | Percent of total N |
|----------------------|---|---------|--------------------|
| NRC | New York City | 62180 | 32.16 |
| | Big 4 Cities | 5910 | 3.06 |
| | Urban-suburban | 14664 | 7.58 |
| | Rural | 12777 | 6.61 |
| | Average Needs | 63712 | 32.95 |
| | Low Needs | 31033 | 16.05 |
| | Charter | 1719 | 0.89 |
| | (unassigned) | 1359 | 0.70 |
| Ethnicity | Asian | 12150 | 6.28 |
| | Black or African American | 38241 | 19.78 |
| | Hispanic or Latino | 31119 | 16.09 |
| | American Indian or Alaska Native | 1097 | 0.57 |
| | Native Hawaiian /Other Pacific Islander | 38 | 0.02 |
| | White | 110708 | 57.26 |
| | Blank (No Response) | 1 | 0 |

Table 8e. Grade 7 Sample Characteristics (N=200,208)

| Demographic Category | | N-count | Percent of total N |
|----------------------|---|---------|--------------------|
| NRC | New York City | 63450 | 31.69 |
| | Big 4 Cities | 7068 | 3.53 |
| | Urban-suburban | 15020 | 7.50 |
| | Rural | 13556 | 6.77 |
| | Average Needs | 66855 | 33.39 |
| | Low Needs | 31113 | 15.54 |
| | Charter | 1416 | 0.71 |
| | (unassigned) | 1730 | 0.86 |
| Ethnicity | Asian | 11830 | 5.91 |
| | Black or African American | 40960 | 20.46 |
| | Hispanic or Latino | 31441 | 15.70 |
| | American Indian or Alaska Native | 1072 | 0.54 |
| | Native Hawaiian /Other Pacific Islander | 44 | 0.02 |
| | White | 114858 | 57.37 |
| | Blank (No Response) | 3 | 0 |

Table 8f. Grade 8 Sample Characteristics (N=202,424)

| Demographic Category | | N-count | Percent of total N |
|----------------------|---|---------|--------------------|
| NRC | New York City | 64507 | 31.87 |
| | Big 4 Cities | 7628 | 3.77 |
| | Urban-suburban | 14717 | 7.27 |
| | Rural | 13550 | 6.69 |
| | Average Needs | 68329 | 33.76 |
| | Low Needs | 31198 | 15.41 |
| | Charter | 859 | 0.42 |
| | (unassigned) | 1636 | 0.81 |
| Ethnicity | Asian | 11781 | 5.82 |
| | Black or African American | 41033 | 20.27 |
| | Hispanic or Latino | 31296 | 15.46 |
| | American Indian or Alaska Native | 1020 | 0.50 |
| | Native Hawaiian /Other Pacific Islander | 29 | 0.01 |
| | White | 117265 | 57.93 |
| | Blank (No Response) | 0 | 0.00 |

Classical Data Analysis

Classical data analysis of the Grades 3-8 ELA Tests consists of four primary elements. One element is the analysis of item level statistical information about student performance. It is important to verify that the items and test forms function as intended. Information on item response patterns, item difficulty (p-value) and item-test correlation (point biserial) is examined thoroughly. If any serious error 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 Difficulty and Response Distribution

Item difficulty and response distribution tables (Table 9a-9f) illustrate student test performance, as observed from both MC and CR item responses. Omit rates signify the 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”, “PCT SEL”, and “% at 5” (in grades 6, and 8 only) 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 0.244 to 0.958. For grade 3, the item p-values were between 0.505 and 0.958 with a mean of 0.76. For grade 4, the item p-values were between 0.459 and 0.921 with a mean of 0.71. For grade 5, the item p-values were between 0.244 and 0.902 with a mean of 0.70. For grade 6, the item p-values were between 0.414 and 0.946 with a mean of 0.69. For grade 7, the item p-values were between 0.298 and 0.935 with a mean of 0.70. For grade 8, the item p-values

were between 0.357 and 0.897 with a mean of 0.72. These statistics are also provided in Table 10, along with other classical test summary statistics.

Table 9a. P-values, Scored Response Distributions, and Point Biserials, 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 | 179409 | 0.895 | 0.04 | 0.04 | 3.54 | *89.38 | 4.23 | 2.77 | -0.24 | *0.37 | -0.19 | -0.19 | 0.37 |
| 2 | 179362 | 0.958 | 0.07 | 0.03 | 2.35 | 1.07 | *95.73 | 0.74 | -0.16 | -0.20 | *0.31 | -0.20 | 0.31 |
| 3 | 179176 | 0.857 | 0.17 | 0.04 | *85.53 | 6.19 | 3.54 | 4.52 | *0.41 | -0.19 | -0.25 | -0.24 | 0.41 |
| 4 | 179037 | 0.902 | 0.25 | 0.04 | 3.52 | *89.99 | 2.33 | 3.87 | -0.26 | *0.46 | -0.24 | -0.26 | 0.46 |
| 5 | 178844 | 0.871 | 0.33 | 0.07 | 7.85 | 1.14 | 3.89 | *86.73 | -0.23 | -0.22 | -0.23 | *0.39 | 0.39 |
| 6 | 179309 | 0.639 | 0.07 | 0.06 | 16.45 | 11.64 | 8.00 | *63.77 | -0.20 | -0.10 | -0.25 | *0.36 | 0.36 |
| 7 | 179279 | 0.836 | 0.11 | 0.04 | 3.53 | 5.84 | *83.47 | 7.01 | -0.27 | -0.22 | *0.42 | -0.22 | 0.42 |
| 8 | 179069 | 0.751 | 0.22 | 0.05 | *74.90 | 5.91 | 15.11 | 3.81 | *0.35 | -0.21 | -0.17 | -0.21 | 0.35 |
| 9 | 178921 | 0.803 | 0.32 | 0.04 | 3.89 | *79.98 | 5.21 | 10.56 | -0.29 | *0.44 | -0.26 | -0.20 | 0.44 |
| 10 | 178791 | 0.929 | 0.39 | 0.03 | *92.52 | 1.72 | 1.50 | 3.83 | *0.33 | -0.17 | -0.23 | -0.18 | 0.33 |
| 11 | 179302 | 0.886 | 0.11 | 0.03 | 2.62 | 5.21 | *88.50 | 3.54 | -0.29 | -0.25 | *0.49 | -0.28 | 0.49 |
| 12 | 179172 | 0.634 | 0.17 | 0.04 | 5.72 | 6.77 | 24.02 | *63.28 | -0.27 | -0.32 | -0.15 | *0.43 | 0.43 |
| 13 | 179130 | 0.917 | 0.21 | 0.03 | *91.46 | 3.37 | 3.69 | 1.24 | *0.34 | -0.13 | -0.27 | -0.19 | 0.34 |
| 14 | 178967 | 0.505 | 0.27 | 0.05 | *50.36 | 12.60 | 8.39 | 28.32 | *0.26 | -0.22 | -0.15 | -0.03 | 0.26 |
| 15 | 178724 | 0.622 | 0.42 | 0.04 | 14.64 | *61.87 | 19.57 | 3.46 | -0.24 | *0.38 | -0.14 | -0.22 | 0.38 |
| 16 | 178612 | 0.839 | 0.50 | 0.02 | 4.01 | 8.85 | *83.45 | 3.16 | -0.25 | -0.24 | *0.46 | -0.28 | 0.46 |
| 17 | 178717 | 0.607 | 0.42 | 0.04 | *60.39 | 11.82 | 7.37 | 19.95 | *0.35 | -0.11 | -0.25 | -0.18 | 0.35 |
| 18 | 177785 | 0.936 | 0.98 | 3.46 | 5.71 | 89.84 | | | | | | | |
| 19 | 178736 | 0.838 | 0.43 | 0.03 | 10.21 | 2.24 | 3.63 | *83.46 | -0.35 | -0.26 | -0.33 | *0.56 | 0.56 |
| 20 | 178466 | 0.537 | 0.58 | 0.02 | 15.57 | 9.52 | *53.40 | 20.90 | -0.02 | -0.17 | *0.36 | -0.30 | 0.36 |
| 21 | 178120 | 0.653 | 0.77 | 0.03 | 5.21 | 8.11 | *64.74 | 21.14 | -0.22 | -0.14 | *0.31 | -0.15 | 0.31 |
| 22 | 179418 | 0.924 | 0.05 | 0.03 | *92.29 | 2.40 | 1.68 | 3.55 | *0.35 | -0.22 | -0.20 | -0.18 | 0.35 |
| 23 | 179399 | 0.835 | 0.05 | 0.03 | 6.01 | 5.00 | 5.48 | *83.42 | -0.27 | -0.05 | -0.17 | *0.31 | 0.31 |
| 24 | 179310 | 0.867 | 0.11 | 0.02 | 1.08 | *86.60 | 4.24 | 7.95 | -0.14 | | -0.22 | -0.09 | 0.24 |
| 25 | 178722 | 0.572 | 0.46 | 20.86 | 43.49 | 35.18 | | | | | | | |
| 26 | 177152 | 0.571 | 1.34 | 22.13 | 40.43 | 36.10 | | | | | | | |
| 27 | 178705 | 0.682 | 0.45 | 0.03 | 4.70 | 6.52 | *67.85 | 20.45 | -0.17 | *0.24 | *0.40 | -0.26 | 0.40 |
| 28 | 179138 | 0.780 | 0.23 | 9.16 | 9.70 | 18.87 | 62.03 | | | | | | |

Table 9b. P-values, Scored Response Distributions, and Point Biserials, 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 | 183272 | 0.752 | 0.02 | 0.02 | 21.35 | 1.10 | 2.32 | *75.19 | -0.37 | -0.16 | -0.19 | *0.45 | 0.45 |
| 2 | 183227 | 0.705 | 0.05 | 0.01 | 14.86 | 4.29 | *70.47 | 10.32 | -0.11 | -0.25 | *0.31 | -0.17 | 0.31 |
| 3 | 183198 | 0.807 | 0.06 | 0.02 | 5.71 | 6.78 | *80.68 | 6.75 | -0.29 | -0.23 | *0.50 | -0.28 | 0.50 |
| 4 | 183234 | 0.722 | 0.04 | 0.02 | 6.83 | *72.11 | 10.59 | 10.41 | -0.25 | *0.39 | -0.25 | -0.13 | 0.39 |
| 5 | 183192 | 0.921 | 0.05 | 0.03 | 5.09 | 1.58 | 1.26 | *91.99 | -0.21 | -0.15 | -0.2 | *0.33 | 0.33 |
| 6 | 183175 | 0.693 | 0.06 | 0.03 | *69.20 | 3.18 | 8.29 | 19.24 | *0.43 | -0.21 | -0.23 | -0.25 | 0.43 |
| 7 | 183169 | 0.783 | 0.06 | 0.03 | 2.60 | *78.24 | 5.44 | 13.63 | -0.15 | *0.38 | -0.24 | -0.23 | 0.38 |
| 8 | 183119 | 0.751 | 0.09 | 0.03 | 7.33 | 6.21 | *74.98 | 11.35 | -0.21 | -0.07 | *0.22 | -0.07 | 0.22 |
| 9 | 183122 | 0.735 | 0.10 | 0.02 | *73.43 | 5.65 | 1.49 | 19.32 | *0.26 | -0.27 | -0.21 | -0.06 | 0.26 |
| 10 | 183191 | 0.826 | 0.05 | 0.03 | 4.52 | *82.49 | 1.84 | 11.07 | -0.19 | *0.27 | -0.15 | -0.14 | 0.27 |
| 11 | 183153 | 0.806 | 0.08 | 0.02 | 5.55 | *80.49 | 5.64 | 8.22 | -0.29 | *0.41 | -0.2 | -0.17 | 0.41 |
| 12 | 183038 | 0.845 | 0.13 | 0.04 | 6.21 | 4.72 | 4.49 | *84.40 | -0.27 | -0.29 | -0.23 | *0.49 | 0.49 |
| 13 | 183093 | 0.578 | 0.12 | 0.02 | 31.26 | 7.17 | *57.76 | 3.67 | -0.11 | -0.22 | *0.31 | -0.24 | 0.31 |
| 14 | 183114 | 0.806 | 0.11 | 0.02 | 10.00 | *80.51 | 3.36 | 5.99 | -0.27 | *0.44 | -0.18 | -0.25 | 0.44 |
| 15 | 183120 | 0.708 | 0.11 | 0.01 | *70.74 | 23.83 | 3.37 | 1.94 | *0.36 | -0.20 | -0.26 | -0.23 | 0.36 |
| 16 | 182975 | 0.737 | 0.17 | 0.03 | 15.67 | 3.55 | 6.99 | *73.59 | -0.16 | -0.24 | -0.24 | *0.37 | 0.37 |
| 17 | 182949 | 0.907 | 0.20 | 0.02 | 3.61 | 2.36 | *90.46 | 3.35 | -0.26 | -0.25 | *0.44 | -0.24 | 0.44 |
| 18 | 182376 | 0.756 | 0.50 | 0.03 | 12.68 | *75.17 | 5.26 | 6.36 | -0.25 | *0.49 | -0.25 | -0.28 | 0.49 |
| 19 | 182270 | 0.651 | 0.54 | 0.04 | *64.69 | 9.77 | 12.05 | 12.91 | *0.31 | -0.2 | -0.25 | -0.02 | 0.31 |
| 20 | 182044 | 0.528 | 0.66 | 0.05 | 13.67 | 10.78 | *52.41 | 22.43 | -0.20 | -0.24 | *0.47 | -0.22 | 0.47 |
| 21 | 181946 | 0.639 | 0.73 | 0.03 | 22.40 | 5.78 | 7.68 | *63.38 | -0.15 | -0.25 | -0.31 | *0.42 | 0.42 |
| 22 | 181724 | 0.705 | 0.84 | 0.04 | *69.89 | 9.98 | 3.35 | 15.90 | *0.36 | -0.22 | -0.25 | -0.15 | 0.36 |
| 23 | 181609 | 0.715 | 0.91 | 0.03 | 6.50 | 16.77 | *70.85 | 4.93 | -0.27 | -0.16 | *0.38 | -0.21 | 0.38 |
| 24 | 180360 | 0.459 | 1.59 | 0.04 | 29.09 | *45.14 | 13.95 | 10.19 | -0.05 | *0.32 | -0.22 | -0.20 | 0.32 |
| 25 | 179895 | 0.635 | 1.83 | 0.05 | 11.45 | 6.36 | 17.99 | *62.32 | -0.17 | -0.24 | -0.23 | *0.42 | 0.42 |
| 26 | 179302 | 0.751 | 2.16 | 0.05 | *73.42 | 5.25 | 14.72 | 4.41 | *0.47 | -0.27 | -0.27 | -0.23 | 0.47 |
| 27 | 179160 | 0.856 | 2.25 | 0.03 | 7.58 | 3.46 | *83.64 | 3.04 | -0.20 | -0.21 | *0.41 | -0.28 | 0.41 |
| 28 | 178695 | 0.736 | 2.52 | 0.02 | 7.42 | 10.39 | *71.70 | 7.96 | -0.28 | -0.16 | *0.45 | -0.29 | 0.45 |
| 29 | 183238 | 0.630 | 0.06 | 0.84 | 11.04 | 36.17 | 39.17 | 12.72 | | | | | |
| 30 | 183256 | 0.731 | 0.05 | 1.00 | 14.64 | 48.35 | 35.96 | | | | | | |
| 31 | 183342 | 0.682 | 0.00 | 0.53 | 6.91 | 30.82 | 42.71 | 19.03 | | | | | |

Table 9c. P-values, Scored Response Distributions, and Point Biserials, 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 | 191023 | 0.832 | 0.06 | 0.01 | 4.26 | *83.18 | 9.15 | 3.34 | -0.19 | *0.35 | -0.23 | -0.14 | 0.35 |
| 2 | 191049 | 0.880 | 0.05 | 0.01 | *87.93 | 0.81 | 10.53 | 0.67 | *0.14 | -0.13 | -0.09 | -0.10 | 0.14 |
| 3 | 190955 | 0.244 | 0.09 | 0.02 | 60.95 | *24.42 | 5.86 | 8.66 | -0.11 | *0.35 | -0.19 | -0.20 | 0.35 |
| 4 | 191020 | 0.892 | 0.06 | 0.01 | 8.07 | 1.28 | *89.17 | 1.40 | -0.21 | -0.20 | *0.34 | -0.22 | 0.34 |
| 5 | 190952 | 0.508 | 0.09 | 0.02 | 9.13 | *50.79 | 31.17 | 8.79 | -0.17 | *0.24 | 0.04 | -0.32 | 0.24 |
| 6 | 190975 | 0.811 | 0.08 | 0.02 | 7.70 | 4.55 | 6.63 | *81.03 | -0.23 | -0.17 | -0.12 | *0.33 | 0.33 |
| 7 | 190958 | 0.902 | 0.09 | 0.01 | *90.10 | 3.10 | 3.46 | 3.23 | *0.39 | -0.20 | -0.26 | -0.21 | 0.39 |
| 8 | 190941 | 0.710 | 0.11 | 0.01 | 5.85 | 20.71 | *70.96 | 2.36 | -0.24 | -0.17 | *0.34 | -0.18 | 0.34 |
| 9 | 190879 | 0.586 | 0.13 | 0.02 | 11.24 | *58.55 | 22.71 | 7.35 | -0.25 | *0.4 | -0.19 | -0.14 | 0.40 |
| 10 | 190889 | 0.730 | 0.13 | 0.01 | 3.70 | *72.87 | 5.05 | 18.24 | -0.24 | *0.21 | -0.20 | -0.01 | 0.21 |
| 11 | 190918 | 0.902 | 0.12 | 0.01 | *90.08 | 1.43 | 2.84 | 5.52 | *0.36 | -0.17 | -0.23 | -0.21 | 0.36 |
| 12 | 189895 | 0.678 | 0.66 | 14.90 | 34.14 | 50.30 | | | | | | | |
| 13 | 190832 | 0.715 | 0.16 | 0.01 | *71.40 | 3.83 | 17.80 | 6.80 | *0.24 | -0.18 | -0.13 | -0.10 | 0.24 |
| 14 | 190776 | 0.814 | 0.18 | 0.02 | 12.70 | 3.40 | 2.45 | *81.24 | -0.27 | -0.16 | -0.18 | *0.38 | 0.38 |
| 15 | 190703 | 0.803 | 0.23 | 0.01 | 7.23 | *80.15 | 5.97 | 6.41 | -0.07 | *0.26 | -0.12 | -0.23 | 0.26 |
| 16 | 190576 | 0.440 | 0.28 | 0.03 | 34.61 | 8.37 | *43.86 | 12.85 | 0.12 | -0.18 | *0.10 | -0.17 | 0.10 |
| 17 | 190614 | 0.601 | 0.26 | 0.03 | 2.80 | 1.52 | *59.93 | 35.45 | -0.21 | -0.19 | *0.18 | -0.07 | 0.18 |
| 18 | 190256 | 0.797 | 0.46 | 0.01 | 9.75 | *79.36 | 4.96 | 5.46 | -0.30 | *0.47 | -0.22 | -0.23 | 0.47 |
| 19 | 190061 | 0.757 | 0.56 | 0.02 | 13.47 | 3.30 | *75.27 | 7.39 | -0.23 | -0.22 | *0.37 | -0.15 | 0.37 |
| 20 | 189967 | 0.804 | 0.61 | 0.02 | 7.11 | 9.98 | 2.36 | *79.93 | -0.29 | -0.23 | -0.22 | *0.45 | 0.45 |
| 21 | 189843 | 0.617 | 0.68 | 0.01 | 16.68 | 7.76 | *61.24 | 13.63 | -0.16 | -0.16 | *0.34 | -0.18 | 0.34 |
| 22 | 191021 | 0.549 | 0.06 | 0.01 | 10.85 | *54.83 | 30.06 | 4.18 | -0.07 | *0.28 | -0.17 | -0.19 | 0.28 |
| 23 | 191036 | 0.855 | 0.05 | 0.01 | 5.80 | *85.45 | 4.83 | 3.85 | -0.34 | *0.46 | -0.19 | -0.21 | 0.46 |
| 24 | 190993 | 0.580 | 0.06 | 0.02 | *57.94 | 12.14 | 17.43 | 12.41 | *0.35 | -0.22 | -0.08 | -0.22 | 0.35 |
| 25 | 190904 | 0.804 | 0.12 | 0.01 | 5.44 | 6.04 | 8.05 | *80.33 | -0.21 | -0.20 | -0.23 | *0.40 | 0.40 |
| 26 | 190932 | 0.694 | 0.12 | 12.22 | 36.62 | 51.04 | | | | | | | |
| 27 | 190655 | 0.612 | 0.26 | 12.82 | 22.08 | 33.39 | 31.44 | | | | | | |

Table 9d. 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 | % at 5 | Pbis Option 1 | Pbis Option 2 | Pbis Option 3 | Pbis Option 4 | Pbis Key |
|------|---------|---------|--------|--------|------------------|------------------|------------------|------------------|--------|---------------|---------------|---------------|---------------|----------|
| 1 | 193316 | 0.940 | 0.02 | 0.00 | 3.59 | 0.69 | 1.69 | *94.01 | | -0.26 | -0.11 | -0.13 | *0.32 | 0.32 |
| 2 | 193180 | 0.537 | 0.08 | 0.01 | 15.20 | *53.62 | 18.10 | 13.00 | | -0.22 | *0.21 | 0.00 | -0.08 | 0.21 |
| 3 | 193311 | 0.946 | 0.01 | 0.01 | *94.55 | 0.64 | 3.07 | 1.71 | | *0.23 | -0.11 | -0.15 | -0.13 | 0.23 |
| 4 | 193232 | 0.887 | 0.04 | 0.02 | 2.37 | 3.19 | 5.71 | *88.68 | | -0.15 | -0.23 | -0.20 | *0.35 | 0.35 |
| 5 | 193247 | 0.776 | 0.04 | 0.01 | 10.44 | 6.14 | 5.80 | *77.56 | | -0.17 | -0.23 | -0.22 | *0.39 | 0.39 |
| 6 | 193206 | 0.571 | 0.05 | 0.03 | *57.09 | 6.53 | 30.81 | 5.50 | | *0.42 | -0.21 | -0.20 | -0.28 | 0.42 |
| 7 | 193181 | 0.747 | 0.08 | 0.01 | *74.64 | 14.62 | 4.72 | 5.92 | | *0.36 | -0.12 | -0.25 | -0.25 | 0.36 |
| 8 | 193199 | 0.849 | 0.06 | 0.02 | 6.20 | 5.32 | 3.56 | *84.84 | | -0.27 | -0.28 | -0.23 | *0.48 | 0.48 |
| 9 | 193190 | 0.654 | 0.06 | 0.02 | *65.38 | 19.75 | 10.86 | 3.93 | | *0.25 | -0.03 | -0.23 | -0.19 | 0.25 |
| 10 | 193114 | 0.414 | 0.10 | 0.02 | 21.30 | *41.35 | 20.22 | 17.00 | | -0.17 | *0.25 | -0.07 | -0.08 | 0.25 |
| 11 | 193164 | 0.797 | 0.08 | 0.02 | 1.14 | 5.24 | 13.88 | *79.64 | | -0.15 | -0.24 | -0.14 | *0.30 | 0.30 |
| 12 | 193160 | 0.817 | 0.08 | 0.02 | 10.09 | *81.64 | 2.73 | 5.43 | | -0.24 | *0.36 | -0.17 | -0.17 | 0.36 |
| 13 | 193109 | 0.577 | 0.11 | 0.02 | *57.68 | 8.60 | 18.57 | 15.03 | | *0.39 | -0.12 | -0.18 | -0.25 | 0.39 |
| 14 | 193169 | 0.686 | 0.08 | 0.01 | 23.47 | *68.53 | 5.90 | 2.01 | | 0.00 | *0.16 | -0.21 | -0.18 | 0.16 |
| 15 | 193143 | 0.646 | 0.10 | 0.01 | 16.53 | 11.02 | *64.56 | 7.78 | | -0.17 | -0.18 | *0.34 | -0.17 | 0.34 |
| 16 | 192641 | 0.659 | 0.34 | 0.03 | 3.86 | 12.42 | 17.71 | *65.63 | | -0.20 | -0.19 | -0.28 | *0.44 | 0.44 |
| 17 | 192675 | 0.633 | 0.33 | 0.02 | 11.47 | *63.11 | 15.62 | 9.45 | | -0.23 | *0.42 | -0.18 | -0.21 | 0.42 |
| 18 | 192609 | 0.557 | 0.36 | 0.02 | 23.46 | 16.26 | *55.44 | 4.45 | | -0.19 | -0.22 | *0.40 | -0.18 | 0.40 |
| 19 | 192496 | 0.731 | 0.41 | 0.03 | *72.76 | 3.68 | 12.52 | 10.59 | | *0.44 | -0.22 | -0.28 | -0.20 | 0.44 |
| 20 | 192362 | 0.544 | 0.46 | 0.05 | 10.48 | 8.69 | 26.24 | *54.09 | | -0.21 | -0.18 | -0.13 | *0.35 | 0.35 |
| 21 | 192110 | 0.489 | 0.62 | 0.03 | 17.89 | *48.62 | 11.30 | 21.55 | | -0.16 | *0.33 | -0.23 | -0.07 | 0.33 |
| 22 | 192256 | 0.704 | 0.54 | 0.03 | 12.81 | *70.03 | 8.49 | 8.09 | | -0.11 | *0.26 | -0.07 | -0.23 | 0.26 |
| 23 | 191870 | 0.802 | 0.74 | 0.02 | 5.74 | *79.55 | 11.13 | 2.81 | | -0.15 | *0.32 | -0.21 | -0.17 | 0.32 |
| 24 | 191682 | 0.763 | 0.84 | 0.03 | *75.60 | 6.49 | 8.53 | 8.51 | | *0.38 | -0.21 | -0.24 | -0.15 | 0.38 |
| 25 | 191516 | 0.591 | 0.93 | 0.02 | 14.85 | *58.50 | 21.41 | 4.29 | | -0.14 | *0.26 | -0.09 | -0.19 | 0.26 |
| 26 | 191412 | 0.740 | 0.99 | 0.02 | 16.12 | *73.31 | 3.68 | 5.89 | | -0.27 | *0.46 | -0.22 | -0.27 | 0.46 |
| 27 | 193242 | 0.697 | 0.06 | 0.41 | 3.89 | 13.04 | 29.45 | 35.80 | 17.36 | | | | | |
| 28 | 193156 | 0.733 | 0.10 | 0.97 | 14.32 | 48.49 | 36.12 | | | | | | | |
| 29 | 193354 | 0.665 | 0.00 | 0.57 | 6.13 | 16.81 | 29.73 | 30.44 | 16.32 | | | | | |

Table 9e. 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 | 199947 | 0.728 | 0.12 | 0.01 | 16.58 | 7.87 | 2.74 | *72.68 | -0.22 | -0.19 | -0.16 | *0.36 | 0.36 |
| 2 | 200127 | 0.907 | 0.03 | 0.01 | 1.64 | 3.65 | *90.65 | 4.03 | -0.21 | -0.33 | *0.46 | -0.23 | 0.46 |
| 3 | 200093 | 0.743 | 0.05 | 0.01 | 9.83 | 2.73 | *74.27 | 13.11 | -0.24 | -0.22 | *0.45 | -0.26 | 0.45 |
| 4 | 199922 | 0.532 | 0.13 | 0.02 | *53.17 | 12.49 | 26.40 | 7.80 | *0.37 | -0.23 | -0.14 | -0.18 | 0.37 |
| 5 | 200053 | 0.819 | 0.06 | 0.02 | *81.79 | 2.55 | 6.63 | 8.95 | *0.32 | -0.13 | -0.10 | -0.27 | 0.32 |
| 6 | 199825 | 0.716 | 0.17 | 0.02 | 2.75 | 8.15 | 17.48 | *71.43 | -0.19 | -0.08 | -0.16 | *0.26 | 0.26 |
| 7 | 200062 | 0.751 | 0.06 | 0.01 | 1.19 | *75.01 | 4.48 | 19.25 | -0.18 | *0.44 | -0.23 | -0.31 | 0.44 |
| 8 | 199944 | 0.655 | 0.11 | 0.02 | *65.38 | 5.58 | 14.55 | 14.36 | *0.48 | -0.26 | -0.22 | -0.26 | 0.48 |
| 9 | 199819 | 0.532 | 0.18 | 0.01 | 18.86 | *53.10 | 24.25 | 3.60 | -0.10 | *0.44 | -0.33 | -0.21 | 0.44 |
| 10 | 199878 | 0.739 | 0.15 | 0.01 | 7.47 | *73.77 | 7.57 | 11.02 | -0.18 | *0.34 | -0.14 | -0.21 | 0.34 |
| 11 | 200045 | 0.907 | 0.07 | 0.01 | *90.59 | 5.81 | 2.36 | 1.16 | *0.35 | -0.27 | -0.16 | -0.14 | 0.35 |
| 12 | 199938 | 0.861 | 0.12 | 0.02 | *85.98 | 3.12 | 6.74 | 4.03 | *0.38 | -0.20 | -0.19 | -0.24 | 0.38 |
| 13 | 199876 | 0.710 | 0.15 | 0.02 | 2.69 | 21.12 | *70.93 | 5.10 | -0.19 | -0.25 | *0.42 | -0.25 | 0.42 |
| 14 | 199998 | 0.895 | 0.10 | 0.01 | 7.46 | *89.38 | 1.81 | 1.25 | -0.09 | *0.23 | -0.20 | -0.17 | 0.23 |
| 15 | 199699 | 0.677 | 0.24 | 0.01 | 7.95 | 3.64 | *67.51 | 20.64 | -0.22 | -0.21 | *0.43 | -0.25 | 0.43 |
| 16 | 199750 | 0.760 | 0.22 | 0.01 | *75.82 | 8.00 | 11.36 | 4.59 | *0.53 | -0.30 | -0.28 | -0.27 | 0.53 |
| 17 | 195488 | 0.702 | 2.36 | 13.83 | 30.49 | 53.32 | | | | | | | |
| 18 | 199247 | 0.597 | 0.46 | 0.02 | *59.45 | 10.13 | 26.66 | 3.28 | *0.30 | -0.21 | -0.13 | -0.17 | 0.30 |
| 19 | 199206 | 0.781 | 0.48 | 0.02 | 5.93 | 10.95 | *77.74 | 4.87 | -0.29 | -0.26 | *0.49 | -0.23 | 0.49 |
| 20 | 199008 | 0.820 | 0.57 | 0.03 | 1.61 | 13.10 | *81.47 | 3.22 | -0.22 | -0.18 | *0.32 | -0.19 | 0.32 |
| 21 | 198692 | 0.599 | 0.72 | 0.03 | 22.25 | *59.45 | 6.74 | 10.80 | -0.12 | *0.25 | -0.11 | -0.14 | 0.25 |
| 22 | 198489 | 0.644 | 0.84 | 0.02 | 19.13 | *63.80 | 8.89 | 7.32 | -0.09 | *0.22 | -0.15 | -0.11 | 0.22 |
| 23 | 195332 | 0.651 | 2.44 | 12.79 | 42.56 | 42.22 | | | | | | | |
| 24 | 196605 | 0.738 | 1.77 | 0.03 | 12.14 | 8.47 | 5.13 | *72.47 | -0.20 | -0.28 | -0.22 | *0.44 | 0.44 |
| 25 | 195614 | 0.578 | 2.28 | 0.02 | 12.46 | 17.62 | *56.43 | 11.19 | -0.16 | -0.19 | *0.30 | -0.07 | 0.30 |
| 26 | 195403 | 0.808 | 2.38 | 0.02 | 3.73 | 12.43 | *78.86 | 2.58 | -0.24 | -0.25 | *0.41 | -0.20 | 0.41 |
| 27 | 194813 | 0.821 | 2.65 | 0.05 | 3.63 | 9.17 | 4.61 | *79.89 | -0.25 | -0.22 | -0.23 | *0.42 | 0.42 |
| 28 | 194687 | 0.528 | 2.74 | 0.02 | 2.82 | *51.37 | 3.78 | 39.27 | -0.20 | *0.25 | -0.22 | -0.10 | 0.25 |
| 29 | 199957 | 0.885 | 0.12 | 0.00 | 2.46 | *88.42 | 6.20 | 2.80 | -0.18 | *0.35 | -0.19 | -0.24 | 0.35 |
| 30 | 199921 | 0.846 | 0.14 | 0.01 | *84.48 | 9.62 | 1.93 | 3.82 | *0.16 | -0.07 | -0.07 | -0.15 | 0.16 |
| 31 | 199959 | 0.935 | 0.12 | 0.00 | 0.51 | 2.77 | *93.41 | 3.18 | -0.12 | -0.19 | *0.29 | -0.18 | 0.29 |
| 32 | 199159 | 0.718 | 0.52 | 7.94 | 40.18 | 51.35 | | | | | | | |
| 33 | 199259 | 0.816 | 0.47 | 4.51 | 27.58 | 67.43 | | | | | | | |
| 34 | 199414 | 0.876 | 0.39 | 0.00 | 1.33 | *87.23 | 6.90 | 4.14 | -0.19 | *0.44 | -0.28 | -0.26 | 0.44 |
| 35 | 199259 | 0.298 | 0.47 | 45.94 | 25.76 | 20.27 | 7.56 | | | | | | |

Table 9f. 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 | % at 5 | Pbis Option 1 | Pbis Option 2 | Pbis Option 3 | Pbis Option 4 | Pbis Key |
|------|---------|---------|--------|--------|------------------|------------------|------------------|------------------|--------|---------------|---------------|---------------|---------------|----------|
| 1 | 202298 | 0.896 | 0.06 | 0.00 | *89.54 | 4.04 | 1.50 | 4.85 | | *0.33 | -0.21 | -0.16 | -0.18 | 0.33 |
| 2 | 202255 | 0.622 | 0.07 | 0.01 | 2.66 | *62.10 | 22.85 | 12.30 | | -0.11 | *0.23 | -0.17 | -0.06 | 0.23 |
| 3 | 202299 | 0.718 | 0.05 | 0.01 | 0.59 | 1.58 | *71.76 | 26.02 | | -0.14 | -0.21 | *0.27 | -0.19 | 0.27 |
| 4 | 202119 | 0.506 | 0.14 | 0.01 | 12.15 | 10.09 | *50.55 | 27.06 | | -0.15 | -0.22 | *0.23 | 0.00 | 0.23 |
| 5 | 202328 | 0.882 | 0.04 | 0.00 | 2.30 | *88.21 | 8.11 | 1.33 | | -0.20 | *0.34 | -0.21 | -0.20 | 0.34 |
| 6 | 202176 | 0.888 | 0.11 | 0.01 | 3.12 | 5.38 | 2.66 | *88.72 | | -0.24 | -0.23 | -0.21 | *0.40 | 0.40 |
| 7 | 202262 | 0.886 | 0.07 | 0.01 | 2.68 | 5.99 | *88.53 | 2.72 | | -0.26 | -0.26 | *0.44 | -0.23 | 0.44 |
| 8 | 202281 | 0.897 | 0.06 | 0.01 | 4.86 | 3.98 | *89.67 | 1.42 | | -0.26 | -0.26 | *0.42 | -0.18 | 0.42 |
| 9 | 202117 | 0.509 | 0.14 | 0.01 | *50.79 | 5.14 | 17.62 | 26.30 | | *0.34 | -0.17 | -0.16 | -0.16 | 0.34 |
| 10 | 202132 | 0.357 | 0.13 | 0.02 | 28.00 | 19.27 | *35.61 | 16.97 | | 0.01 | -0.15 | *0.34 | -0.28 | 0.34 |
| 11 | 202242 | 0.844 | 0.08 | 0.01 | 4.23 | *84.37 | 6.49 | 4.83 | | -0.20 | *0.40 | -0.20 | -0.25 | 0.40 |
| 12 | 202102 | 0.501 | 0.14 | 0.02 | 6.00 | 27.20 | *50.01 | 16.63 | | -0.24 | 0.02 | *0.20 | -0.13 | 0.20 |
| 13 | 202225 | 0.853 | 0.09 | 0.01 | *85.21 | 2.76 | 7.22 | 4.71 | | *0.41 | -0.23 | -0.23 | -0.22 | 0.41 |
| 14 | 202236 | 0.869 | 0.08 | 0.01 | *86.82 | 3.57 | 2.93 | 6.60 | | *0.38 | -0.25 | -0.23 | -0.18 | 0.38 |
| 15 | 202159 | 0.595 | 0.12 | 0.01 | *59.42 | 13.27 | 10.72 | 16.46 | | *0.25 | -0.11 | -0.18 | -0.08 | 0.25 |
| 16 | 202194 | 0.796 | 0.10 | 0.01 | 2.24 | 7.79 | *79.51 | 10.35 | | -0.2 | -0.22 | *0.29 | -0.09 | 0.29 |
| 17 | 202082 | 0.665 | 0.16 | 0.01 | 6.09 | 22.36 | *66.35 | 5.03 | | -0.21 | -0.15 | *0.36 | -0.25 | 0.36 |
| 18 | 202027 | 0.550 | 0.18 | 0.02 | 25.72 | 9.09 | 10.11 | *54.88 | | -0.19 | -0.15 | -0.07 | *0.30 | 0.30 |
| 19 | 202115 | 0.874 | 0.14 | 0.01 | 2.42 | 7.43 | *87.27 | 2.72 | | -0.24 | -0.19 | *0.36 | -0.20 | 0.36 |
| 20 | 202047 | 0.865 | 0.18 | 0.01 | 6.33 | *86.34 | 2.78 | 4.37 | | -0.24 | *0.45 | -0.23 | -0.29 | 0.45 |
| 21 | 201639 | 0.432 | 0.36 | 0.02 | *43.08 | 22.49 | 13.21 | 20.84 | | *0.17 | 0.10 | -0.17 | -0.16 | 0.17 |
| 22 | 201550 | 0.692 | 0.41 | 0.02 | *68.90 | 9.13 | 13.59 | 7.95 | | *0.51 | -0.23 | -0.35 | -0.19 | 0.51 |
| 23 | 201247 | 0.578 | 0.56 | 0.02 | *57.42 | 14.68 | 20.64 | 6.68 | | *0.33 | -0.24 | -0.12 | -0.12 | 0.33 |
| 24 | 201383 | 0.754 | 0.50 | 0.02 | 13.40 | *75.04 | 4.44 | 6.60 | | -0.22 | *0.43 | -0.22 | -0.25 | 0.43 |
| 25 | 201084 | 0.841 | 0.61 | 0.05 | 3.55 | 4.33 | 7.97 | *83.50 | | -0.20 | -0.25 | -0.20 | *0.39 | 0.39 |
| 26 | 201025 | 0.659 | 0.68 | 0.01 | 13.45 | *65.42 | 18.43 | 2.01 | | -0.15 | *0.32 | -0.18 | -0.20 | 0.32 |
| 27 | 202179 | 0.730 | 0.12 | 0.34 | 3.00 | 11.00 | 26.67 | 34.97 | 23.91 | | | | | |
| 28 | 202199 | 0.770 | 0.11 | 0.78 | 10.41 | 45.73 | 42.97 | | | | | | | |
| 29 | 202424 | 0.714 | 0.00 | 0.35 | 3.73 | 13.54 | 26.66 | 32.25 | 23.48 | | | | | |

Point-Biserial Correlation Coefficients

Point biserial statistics are used to examine item-test correlations or item discrimination. In the Tables 9a-9f, point biserial correlation coefficients were computed for each answer option. Point biseri-als for the correct answer option are denoted with an asterisk (*) and are repeated in the 'Pbis Key' field. The point biserial correlation is a measure of internal consistency that ranges between +/-1. It indicates a correlation of students' responses to an item relative to their performance on the rest of the test. The criterion for point biserial for the correct answer option used for New York State test was 0.15. The point biseri-als for the correct answer option that are equal to or greater than 0.15 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. Grades 3, 4, 6, 7 and 8 did not have any item answer keys flagged for point biseri-als. Grade 5 had two flagged for point biserial values of the answer keys being less than 0.15: item 2 (0.14) and item 16 (0.10). Point biseri-als for correct answer options (pbis*) on the tests ranged from 0.10 to 0.56. For grade 3, the pbis* were between 0.24 and 0.56. For grade 4, the

pbis* were between 0.22 and 0.50. For grade 5, the pbis* were between 0.10 and 0.47. For grade 6, pbis* were between 0.16 and 0.48. For grade 7, the pbis* were between 0.16 and 0.53. For grade 8, the pbis* were between 0.17 and 0.51.

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 9a to 9f 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 ELA items except item 3 on grade 5 ELA test. Approximately 24% of students answered this item correctly while close to 61% of student 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. Two grade 5 items were flagged for positive point biserial values on distracter (incorrect) answer options (items 2 and 16). Only one test item was flagged for point biserials of both a distracter and the answer option, grade 5 item 16; however, grade 5 item 16 had very similar data and flags (for p-value and point biserials) from the 2005 field test administration. Grade 8 items 10, 12, and 21 were flagged for positive point biserial of a distracter answer option (0.01, 0.02, and 0.10, respectively). It should be noted that none of the point biserials of distracter options on any 2006 NYSTP ELA test exceeded the point biserial of the corresponding answer key option. There were no flags for point biserials of distracters in grades 3, 4, 6, and 7.

Test Statistics and Reliability Coefficients

Test statistics including raw score mean and standard deviation are presented in Table 10, below. Please note, that for grades 4 and 8 both weighted and unweighted test statistics are provided. Grade 4 and 8 CR items were weighted by a 1.38 factor to increase proportion of score points obtainable from these items. Weighting CR items on these two grades resulted in better alignment of proportions of test raw score points obtainable from MC and CR items between 2005 and 2006 ELA operational tests for these grades. More information on weighting CR items and the effect on test content is provided in Section VI (IRT Scaling). Reliability coefficients provide measures of internal consistency that range from zero to one. Two reliability coefficients: Cronbach alpha and Feldt-Raju were computed for the Grades 3-8 ELA Tests. Both types of reliability estimates are appropriate to use when a test contains both MC and CR items. Calculated Cronbach reliabilities ranged from 0.81 to 0.88. Feldt-Raju reliability coefficients ranged from 0.82 to 0.89. The lowest reliability was observed for the grade 5 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 4 test. All reliabilities met or exceeded 0.80, across statistics, which is a good indication that the NYSTP 3-8 ELA Tests are acceptably reliable. High reliability indicates that scores are consistent and not unduly influenced by random error. More information on test reliability and standard error of measurement is provided in Section VIII (Reliability) of this report.

Table 10. NYSTP ELA 2006 Test Form Statistics and Reliability

| Grade | Max RS | RS Mean | RS SD | P-value Mean | Cronbach Alpha | Feldt-Raju Alpha |
|-------|----------------|----------------------|--------------------|--------------|----------------|------------------|
| 3 | 33 | 25.19 | 5.75 | 0.76 | 0.85 | 0.86 |
| 4 | 39 (43 WGT) | 27.83 (30.65 WGT) | 6.99 (7.57 WGT) | 0.71 | 0.88 | 0.89 |
| 5 | 31 | 21.67 | 5.40 | 0.70 | 0.81 | 0.82 |
| 6 | 39 | 27.01 | 6.63 | 0.69 | 0.85 | 0.86 |
| 7 | 41 | 28.84 | 7.20 | 0.70 | 0.87 | 0.88 |
| 8 | 39 (44 WGT) | 28.01 (31.64 WGT) | 6.38 (7.24 WGT) | 0.72 | 0.84 | 0.86 |

Note: WGT=weighted results

Speededness

Speededness is the term used to refer to interference in test score observation due to insufficient testing time. Test developers considered speededness in the development of the NYSTP tests. NYSED believes that achievement tests should not be speeded; little or no useful instructional information can be obtained from the fact that a student 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 9a-9f show the omit rates for items on the Grades 3-8 ELA Tests. These results provide no evidence of speededness on these tests.

Differential Item Functioning

Classical Differential Item Functioning (DIF) was evaluated using two methods. First, the standardized mean difference (SMD) was computed for all items. The SMD statistic (Dorans, Schmitt & Bleistein, 1992) compares the mean scores of reference and focal groups, after adjusting for ability differences. A moderate amount of significant DIF, for or against the focal group, is represented by an SMD with an absolute value between .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-Haenszel 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), and ethnicity (focal groups: African-American, Hispanic, and Asian; reference group: White). 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 11 shows the number of cases for subgroups.

Table 11. NYSTP ELA 2006 Classical DIF Sample N-Counts

| Grade | Ethnicity | | | | Gender | | Need Resource Category | |
|-------|-----------|----------|-------|-------|--------|------|------------------------|------|
| | Black | Hispanic | Asian | White | Female | Male | High | Low |
| 3 | 1521 | 1111 | 500 | 3912 | 3522 | 3522 | 3669 | 3300 |
| 4 | 1415 | 1092 | 500 | 3998 | 3441 | 3564 | 3495 | 3458 |
| 5 | 1332 | 1183 | 500 | 4009 | 3521 | 3503 | 3518 | 3402 |
| 6 | 1398 | 1165 | 500 | 3957 | 3444 | 3576 | 3524 | 3412 |
| 7 | 1446 | 1218 | 500 | 3923 | 3514 | 3573 | 3570 | 3440 |
| 8 | 1441 | 1082 | 500 | 4062 | 3458 | 3627 | 3486 | 3525 |

Table 12 presents the number of items flagged for DIF by either of the classical methods described earlier. It should be noted that items showing statistically significant DIF do not necessarily pose bias. In addition to item bias, DIF may be attributed to item impact or type one error. All items that were flagged for significant DIF were carefully examined by multiple reviewers during operational item selection for possible item bias. Only those items that were determined free of bias were included in the operational tests.

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

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

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

Section VI: IRT Scaling

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 ELA tests. It should be noted that the scaling was done on nearly the total New York State population of students in public schools and exclusion of some cases during the data cleaning had very minimal or no effect on parameter estimation.

Calibration Process

The IRT model parameters were estimated using CTB/McGraw-Hill's PARDUX software (Burket, 2002). PARDUX estimates parameters simultaneously for MC and CR items using marginal maximum likelihood procedures implemented via the 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 ELA 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 ELA 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 ELA Tests, all calibration estimation results are reasonable.

Table 13. NYSTP ELA 2006 Calibration Results

| Grade | Largest 'a' parameter | 'b' parameter range | | # items with Default 'c' | Theta Mean | Theta Standard Deviation | N students |
|-------|-----------------------|---------------------|-------|--------------------------|------------|--------------------------|------------|
| 3 | 2.342 | -4.195 | 0.511 | 1 | -0.01 | 1.292 | 179552 |
| 4 | 2.310 | -3.558 | 0.709 | 3 | -0.10 | 1.191 | 183342 |
| 5 | 2.007 | -3.135 | 1.744 | 4 | -0.01 | 1.260 | 191161 |
| 6 | 2.007 | -3.675 | 1.162 | 4 | -0.09 | 1.181 | 193354 |
| 7 | 2.193 | -3.899 | 0.361 | 5 | -0.05 | 1.182 | 200208 |
| 8 | 2.179 | -3.121 | 1.261 | 3 | -0.01 | 1.188 | 202424 |

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_i was transformed to Z_{Q_i} where

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

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 $Z_{Q_i}Crit$ used to flag items was calculated using the expression

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

where N is the calibration sample size.

Items were considered to have poor fit if the value of obtained Z_{Q_i} was greater than the value of Z_{Q_i} critical. If the obtained Z_{Q_i} was less than Z_{Q_i} critical the items were rated as having acceptable fit. It should be noted that all items in the NYSTP 2006 ELA test demonstrated good model fit, further supporting use of the chosen models. No items exhibited poor item-model fit statistics. These statistics are presented in Appendix F.

Local Independence

In using IRT models, one of the assumptions made is that the items are locally independent. That is, a student's response on one item is not dependent upon 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 ELA tests and only one pair of items was found to be locally dependent. Grade 4 items 26 and 27 (both MC items from the same passage) were found to be locally dependent ($Q_3 = 0.297$). The magnitude of this statistic was not sufficient to warrant concern.

Scaling

The scaling of the Grades 3-8 ELA 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 ($M1$ and $M2$) were used to transform the New York State ELA item parameters in the original theta metric estimated during the item calibration process to the scale score metric. These constants are presented in Table 14.

Table 14. NYSTP ELA 2006 Initial Transformation Constants

| Grade | $M1$ | $M2$ |
|-------|------|------|
| 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 'Meeting 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 14)

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 15 presents the final transformation parameters for New York State Grades 3-8 ELA Tests.

Table 15. NYSTP ELA 2006 Final Transformation Constants

| Grade | $M1_{new}$ | $M2_{new}$ |
|-------|------------|------------|
| 3 | 32.0147 | 671.3431 |
| 4 | 33.6458 | 671.3090 |
| 5 | 30.8209 | 665.4105 |
| 6 | 33.5152 | 662.2889 |
| 7 | 32.8933 | 656.5787 |
| 8 | 32.7136 | 652.1809 |

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 15) and the following procedure:

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

$$Cut_{\theta} = (Cut_{old} - M2_{old}) / M1_{old}$$

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.**

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

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 ELA 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 16a-16f. Descriptions of what each of the parameter variables mean is presented in the subsection depicting the IRT models and rationale.

Table 16a. Grade 3 2006 Operational Item Parameter Estimates

| Item | Max Pts | a par/ alpha | b par/ gamma1 | c par/ gamma2 | gamma3 |
|------|---------|--------------|---------------|---------------|---------|
| 1 | 1 | 0.02308 | 600.5234 | 0.0733 | |
| 2 | 1 | 0.02668 | 578.8792 | 0.0733 | |
| 3 | 1 | 0.02408 | 611.2498 | 0.0213 | |
| 4 | 1 | 0.03301 | 609.8125 | 0.0549 | |
| 5 | 1 | 0.02325 | 608.7785 | 0.0733 | |
| 6 | 1 | 0.01723 | 645.6807 | 0.0131 | |
| 7 | 1 | 0.02439 | 617.3813 | 0.0421 | |
| 8 | 1 | 0.01781 | 628.4343 | 0.0733 | |
| 9 | 1 | 0.02514 | 626.8038 | 0.0733 | |
| 10 | 1 | 0.02397 | 590.4236 | 0.0733 | |
| 11 | 1 | 0.03354 | 613.3588 | 0.0268 | |
| 12 | 1 | 0.02781 | 658.8746 | 0.1338 | |
| 13 | 1 | 0.02339 | 593.6426 | 0.0733 | |
| 14 | 1 | 0.01845 | 687.6444 | 0.2029 | |
| 15 | 1 | 0.02275 | 660.4994 | 0.1417 | |
| 16 | 1 | 0.02879 | 622.9772 | 0.0751 | |
| 17 | 1 | 0.02350 | 666.8063 | 0.1911 | |
| 18 | 2 | 0.06156 | 37.2814 | 37.2298 | |
| 19 | 1 | 0.04303 | 630.1678 | 0.0830 | |
| 20 | 1 | 0.02099 | 669.1681 | 0.0811 | |
| 21 | 1 | 0.01544 | 646.2624 | 0.0733 | |
| 22 | 1 | 0.02606 | 603.7823 | 0.2988 | |
| 23 | 1 | 0.01575 | 601.9543 | 0.0733 | |
| 24 | 1 | 0.01353 | 589.8869 | 0.2000 | |
| 25 | 2 | 0.02161 | 13.4245 | 14.8004 | |
| 26 | 2 | 0.02435 | 15.3493 | 16.5776 | |
| 27 | 1 | 0.02527 | 655.6339 | 0.2034 | |
| 28 | 3 | 0.02086 | 13.2024 | 12.9605 | 12.7981 |

Table 16b. Grade 4 2006 Operational Item Parameter Estimates

| Item | Max Pts | a par/ alpha | b par/ gamma1 | c par/ gamma2 | gamma3 | gamma4 |
|------|---------|-----------------|------------------|------------------|---------|---------|
| 1 | 1 | 0.02348 | 636.0706 | 0.1389 | | |
| 2 | 1 | 0.01411 | 635.8158 | 0.1389 | | |
| 3 | 1 | 0.03057 | 628.8612 | 0.0979 | | |
| 4 | 1 | 0.01769 | 630.1777 | 0.0288 | | |
| 5 | 1 | 0.02146 | 587.3539 | 0.1389 | | |
| 6 | 1 | 0.02491 | 651.1445 | 0.1875 | | |
| 7 | 1 | 0.01910 | 625.2031 | 0.1389 | | |
| 8 | 1 | 0.00974 | 615.5748 | 0.2000 | | |
| 9 | 1 | 0.01161 | 627.6515 | 0.2000 | | |
| 10 | 1 | 0.01376 | 604.9571 | 0.2000 | | |
| 11 | 1 | 0.02168 | 622.1011 | 0.1099 | | |
| 12 | 1 | 0.03370 | 625.9250 | 0.1700 | | |
| 13 | 1 | 0.01694 | 671.9748 | 0.1997 | | |
| 14 | 1 | 0.02385 | 623.6915 | 0.0922 | | |
| 15 | 1 | 0.01968 | 647.8500 | 0.2192 | | |
| 16 | 1 | 0.01818 | 634.4384 | 0.1389 | | |
| 17 | 1 | 0.03319 | 608.2625 | 0.1243 | | |
| 18 | 1 | 0.04039 | 650.7969 | 0.2843 | | |
| 19 | 1 | 0.02274 | 671.7303 | 0.3342 | | |
| 20 | 1 | 0.03629 | 674.0767 | 0.1442 | | |
| 21 | 1 | 0.02418 | 658.1336 | 0.1501 | | |
| 22 | 1 | 0.02032 | 649.6003 | 0.2143 | | |
| 23 | 1 | 0.01894 | 638.9962 | 0.1063 | | |
| 24 | 1 | 0.02287 | 689.5362 | 0.1623 | | |
| 25 | 1 | 0.02202 | 655.8521 | 0.1013 | | |
| 26 | 1 | 0.03602 | 651.0165 | 0.2512 | | |
| 27 | 1 | 0.02521 | 620.6647 | 0.1499 | | |
| 28 | 1 | 0.02860 | 647.8360 | 0.1810 | | |
| 29 | 4 | 0.03553 | 19.0698 | 21.5644 | 23.6277 | 25.6934 |
| 30 | 3 | 0.03611 | 19.4681 | 22.1635 | 24.7132 | |
| 31 | 4 | 0.04055 | 21.7283 | 24.1422 | 26.4981 | 28.6991 |

Table 16c. Grade 5 2006 Operational Item Parameter Estimates

| Item | Max Pts | a par/ alpha | b par/ gamma1 | c par/ gamma2 | gamma3 |
|------|---------|-----------------|------------------|------------------|---------|
| 1 | 1 | 0.02667 | 630.5118 | 0.3258 | |
| 2 | 1 | 0.00907 | 546.2111 | 0.2000 | |
| 3 | 1 | 0.02966 | 699.9956 | 0.0344 | |
| 4 | 1 | 0.02529 | 606.9121 | 0.1996 | |
| 5 | 1 | 0.01108 | 663.6760 | 0.0159 | |
| 6 | 1 | 0.02195 | 627.9653 | 0.2665 | |
| 7 | 1 | 0.03369 | 610.6783 | 0.1472 | |
| 8 | 1 | 0.02013 | 641.6817 | 0.1749 | |
| 9 | 1 | 0.02767 | 663.2133 | 0.1617 | |
| 10 | 1 | 0.01100 | 625.0060 | 0.2000 | |
| 11 | 1 | 0.02808 | 607.3604 | 0.1996 | |
| 12 | 2 | 0.02714 | 16.7028 | 17.7124 | |
| 13 | 1 | 0.02247 | 664.8525 | 0.4365 | |
| 14 | 1 | 0.02453 | 623.3248 | 0.1378 | |
| 15 | 1 | 0.01472 | 613.4265 | 0.2000 | |
| 16 | 1 | 0.00614 | 747.6031 | 0.2000 | |
| 17 | 1 | 0.01264 | 677.9700 | 0.2958 | |
| 18 | 1 | 0.03830 | 636.1809 | 0.1809 | |
| 19 | 1 | 0.02596 | 640.4509 | 0.2253 | |
| 20 | 1 | 0.03165 | 629.2469 | 0.0957 | |
| 21 | 1 | 0.02683 | 668.2941 | 0.2772 | |
| 22 | 1 | 0.02236 | 681.2870 | 0.2790 | |
| 23 | 1 | 0.03729 | 623.4698 | 0.1416 | |
| 24 | 1 | 0.02880 | 671.4012 | 0.2540 | |
| 25 | 1 | 0.03466 | 641.9619 | 0.3558 | |
| 26 | 2 | 0.03528 | 21.5357 | 23.1782 | |
| 27 | 3 | 0.02660 | 16.4777 | 17.0892 | 18.0329 |

Table 16d. Grade 6 2006 Operational Item Parameter Estimates

| Item | Max Pts | a par/ alpha | b par/ gamma1 | c par/ gamma2 | gamma3 | gamma4 | gamma5 |
|------|------------|-----------------|------------------|------------------|---------|---------|---------|
| 1 | 1 | 0.02740 | 583.3758 | 0.2000 | | | |
| 2 | 1 | 0.00856 | 659.0174 | 0.0779 | | | |
| 3 | 1 | 0.01844 | 558.9557 | 0.2000 | | | |
| 4 | 1 | 0.02310 | 595.1591 | 0.1376 | | | |
| 5 | 1 | 0.02413 | 630.3820 | 0.2558 | | | |
| 6 | 1 | 0.02746 | 660.5319 | 0.1689 | | | |
| 7 | 1 | 0.01802 | 624.2852 | 0.1559 | | | |
| 8 | 1 | 0.03521 | 616.1484 | 0.1395 | | | |
| 9 | 1 | 0.01159 | 642.7513 | 0.2000 | | | |
| 10 | 1 | 0.01959 | 697.1758 | 0.2052 | | | |
| 11 | 1 | 0.01456 | 599.1006 | 0.0687 | | | |
| 12 | 1 | 0.02287 | 622.8094 | 0.2986 | | | |
| 13 | 1 | 0.01939 | 652.9505 | 0.0893 | | | |
| 14 | 1 | 0.00682 | 619.0731 | 0.2000 | | | |
| 15 | 1 | 0.02631 | 662.4738 | 0.3330 | | | |
| 16 | 1 | 0.02940 | 651.1423 | 0.2181 | | | |
| 17 | 1 | 0.02512 | 651.1233 | 0.1699 | | | |
| 18 | 1 | 0.03138 | 667.3766 | 0.2248 | | | |
| 19 | 1 | 0.02971 | 640.1871 | 0.2271 | | | |
| 20 | 1 | 0.02374 | 670.2170 | 0.2171 | | | |
| 21 | 1 | 0.02501 | 679.6954 | 0.2184 | | | |
| 22 | 1 | 0.01252 | 626.7135 | 0.1559 | | | |
| 23 | 1 | 0.01637 | 603.4102 | 0.0523 | | | |
| 24 | 1 | 0.01888 | 616.4393 | 0.0351 | | | |
| 25 | 1 | 0.01206 | 656.2128 | 0.1559 | | | |
| 26 | 1 | 0.03522 | 642.1313 | 0.2528 | | | |
| 27 | 5 | 0.03484 | 18.2795 | 20.1748 | 21.3625 | 22.7394 | 24.4299 |
| 28 | 3 | 0.04220 | 22.8036 | 25.6597 | 28.4534 | | |
| 29 | 5 | 0.03422 | 17.9877 | 20.1920 | 21.3744 | 22.6206 | 23.9796 |

Table 16e. Grade 7 2006 Operational Item Parameter Estimates

| Item | Max Pts | a par/ alpha | b par/ gamma 1 | c par/ gamma 2 | gamma 3 |
|------|---------|-----------------|-------------------|-------------------|---------|
| 1 | 1 | 0.02087 | 631.8830 | 0.2217 | |
| 2 | 1 | 0.03922 | 598.1007 | 0.0548 | |
| 3 | 1 | 0.03100 | 633.6187 | 0.2202 | |
| 4 | 1 | 0.02044 | 658.4709 | 0.1162 | |
| 5 | 1 | 0.02040 | 617.2150 | 0.3177 | |
| 6 | 1 | 0.01632 | 644.1584 | 0.3546 | |
| 7 | 1 | 0.02465 | 622.9119 | 0.0987 | |
| 8 | 1 | 0.03621 | 646.6580 | 0.1955 | |
| 9 | 1 | 0.02636 | 655.9073 | 0.0867 | |
| 10 | 1 | 0.02115 | 634.5809 | 0.2856 | |
| 11 | 1 | 0.02533 | 590.2272 | 0.1814 | |
| 12 | 1 | 0.02375 | 599.5237 | 0.1184 | |
| 13 | 1 | 0.03024 | 641.8343 | 0.2666 | |
| 14 | 1 | 0.01370 | 563.7889 | 0.2000 | |
| 15 | 1 | 0.02814 | 642.5233 | 0.1988 | |
| 16 | 1 | 0.03846 | 629.0345 | 0.1321 | |
| 17 | 2 | 0.03140 | 19.0672 | 20.0755 | |
| 18 | 1 | 0.01560 | 652.8855 | 0.1814 | |
| 19 | 1 | 0.03127 | 622.3257 | 0.1040 | |
| 20 | 1 | 0.01756 | 605.5117 | 0.2000 | |
| 21 | 1 | 0.01411 | 659.2856 | 0.2318 | |
| 22 | 1 | 0.01094 | 642.4467 | 0.2000 | |
| 23 | 2 | 0.01891 | 10.8976 | 12.4568 | |
| 24 | 1 | 0.03579 | 642.1110 | 0.2952 | |
| 25 | 1 | 0.01684 | 656.8645 | 0.1573 | |
| 26 | 1 | 0.03384 | 634.4266 | 0.3591 | |
| 27 | 1 | 0.02970 | 626.3706 | 0.2836 | |
| 28 | 1 | 0.01356 | 672.2538 | 0.1814 | |
| 29 | 1 | 0.02332 | 593.4963 | 0.1563 | |
| 30 | 1 | 0.00878 | 553.2646 | 0.2000 | |
| 31 | 1 | 0.02298 | 575.2849 | 0.2000 | |
| 32 | 2 | 0.02852 | 16.3555 | 18.4560 | |
| 33 | 2 | 0.02536 | 14.0788 | 15.6164 | |
| 34 | 1 | 0.03377 | 608.1914 | 0.1894 | |
| 35 | 3 | 0.02815 | 18.7871 | 19.0112 | 20.2675 |

Table 16f. Grade 8 2006 Operational Item Parameter Estimates

| Item | Max Pts | a par/ alpha | b par/ gamma1 | c par/ gamma2 | gamma3 | gamma4 | gamma5 |
|------|---------|-----------------|------------------|------------------|---------|---------|---------|
| 1 | 1 | 0.02147 | 582.2722 | 0.1413 | | | |
| 2 | 1 | 0.01451 | 661.3540 | 0.3217 | | | |
| 3 | 1 | 0.01271 | 614.5122 | 0.1532 | | | |
| 4 | 1 | 0.01172 | 676.3866 | 0.2000 | | | |
| 5 | 1 | 0.02095 | 586.7340 | 0.1664 | | | |
| 6 | 1 | 0.03125 | 602.5144 | 0.2720 | | | |
| 7 | 1 | 0.03272 | 598.3033 | 0.1209 | | | |
| 8 | 1 | 0.03135 | 593.6239 | 0.1049 | | | |
| 9 | 1 | 0.02451 | 666.6305 | 0.2077 | | | |
| 10 | 1 | 0.02554 | 681.2286 | 0.1121 | | | |
| 11 | 1 | 0.02318 | 596.0840 | 0.0409 | | | |
| 12 | 1 | 0.01013 | 674.7971 | 0.1622 | | | |
| 13 | 1 | 0.02629 | 601.5170 | 0.1336 | | | |
| 14 | 1 | 0.02675 | 602.3523 | 0.2358 | | | |
| 15 | 1 | 0.01230 | 649.0274 | 0.1785 | | | |
| 16 | 1 | 0.01505 | 601.9833 | 0.2000 | | | |
| 17 | 1 | 0.01917 | 634.5073 | 0.1536 | | | |
| 18 | 1 | 0.01417 | 650.5356 | 0.0996 | | | |
| 19 | 1 | 0.02281 | 592.9301 | 0.1664 | | | |
| 20 | 1 | 0.03348 | 606.0294 | 0.1747 | | | |
| 21 | 1 | 0.00996 | 707.0299 | 0.2000 | | | |
| 22 | 1 | 0.03918 | 636.7089 | 0.1654 | | | |
| 23 | 1 | 0.01966 | 653.7186 | 0.1845 | | | |
| 24 | 1 | 0.02869 | 626.2669 | 0.1930 | | | |
| 25 | 1 | 0.02850 | 615.4030 | 0.3040 | | | |
| 26 | 1 | 0.02071 | 648.0469 | 0.2852 | | | |
| 27 | 5 | 0.04618 | 24.4637 | 26.4693 | 27.9281 | 29.5662 | 31.3135 |
| 28 | 3 | 0.04564 | 24.5060 | 27.0584 | 29.9958 | | |
| 29 | 5 | 0.04225 | 22.1507 | 24.2744 | 25.8170 | 27.1934 | 28.6443 |

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 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

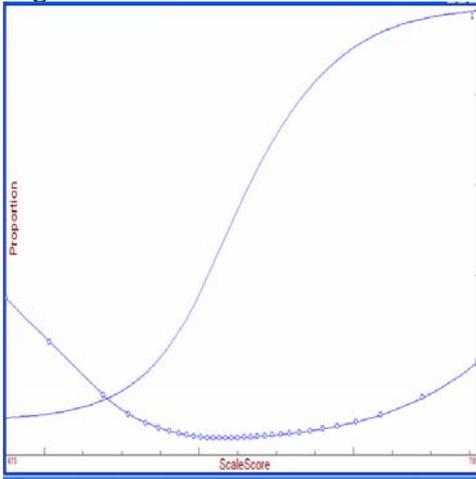


Figure 2. Grade 4 2006 OP TCC and SE

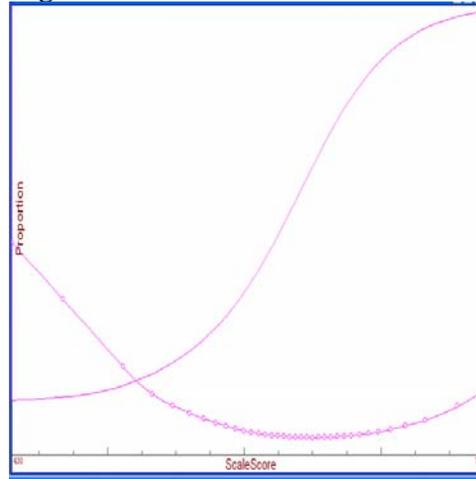


Figure 3. Grade 5 2006 OP TCC and SE

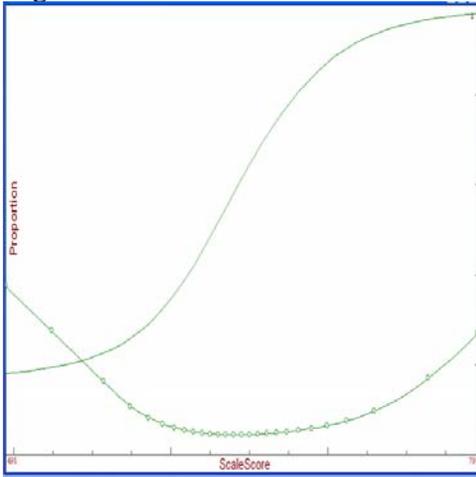


Figure 4. Grade 6 2006 OP TCC and SE

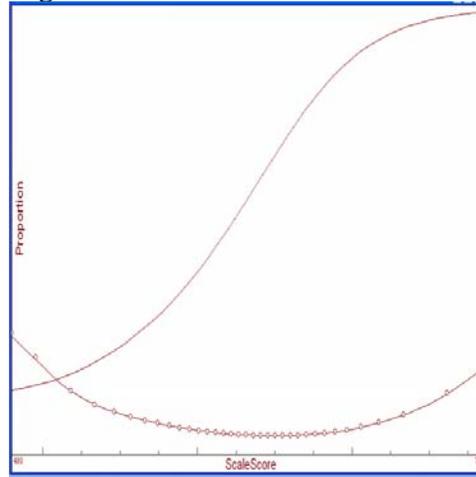


Figure 5. Grade 7 2006 OP TCC and SE

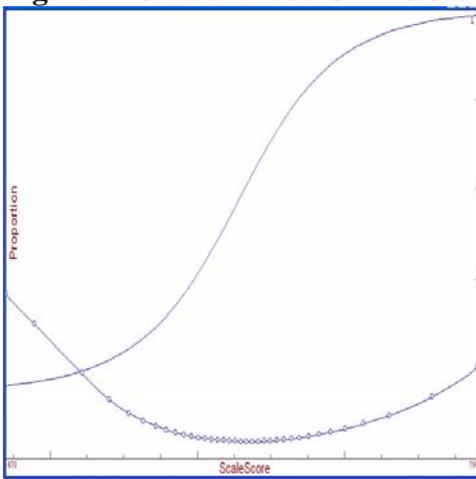
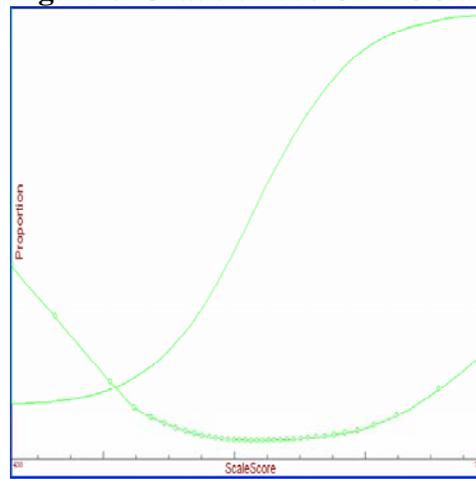


Figure 6. Grade 8 2006 OP TCC and SE



Equating

The Grades 3-8 ELA testing program is considered to be a new family of tests on new scales with new proficiency standards set in 2006. Therefore, no direct equating between years 2005 and 2006 was performed for the grades 4 and 8 assessments. Instead, an equipercentile linking of these assessments was conducted and is described in detail in Section X: Special Studies. It should be noted that there is no history for the grades 3, 5, 6 and 7 ELA state assessments. The new ELA 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 who 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 2006 Grades 3-8 ELA 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 ELA 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 ELA tests have a maximum raw score higher than 30 points. In the inverse TCC method, a student's trait estimate is taken to be the trait value 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.

Weighting Constructed Response Items in Grades 4 and 8

A weight factor of 1.38 was applied to all CR items in grades 4 and 8. The CR items were weighted in order to align proportions of raw score points obtainable from MC and CR items on 2005 and 2006 ELA grade 4 and 8 tests. Aligning proportions of raw score points between the two tests was believed to provide better continuity between the two testing programs. Tables 17 to 20 present number of score points obtainable from MC and CR items on 2005 and 2006 grade 4 and 8 tests. The abbreviations describing CR items are as follows: R/W is Reading/Writing, L/W is Listening/Writing, IW is Independent Writing (2005 only), and WM is Writing Mechanics. Target pts (points) refer to the number of test point obtainable from each standard as specified by the test blueprint. Target pts% (points percent) refers to the proportion of test points obtainable from each content standard as specified in the test blueprint. It is desirable that the target and actual percent of score points obtainable from content standards do not differ by more than 10%. CTB/McGraw-Hill's content specialists' goal is to restrict this difference to 5% or less.

Table 17. ELA grade 4 MC and CR point distribution in 2005 by Learning Standards

| Standard | MC pts | CR (R/W) pts | CR (L/W) pts | CR (IW) pts | CR (WM) pts | Total pts | Target pts | Total pts % | Target pts % |
|----------|--------|--------------|--------------|-------------|-------------|-----------|------------|-------------|--------------|
| 1 | 14 | | | | | 14 | 16 | 33% | 38% |
| 2 | 10 | | 4 | 3 | | 17 | 16 | 40% | 38% |
| 3 | 4 | 4 | | | | 8 | 7 | 19% | 17% |
| | | | | | 3 | 3 | 3 | 7% | 7% |
| Totals | 28 | 4 | 4 | 3 | 3 | 42 | 42 | 100% | 100% |

Table 18. ELA grade 4 MC and CR point distribution in 2006 by Learning Standards

| Standard | MC pts | CR (R/W) pts | CR (L/W) pts | CR (WM) pts | Total pts | Target pts | Total pts % | Target pts % | Total pts WGT* | Total pts % WGT* |
|----------|--------|--------------|--------------|-------------|-----------|------------|-------------|--------------|----------------|------------------|
| 1 | 13 | | | | 13 | 13 | 33% | 33% | 13 | 30% |
| 2 | 11 | | 4 | | 15 | 16 | 38% | 41% | 16.5 | 38% |
| 3 | 4 | 4 | | | 8 | 7 | 21% | 18% | 9.5 | 22% |
| | | | | 3 | 3 | 3 | 8% | 8% | 4 | 9% |
| Totals | 28 | 4 | 4 | 3 | 39 | 39 | 100% | 100% | 43 | 100% |

Note: WGT = weighted results

Table 19. ELA grade 8 MC and CR point distribution in 2005 by Learning Standards

| Standard | MC pts | CR (R/W) pts | CR (L/W) pts | CR (IW) pts | CR (WM) pts | Total pts | Target pts | Total pts % | Target pts % |
|----------|--------|--------------|--------------|-------------|-------------|-----------|------------|-------------|--------------|
| 1 | 11 | | 6 | 3 | | 20 | 25 | 47% | 58% |
| 2 | 9 | | | | | 9 | 6 | 21% | 14% |
| 3 | 5 | 6 | | | | 11 | 9 | 26% | 21% |
| | | | | | 3 | 3 | 3 | 7% | 7% |
| Totals | 25 | 6 | 6 | 3 | 3 | 43 | 43 | 100.0% | 100% |

Table 20. ELA grade 8 MC and CR point distribution in 2006 by Learning Standards

| Standard | MC pts | CR (R/W) pts | CR (L/W) pts | CR (WM) pts | Total pts | Target pts | Total pts % | Target pts % | Total pts WGT* | Total pts % WGT* |
|----------|--------|--------------|--------------|-------------|-----------|------------|-------------|--------------|----------------|------------------|
| 1 | 10 | | 5 | | 15 | 14 | 38% | 36% | 17 | 39% |
| 2 | 13 | | | | 13 | 14 | 33% | 36% | 13 | 30% |
| 3 | 3 | 5 | | | 8 | 8 | 21% | 21% | 10 | 23% |
| | | | | 3 | 3 | 3 | 8% | 8% | 4 | 9% |
| Totals | 26 | 5 | 5 | 3 | 39 | 39 | 100% | 100% | 44 | 100.0 |

Note: WGT = weighted results

Weighted CR items in grades 4 and 8 resulted in better alignment of proportions of MC and CR score points between 2005 and 2006 tests (see Table 21) and had no significant effect on the test blueprint (content). For both of the grades and all learning standards (except standard 2, grade 8), the difference between target percent and actual percent of score points measuring each standard was less than 5%. For standard 2 in grade 8, the difference between target % and actual percent of score points was about 6%, which is still very small.

Table 21. New York State ELA grades 4 and 8 MC and CR proportions in 2005 and 2006

| Year | Grade | MC pts % | CR pts % | MC pts % WGT* | CR pts % WGT* |
|------|-------|----------|----------|------------------|------------------|
| 2005 | 4 | 67% | 33% | | |
| 2006 | 4 | 72% | 28% | 65% | 35% |
| | | | | | |
| 2005 | 8 | 58% | 42% | | |
| 2006 | 8 | 67% | 33% | 59% | 41% |

Note: WGT = weighted results

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

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 22, below.

Table 22. NYSTP ELA 2006 Minimum and Maximum Scale Scores

| Grade | LOSS | HOSS |
|-------|------|------|
| 3 | 475 | 780 |
| 4 | 430 | 775 |
| 5 | 495 | 795 |
| 6 | 480 | 785 |
| 7 | 470 | 790 |
| 8 | 430 | 790 |

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 ELA 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 and further adjustments by NYSED 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 23a. Grade 3 Raw Score to Scale Score (with Standard Error)

| Raw Score | Scale Score | Standard Error |
|-----------|-------------|----------------|
| 0 | 475 | 102 |
| 1 | 475 | 102 |
| 2 | 475 | 102 |
| 3 | 503 | 74 |
| 4 | 538 | 39 |
| 5 | 554 | 26 |
| 6 | 565 | 21 |
| 7 | 574 | 17 |
| 8 | 581 | 15 |
| 9 | 587 | 14 |
| 10 | 592 | 13 |
| 11 | 597 | 12 |
| 12 | 601 | 12 |
| 13 | 605 | 11 |
| 14 | 609 | 11 |
| 15 | 613 | 11 |
| 16 | 617 | 11 |
| 17 | 621 | 11 |
| 18 | 625 | 11 |
| 19 | 629 | 11 |
| 20 | 633 | 12 |
| 21 | 638 | 12 |
| 22 | 642 | 12 |
| 23 | 647 | 13 |
| 24 | 653 | 13 |
| 25 | 659 | 14 |
| 26 | 665 | 15 |
| 27 | 672 | 16 |
| 28 | 680 | 17 |

(Continued on next page)

Table 23a. Grade 3 Raw Score to Scale Score (with Standard Error) (cont.)

| Raw Score | Scale Score | Standard Error |
|-----------|-------------|----------------|
| 29 | 690 | 19 |
| 30 | 701 | 21 |
| 31 | 717 | 26 |
| 32 | 744 | 37 |
| 33 | 780 | 61 |

Table 23b. Grade 4 Raw Score to Scale Score (with Standard Error)

| Weighted Raw Score | Scale Score | Standard Error |
|--------------------|-------------|----------------|
| 0 | 430 | 131 |
| 1 | 430 | 131 |
| 2 | 430 | 131 |
| 3 | 430 | 131 |
| 4 | 430 | 131 |
| 5 | 461 | 99 |
| 6 | 504 | 57 |
| 7 | 524 | 39 |
| 8 | 538 | 32 |
| 9 | 550 | 28 |
| 10 | 560 | 25 |
| 11 | 569 | 22 |
| 12 | 577 | 20 |
| 13 | 584 | 18 |
| 14 | 590 | 17 |
| 15 | 596 | 16 |
| 16 | 601 | 15 |
| 17 | 606 | 14 |
| 18 | 611 | 14 |
| 19 | 616 | 13 |
| 20 | 620 | 13 |
| 21 | 624 | 12 |
| 22 | 628 | 12 |
| 23 | 632 | 12 |
| 24 | 636 | 12 |
| 25 | 640 | 11 |
| 26 | 644 | 11 |
| 27 | 648 | 11 |
| 28 | 652 | 11 |
| 29 | 656 | 11 |
| 30 | 660 | 11 |
| 31 | 664 | 12 |

(Continued on next page)

Table 23b. Grade 4 Raw Score to Scale Score (with Standard Error) (cont.)

| Weighted Raw Score | Scale Score | Standard Error |
|--------------------|-------------|----------------|
| 32 | 668 | 12 |
| 33 | 673 | 12 |
| 34 | 678 | 13 |
| 35 | 683 | 13 |
| 36 | 689 | 14 |
| 37 | 695 | 15 |
| 38 | 703 | 16 |
| 39 | 711 | 17 |
| 40 | 721 | 20 |
| 41 | 735 | 23 |
| 42 | 756 | 32 |
| 43 | 775 | 43 |

Table 23c. Grade 5 Raw Score to Scale Score (with Standard Error)

| Raw Score | Scale Score | Standard Error |
|-----------|-------------|----------------|
| 0 | 495 | 110 |
| 1 | 495 | 110 |
| 2 | 495 | 110 |
| 3 | 495 | 110 |
| 4 | 495 | 110 |
| 5 | 495 | 110 |
| 6 | 524 | 81 |
| 7 | 557 | 48 |
| 8 | 574 | 32 |
| 9 | 586 | 24 |
| 10 | 595 | 20 |
| 11 | 602 | 18 |
| 12 | 609 | 16 |
| 13 | 614 | 15 |
| 14 | 620 | 14 |
| 15 | 625 | 13 |
| 16 | 630 | 13 |
| 17 | 635 | 13 |
| 18 | 640 | 13 |
| 19 | 645 | 13 |
| 20 | 650 | 13 |
| 21 | 655 | 13 |
| 22 | 661 | 14 |
| 23 | 667 | 14 |
| 24 | 674 | 15 |

(Continued on next page)

Table 23c. Grade 5 Raw Score to Scale Score (with Standard Error) (cont.)

| Raw Score | Scale Score | Standard Error |
|-----------|-------------|----------------|
| 25 | 681 | 16 |
| 26 | 690 | 17 |
| 27 | 699 | 19 |
| 28 | 712 | 22 |
| 29 | 729 | 29 |
| 30 | 764 | 50 |
| 31 | 795 | 79 |

Table 23d. Grade 6 Raw Score to Scale Score (with Standard Error)

| Raw Score | Scale Score | Standard Error |
|-----------|-------------|----------------|
| 0 | 480 | 79 |
| 1 | 480 | 79 |
| 2 | 480 | 79 |
| 3 | 480 | 79 |
| 4 | 480 | 79 |
| 5 | 480 | 79 |
| 6 | 495 | 64 |
| 7 | 518 | 41 |
| 8 | 534 | 33 |
| 9 | 546 | 28 |
| 10 | 557 | 25 |
| 11 | 566 | 22 |
| 12 | 574 | 20 |
| 13 | 582 | 19 |
| 14 | 589 | 18 |
| 15 | 595 | 17 |
| 16 | 601 | 16 |
| 17 | 606 | 15 |
| 18 | 612 | 14 |
| 19 | 617 | 14 |
| 20 | 622 | 13 |
| 21 | 627 | 13 |
| 22 | 632 | 13 |
| 23 | 636 | 13 |
| 24 | 641 | 13 |
| 25 | 646 | 12 |
| 26 | 650 | 12 |
| 27 | 655 | 12 |
| 28 | 660 | 13 |
| 29 | 665 | 13 |

(Continued on next page)

Table 23d. Grade 6 Raw Score to Scale Score (with Standard Error) (cont.)

| Raw Score | Scale Score | Standard Error |
|-----------|-------------|----------------|
| 30 | 671 | 13 |
| 31 | 676 | 13 |
| 32 | 682 | 14 |
| 33 | 689 | 15 |
| 34 | 697 | 16 |
| 35 | 706 | 18 |
| 36 | 717 | 21 |
| 37 | 733 | 26 |
| 38 | 762 | 40 |
| 39 | 785 | 58 |

Table 23e. Grade 7 Raw Score to Scale Score (with Standard Error)

| Raw Score | Scale Score | Standard Error |
|-----------|-------------|----------------|
| 0 | 470 | 107 |
| 1 | 470 | 107 |
| 2 | 470 | 107 |
| 3 | 470 | 107 |
| 4 | 470 | 107 |
| 5 | 470 | 107 |
| 6 | 470 | 107 |
| 7 | 489 | 88 |
| 8 | 522 | 56 |
| 9 | 540 | 38 |
| 10 | 553 | 30 |
| 11 | 563 | 24 |
| 12 | 572 | 21 |
| 13 | 579 | 19 |
| 14 | 585 | 17 |
| 15 | 591 | 15 |
| 16 | 596 | 14 |
| 17 | 601 | 14 |
| 18 | 605 | 13 |
| 19 | 610 | 13 |
| 20 | 614 | 12 |
| 21 | 618 | 12 |
| 22 | 622 | 12 |
| 23 | 626 | 11 |
| 24 | 630 | 11 |
| 25 | 633 | 11 |
| 26 | 637 | 11 |

(Continued on next page)

Table 23e. Grade 7 Raw Score to Scale Score (with Standard Error) (cont.)

| Raw Score | Scale Score | Standard Error |
|-----------|-------------|----------------|
| 27 | 641 | 11 |
| 28 | 645 | 11 |
| 29 | 650 | 12 |
| 30 | 654 | 12 |
| 31 | 659 | 12 |
| 32 | 664 | 13 |
| 33 | 669 | 14 |
| 34 | 675 | 15 |
| 35 | 682 | 16 |
| 36 | 691 | 17 |
| 37 | 700 | 20 |
| 38 | 713 | 23 |
| 39 | 730 | 28 |
| 40 | 759 | 41 |
| 41 | 790 | 60 |

Table 23f. Grade 8 Raw Score to Scale Score (with Standard Error)

| Weighted Raw Score | Scale Score | Standard Error |
|--------------------|-------------|----------------|
| 0 | 430 | 119 |
| 1 | 430 | 119 |
| 2 | 430 | 119 |
| 3 | 430 | 119 |
| 4 | 430 | 119 |
| 5 | 458 | 91 |
| 6 | 499 | 50 |
| 7 | 516 | 33 |
| 8 | 529 | 27 |
| 9 | 538 | 23 |
| 10 | 547 | 21 |
| 11 | 554 | 19 |
| 12 | 560 | 17 |
| 13 | 566 | 16 |
| 14 | 572 | 15 |
| 15 | 577 | 14 |
| 16 | 582 | 13 |
| 17 | 586 | 13 |
| 18 | 590 | 12 |
| 19 | 594 | 12 |
| 20 | 598 | 12 |
| 21 | 602 | 12 |

(Continued on next page)

Table 23f. Grade 8 Raw Score to Scale Score (with Standard Error) (cont.)

| Weighted Raw Score | Scale Score | Standard Error |
|--------------------|-------------|----------------|
| 22 | 606 | 12 |
| 23 | 610 | 11 |
| 24 | 614 | 11 |
| 25 | 618 | 11 |
| 26 | 622 | 11 |
| 27 | 626 | 12 |
| 28 | 630 | 12 |
| 29 | 634 | 12 |
| 30 | 638 | 12 |
| 31 | 643 | 12 |
| 32 | 647 | 12 |
| 33 | 652 | 13 |
| 34 | 657 | 13 |
| 35 | 662 | 14 |
| 36 | 668 | 14 |
| 37 | 674 | 15 |
| 38 | 681 | 16 |
| 39 | 689 | 17 |
| 40 | 698 | 19 |
| 41 | 710 | 23 |
| 42 | 728 | 29 |
| 43 | 760 | 47 |
| 44 | 790 | 69 |

Standard Performance Index

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

Because objectives on all achievement tests are measured by relatively small numbers of items, CTB/McGraw-Hill's scoring system looks not only at how many of those items the student answered correctly but at additional information as well. In technical terms, the procedure CTB/McGraw-Hill uses to calculate the SPI is based on a combination of Item Response Theory (IRT) and Bayesian methodology. In non-technical terms, the procedure takes into consideration the number of items related to the objective that the student answered correctly, the difficulty level of those items, as well as the student's

performance on the rest of the test in which the objective is found. This use of additional information increases the accuracy of the SPI. Details on the SPI derivation procedure are provided in Appendix G.

For the 2006 Grades 3-8 ELA Tests, the performance on objectives was tied to the Level III cut score by computing the SPI target ranges. The expected SPI cuts were computed for the scale scores that are 1 standard error above and 1 standard error below the Level III cut (scale score of 650 for all grades). Table 24 presents SPI target ranges. The objectives in these tables are denoted as follows: 1 – Information and Understanding, 2 – Literary Response and Expression, and 3 – Critical Analysis and Evaluation.

Table 24. SPI Target Ranges

| Grade | Objective | No. Items | Total Points | Level III cut SPI target range |
|-------|-----------|-----------|--------------|--------------------------------|
| 3 | 1 | 9 | 9 | 72-86 |
| | 2 | 12 | 15 | 60-74 |
| | 3 | 6 | 6 | 58-74 |
| 4 | 1 | 13 | 13 | 58-73 |
| | 2 | 12 | 15 | 55-68 |
| | 3 | 5 | 8 | 60-71 |
| 5 | 1 | 12 | 13 | 55-73 |
| | 2 | 9 | 9 | 61-73 |
| | 3 | 5 | 6 | 57-73 |
| 6 | 1 | 12 | 12 | 50-66 |
| | 2 | 12 | 16 | 63-75 |
| | 3 | 4 | 8 | 67-77 |
| 7 | 1 | 15 | 16 | 72-85 |
| | 2 | 14 | 15 | 63-77 |
| | 3 | 5 | 7 | 68-81 |
| 8 | 1 | 11 | 15 | 68-82 |
| | 2 | 13 | 13 | 65-74 |
| | 3 | 4 | 8 | 69-82 |

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

chance of helping those students increase their skills in the areas measured by the test. SPI reports provide students, parents, and educators the opportunity to identify and target specific areas within the broader content domain for improving student academic performance.

IRT DIF Statistics

In addition to classical DIF analysis, an IRT based Linn-Harnisch statistical procedure was used to detect DIF on the Grades 3-8 ELA Tests (Linn & 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., African-American) 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:

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

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

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

and the overall group difference ($D_{i\cdot}$) between observed and expected proportion correct for item i in the complete group (over all deciles) is

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

These indices are indicators of the degree to which members of a specific subgroup perform better or worse than expected on each item. Differences for decile groups provide an index for each of the ten regions on the scale score (θ) scale. The decile group difference (D_{ig}) can be either positive or negative. When the difference (D_{ig}) is greater than or equal to 0.100, or less than or equal to -0.100, the item is flagged for potential DIF.

The following groups were analyzed using the IRT based DIF analysis: Female, Male, Asian, Black or African-American, Hispanic or Latino, American Indian/Alaska Native, White, High-Needs Districts (by NRC code), and Low-Needs Districts (by NRC code). 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-counts for the Native Hawaiian/Other Pacific Islander subgroup were insufficient for IRT DIF analyses. The N counts for all other groups were over 200. As shown in Table 25, one item was flagged for DIF in the grades 3 and 4 tests, 2 items were flagged in the grade 5 test and 3 items were flagged in the grade 6 test by the Linn-Harnisch method. No items were flagged for DIF in grades 7 or 8. A detailed list of flagged items including DIF direction and magnitude is presented in Appendix E.

Table 25. Number of Items Flagged for DIF by the Linn-Harnisch Method

| Grade | Number of Flagged Items |
|-------|-------------------------|
| 3 | 1 |
| 4 | 1 |
| 5 | 2 |
| 6 | 3 |

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 ELA Tests occurred in Albany from June 6th through June 9th 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 1, 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 June 12th 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 English Language Arts* and *New York State ELA 2006 Measurement Review Technical Report 2006 for English Language Arts* for more details.

Description of Standard Setting Process

The NYSTP ELA Standard Setting was a multi-step process during which New York State educators and policy makers set the new performance standards for the Grade 3-8 ELA 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 from across the New York State. 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 English/Language Arts (Grades 3-8 ELA Tests). For more details on Measurement Meeting 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 26.

Table 26. Measurement Review Forum based Recommended Impact Data

| Grade | % of Students in Each Performance Level | | | | % Levels III and IV |
|-------|---|----------|-----------|----------|---------------------|
| | Level I | Level II | Level III | Level IV | |
| 3 | 7.9% | 22.2% | 51.3% | 18.6% | 69.9% |
| 4 | 6.7% | 21.6% | 51.3% | 20.4% | 71.7% |
| 5 | 6.9% | 25.2% | 49.3% | 18.6% | 67.9% |
| 6 | 7.4% | 28.0% | 48.2% | 16.4% | 64.6% |
| 7 | 7.6% | 29.8% | 47.0% | 15.6% | 62.6% |
| 8 | 7.7% | 32.1% | 46.1% | 14.1% | 60.2% |

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 27.

Table 27. 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 | 394 | 416 | 467 | 7.1% | 11.7% | 48.4% | 32.8% | 81.2% |
| 4 | 442 | 473 | 526 | 6.3% | 18.5% | 56.9% | 18.4% | 75.3% |
| 5 | 508 | 527 | 564 | 12.9% | 14.8% | 38.5% | 33.9% | 72.3% |
| 6 | 528 | 570 | 616 | 3.3% | 19.1% | 49.7% | 27.9% | 77.6% |
| 7 | 605 | 633 | 679 | 11.5% | 23.4% | 47.2% | 18.0% | 65.2% |
| 8 | 658 | 685 | 729 | 11.1% | 24.9% | 46.2% | 17.9% | 64.1% |

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 28).

Table 28. 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 | 394 | 424 | 475 | 7.1% | 19.0% | 50.5% | 23.3% | 73.9% |
| 4 | 445 | 478 | 526 | 7.5% | 20.4% | 53.7% | 18.4% | 72.1% |
| 5 | 495 | 527 | 583 | 6.3% | 21.4% | 53.4% | 18.9% | 72.3% |
| 6 | 543 | 579 | 625 | 7.2% | 22.9% | 53.2% | 16.7% | 69.9% |
| 7 | 605 | 633 | 679 | 11.5% | 23.4% | 47.2% | 18.0% | 65.2% |
| 8 | 658 | 685 | 729 | 11.1% | 24.9% | 46.2% | 17.9% | 64.1% |

4. Measurement Review Forum – The same participants who attended the Measurement Review meeting 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. Fourteen participants who attended the second Measurement Review Forum worked in two groups (7 persons in each group) and reviewed NYSTP ELA 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 both Measurement Review Forum groups are presented in Tables 29 and 30.

Table 29. Measurement Review Forum (Group 1) 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 | 394 | 424 | 493 | 7.1% | 19.0% | 59.3% | 14.5% | 73.9% |
| 4 | 442 | 478 | 526 | 6.3% | 21.7% | 53.7% | 18.4% | 72.1% |
| 5 | 495 | 535 | 583 | 6.3% | 26.4% | 48.4% | 18.9% | 67.3% |
| 6 | 543 | 585 | 625 | 7.2% | 27.3% | 48.9% | 16.7% | 65.6% |
| 7 | 595 | 640 | 690 | 6.4% | 32.5% | 48.6% | 12.4% | 61.1% |
| 8 | 647 | 695 | 742 | 6.4% | 38.9% | 41.6% | 13.0% | 54.6% |

Table 30. Measurement Review Forum (Group 2) 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 | 394 | 424 | 493 | 7.1% | 19.0% | 59.3% | 14.5% | 73.9% |
| 4 | 445 | 478 | 536 | 7.5% | 20.4% | 59.0% | 13.1% | 72.1% |
| 5 | 495 | 535 | 595 | 6.3% | 26.4% | 55.0% | 12.3% | 67.3% |
| 6 | 543 | 589 | 639 | 7.2% | 32.0% | 49.1% | 11.7% | 60.8% |
| 7 | 599 | 644 | 701 | 7.9% | 35.4% | 48.9% | 7.8% | 56.7% |
| 8 | 651 | 695 | 753 | 7.8% | 37.6% | 46.0% | 8.6% | 54.6% |

5. NYSED final recommendation – At this stage NYSED reviewed historical results, results from Vertical Articulation Panel and the Measurement Forum and adjusted cut scores for some grades to allow for smooth trends from grade to grade and consistency with the past student performance. During the final adjustment process various educational policies were taken into consideration. The adjustments to cut scores were within 1 Standard Error of Measurement. The final cut score and impact data for ELA 3-8 tests are presented in Table 31.

Table 31. Final NYSED Approved Cut Scores and Impact Data

| Grade | Level II Cut Score | Level III Cut Score | Level IV Cut Score | Adjust- ment | % of Students in Each Performance Level | | | | % Level III and IV |
|-------|-----------------------------|------------------------------|-----------------------------|-----------------|--|-------------|--------------|-------------|--------------------------|
| | | | | | Level I | Level II | Level III | Level IV | |
| 3 | 399 | 430 | 505 | +5 SEM | 8.4% | 22.4% | 62.0% | 7.1% | 69.1% |
| 4 | 448 | 481 | 540 | + .25 SEM | 8.9% | 22.4 | 60.0% | 8.7% | 68.8% |
| 5 | 495 | 535 | 595 | n/a | 6.3% | 26.4% | 55.0% | 12.3% | 67.3% |
| 6 | 543 | 589 | 639 | n/a | 7.2% | 32.0% | 49.1% | 11.7% | 60.8% |
| 7 | 599 | 644 | 701 | n/a | 7.9% | 35.4% | 48.9% | 7.8% | 56.7% |
| 8 | 654 | 698 | 758 | + .25 SEM | 9.3% | 41.1% | 44.7% | 4.9% | 49.6% |

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 ELA 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 ELA 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 ELA 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 31 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 below in Table 32.

Table 32. Final Cut Scores NYSTP ELA

| Grade | Final ELA Cut Scores | | |
|-------|----------------------|-----------|----------|
| | Level II | Level III | Level IV |
| 3 | 616 | 650 | 730 |
| 4 | 612 | 650 | 716 |
| 5 | 608 | 650 | 711 |
| 6 | 598 | 650 | 705 |
| 7 | 600 | 650 | 712 |
| 8 | 602 | 650 | 715 |

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 ELA 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).

Reliability for Total Test

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

Table 33. ELA 3-8 Tests Reliability and Standard Error of Measurement

| Grade | N | # Items | # RS points | Cronbach's Alpha | SEM of Cronbach's | Feldt-Raju Alpha | SEM of Feldt-Raju |
|-------|--------|---------|-------------|------------------|-------------------|------------------|-------------------|
| 3 | 185533 | 28 | 33 | 0.85 | 2.28 | 0.86 | 2.18 |
| 4 | 190847 | 31 | 39 | 0.88 | 2.45 | 0.89 | 2.36 |
| 5 | 201138 | 27 | 31 | 0.82 | 2.37 | 0.83 | 2.28 |
| 6 | 204104 | 29 | 39 | 0.85 | 2.63 | 0.87 | 2.46 |
| 7 | 210518 | 35 | 41 | 0.87 | 2.65 | 0.88 | 2.56 |
| 8 | 212138 | 29 | 39 | 0.85 | 2.55 | 0.87 | 2.36 |

All of the coefficients for total test reliability are in the range of 0.82-0.89, which indicates high internal consistency. As expected, the lowest reliabilities were found for shortest tests (i.e., grade 5) and the highest reliabilities are associated with the longer tests (grades 4, 7 and 8).

Reliability of 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 estimates for the overall test form. Table 34 presents reliabilities for the MC subsets.

Table 34. Reliability and Standard Error of Measurement – MC Items Only

| Grade | N | # Items | Cronbach's Alpha | SEM of Cronbach's | Feldt-Raju Alpha | SEM of Feldt-Raju |
|-------|--------|---------|------------------|-------------------|------------------|-------------------|
| 3 | 185533 | 24 | 0.83 | 1.77 | 0.83 | 1.75 |
| 4 | 190847 | 28 | 0.86 | 2.09 | 0.86 | 2.08 |
| 5 | 201138 | 24 | 0.78 | 1.95 | 0.78 | 1.95 |
| 6 | 204104 | 26 | 0.81 | 2.10 | 0.82 | 2.09 |
| 7 | 210518 | 30 | 0.85 | 2.14 | 0.85 | 2.12 |
| 8 | 212138 | 26 | 0.81 | 2.02 | 0.81 | 2.02 |

Reliability of CR items

Reliability coefficients were also computed for the subsets of CR items. It should be noted that the Grades 3-8 ELA Tests include only 3 to 5 CR items, depending on grade level, and the results presented in Table 35 should be interpreted with caution.

Table 35. Reliability and Standard Error of Measurement - CR Items Only

| Grade | N | # Items | # RS Points | Cronbach's Alpha | SEM of Cronbach's | Feldt-Raju Alpha | SEM of Feldt-Raju |
|-------|--------|---------|-------------|------------------|-------------------|------------------|-------------------|
| 3 | 185533 | 4 | 9 | 0.57 | 1.35 | 0.62 | 1.27 |
| 4 | 190847 | 3 | 11 | 0.76 | 1.01 | 0.76 | 1.00 |
| 5 | 201138 | 3 | 7 | 0.57 | 1.21 | 0.60 | 1.17 |
| 6 | 204104 | 3 | 13 | 0.78 | 1.18 | 0.81 | 1.11 |
| 7 | 210518 | 5 | 11 | 0.64 | 1.46 | 0.67 | 1.39 |
| 8 | 212138 | 3 | 13 | 0.80 | 1.14 | 0.83 | 1.06 |

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

Test Reliability for NCLB reporting categories

In this section, reliability coefficients that were estimated for the population and NCLB reporting subgroups are presented. The reporting categories include the following: gender, ethnicity, Needs Resource Code (NRC), Limited English Proficiency (LEP) status, Disability status (all students with disabilities (SWD) together), and all students using test accommodations (SUA). As shown in Tables 36a-36f, 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 greater than 0.80, with the following exceptions: grade 3 Asian, grade 3 NRC=6 (Low Needs

districts), grade 5 Asian, grade 5 NRC=5 (Average Needs districts), grade 5 NRC=6, grade 5 LEP, grade 6 LEP, and grade 8 NRC=6. Feldt-Raju reliability coefficients, which tend to be larger than the Cronbach's Alpha estimates for the same group, were all larger than 0.80 with the following exceptions: grade 5 NRC=6, and grade 5 LEP. All other test reliability alpha statistics were in the 0.80-0.90 range, indicating very good test internal consistency (reliability) for analyzed subgroups of examinees.

Table 36a. 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 | 185533 | 0.85 | 2.28 | 0.86 | 2.18 |
| Gender | Female | 90870 | 0.84 | 2.20 | 0.85 | 2.12 |
| | Male | 94663 | 0.86 | 2.35 | 0.87 | 2.23 |
| Ethnicity | Asian | 11810 | 0.80 | 1.99 | 0.81 | 1.93 |
| | Black or African-American | 38854 | 0.85 | 2.50 | 0.86 | 2.38 |
| | Hispanic or Latino | 30441 | 0.85 | 2.41 | 0.86 | 2.30 |
| | American Indian or Alaska Native | 1002 | 0.86 | 2.46 | 0.87 | 2.35 |
| | Native Hawaiian/ Other Pacific Islander | 36 | 0.91 | 2.26 | 0.93 | 2.05 |
| | White | 103377 | 0.83 | 2.17 | 0.84 | 2.08 |
| | NRC | New York City | 61484 | 0.86 | 2.38 | 0.88 |
| | Big 4 Cites | 7351 | 0.85 | 2.53 | 0.87 | 2.42 |
| | High Needs Urban-Suburban | 14570 | 0.84 | 2.38 | 0.85 | 2.28 |
| | High Needs Rural | 11503 | 0.83 | 2.34 | 0.85 | 2.25 |
| | Average Needs | 58464 | 0.82 | 2.19 | 0.84 | 2.11 |
| | Low Needs | 29583 | 0.79 | 2.01 | 0.80 | 1.94 |
| | Charter | 1898 | 0.83 | 2.48 | 0.84 | 2.39 |
| SWD | All codes | 23259 | 0.86 | 2.68 | 0.87 | 2.55 |
| SUA | All codes | 23803 | 0.86 | 2.67 | 0.87 | 2.54 |
| LEP | LEP = Y | 3639 | 0.87 | 2.63 | 0.89 | 2.48 |

Table 36b. 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 | 190847 | 0.88 | 2.45 | 0.89 | 2.36 |
| Gender | Female | 93374 | 0.88 | 2.41 | 0.88 | 2.33 |
| | Male | 97473 | 0.89 | 2.48 | 0.89 | 2.38 |
| Ethnicity | Asian | 12595 | 0.85 | 2.28 | 0.86 | 2.20 |
| | Black or African-American | 37458 | 0.88 | 2.59 | 0.88 | 2.50 |
| | Hispanic or Latino | 33426 | 0.87 | 2.56 | 0.88 | 2.48 |
| | American Indian or Alaska Native | 957 | 0.88 | 2.57 | 0.89 | 2.48 |
| | Native Hawaiian/ Other Pacific Islander | 37 | 0.91 | 2.36 | 0.92 | 2.19 |
| | White | 106369 | 0.87 | 2.36 | 0.88 | 2.28 |
| | NRC | New York City | 64468 | 0.88 | 2.53 | 0.89 |
| NRC | Big 4 Cites | 7207 | 0.89 | 2.62 | 0.90 | 2.52 |
| | High Needs Urban-Suburban | 14939 | 0.88 | 2.52 | 0.89 | 2.44 |
| | High Needs Rural | 11686 | 0.88 | 2.51 | 0.88 | 2.43 |
| | Average Needs | 59860 | 0.87 | 2.38 | 0.88 | 2.30 |
| | Low Needs | 30602 | 0.84 | 2.22 | 0.85 | 2.15 |
| | Charter | 1463 | 0.87 | 2.58 | 0.88 | 2.51 |
| | SWD | All codes | 25698 | 0.88 | 2.69 | 0.88 |
| SUA | All codes | 29343 | 0.88 | 2.68 | 0.89 | 2.60 |
| LEP | LEP = Y | 4294 | 0.87 | 2.71 | 0.88 | 2.61 |

Table 36c. 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 | 201138 | 0.82 | 2.37 | 0.83 | 2.28 |
| Gender | Female | 99056 | 0.81 | 2.33 | 0.83 | 2.25 |
| | Male | 102082 | 0.83 | 2.40 | 0.84 | 2.31 |
| Ethnicity | Asian | 13211 | 0.79 | 2.21 | 0.81 | 2.14 |
| | Black or African-American | 39898 | 0.80 | 2.51 | 0.82 | 2.43 |
| | Hispanic or Latino | 37563 | 0.80 | 2.49 | 0.81 | 2.41 |
| | American Indian or Alaska Native | 1048 | 0.81 | 2.48 | 0.82 | 2.38 |
| | Native Hawaiian/ Other Pacific Islander | 49 | 0.82 | 2.34 | 0.83 | 2.24 |
| | White | 109365 | 0.80 | 2.26 | 0.82 | 2.19 |
| | NRC | New York City | 68999 | 0.82 | 2.45 | 0.83 |
| NRC | Big 4 Cites | 8004 | 0.82 | 2.54 | 0.84 | 2.44 |
| | High Needs Urban-Suburban | 15536 | 0.81 | 2.45 | 0.82 | 2.37 |
| | High Needs Rural | 12130 | 0.80 | 2.40 | 0.81 | 2.32 |
| | Average Needs | 62193 | 0.79 | 2.29 | 0.81 | 2.22 |
| | Low Needs | 31165 | 0.75 | 2.11 | 0.76 | 2.06 |
| | Charter | 2326 | 0.83 | 2.46 | 0.84 | 2.39 |
| | SWD | All codes | 28262 | 0.80 | 2.60 | 0.81 |
| SUA | All codes | 33019 | 0.80 | 2.59 | 0.82 | 2.51 |
| LEP | LEP = Y | 6662 | 0.77 | 2.61 | 0.78 | 2.54 |

Table 36d. 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 | 204104 | 0.85 | 2.63 | 0.87 | 2.46 |
| Gender | Female | 99692 | 0.85 | 2.58 | 0.86 | 2.43 |
| | Male | 104412 | 0.86 | 2.65 | 0.87 | 2.48 |
| Ethnicity | Asian | 12829 | 0.83 | 2.45 | 0.85 | 2.31 |
| | Black or African-American | 40882 | 0.83 | 2.76 | 0.85 | 2.60 |
| | Hispanic or Latino | 37726 | 0.83 | 2.74 | 0.85 | 2.59 |
| | American Indian or Alaska Native | 1136 | 0.85 | 2.72 | 0.87 | 2.54 |
| | Native Hawaiian/ Other Pacific Islander | 38 | 0.85 | 2.53 | 0.88 | 2.29 |
| | White | 111492 | 0.84 | 2.53 | 0.86 | 2.36 |
| | NRC | New York City | 69215 | 0.84 | 2.71 | 0.86 |
| NRC | Big 4 Cites | 8209 | 0.84 | 2.79 | 0.86 | 2.62 |
| | High Needs Urban-Suburban | 15964 | 0.84 | 2.69 | 0.86 | 2.53 |
| | High Needs Rural | 12923 | 0.84 | 2.64 | 0.86 | 2.47 |
| | Average Needs | 64345 | 0.83 | 2.54 | 0.85 | 2.39 |
| | Low Needs | 31219 | 0.81 | 2.37 | 0.83 | 2.25 |
| | Charter | 1456 | 0.83 | 2.67 | 0.84 | 2.54 |
| | SWD | All codes | 28105 | 0.81 | 2.80 | 0.83 |
| SUA | All codes | 30674 | 0.82 | 2.80 | 0.84 | 2.66 |
| LEP | LEP = Y | 5695 | 0.79 | 2.86 | 0.81 | 2.69 |

Table 36e. 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 | 210518 | 0.87 | 2.65 | 0.88 | 2.56 |
| Gender | Female | 102028 | 0.86 | 2.60 | 0.87 | 2.52 |
| | Male | 108489 | 0.88 | 2.68 | 0.89 | 2.59 |
| Ethnicity | Asian | 12520 | 0.84 | 2.52 | 0.86 | 2.42 |
| | Black or African-American | 43020 | 0.86 | 2.79 | 0.87 | 2.72 |
| | Hispanic or Latino | 38298 | 0.86 | 2.78 | 0.87 | 2.70 |
| | American Indian or Alaska Native | 1105 | 0.86 | 2.76 | 0.87 | 2.67 |
| | Native Hawaiian/ Other Pacific Islander | 44 | 0.86 | 2.60 | 0.87 | 2.52 |
| | White | 115528 | 0.86 | 2.54 | 0.87 | 2.44 |
| | NRC | New York City | 70517 | 0.87 | 2.75 | 0.88 |
| NRC | Big 4 Cites | 9186 | 0.86 | 2.80 | 0.87 | 2.74 |
| | High Needs Urban-Suburban | 16203 | 0.86 | 2.72 | 0.87 | 2.65 |
| | High Needs Rural | 13676 | 0.86 | 2.67 | 0.87 | 2.59 |
| | Average Needs | 67355 | 0.85 | 2.56 | 0.86 | 2.47 |
| | Low Needs | 31333 | 0.82 | 2.39 | 0.83 | 2.30 |
| | Charter | 1234 | 0.86 | 2.77 | 0.87 | 2.68 |
| | SWD | All codes | 28509 | 0.85 | 2.87 | 0.86 |
| SUA | All codes | 30776 | 0.85 | 2.86 | 0.86 | 2.81 |
| LEP | LEP = Y | 6393 | 0.82 | 2.90 | 0.83 | 2.84 |

Table 36f. 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 | 212138 | 0.85 | 2.55 | 0.87 | 2.36 |
| Gender | Female | 103748 | 0.84 | 2.48 | 0.86 | 2.31 |
| | Male | 108390 | 0.85 | 2.60 | 0.87 | 2.40 |
| Ethnicity | Asian | 12405 | 0.83 | 2.38 | 0.86 | 2.21 |
| | Black or African-American | 42862 | 0.83 | 2.70 | 0.85 | 2.53 |
| | Hispanic or Latino | 37497 | 0.83 | 2.67 | 0.85 | 2.50 |
| | American Indian or Alaska Native | 1045 | 0.83 | 2.67 | 0.85 | 2.50 |
| | Native Hawaiian/ Other Pacific Islander | 29 | 0.85 | 2.43 | 0.87 | 2.25 |
| | White | 118298 | 0.83 | 2.42 | 0.85 | 2.26 |
| | NRC | New York City | 70763 | 0.85 | 2.65 | 0.87 |
| Big 4 Cites | | 9462 | 0.84 | 2.74 | 0.85 | 2.58 |
| High Needs Urban-Suburban | | 15827 | 0.84 | 2.62 | 0.86 | 2.45 |
| High Needs Rural | | 13640 | 0.84 | 2.55 | 0.86 | 2.38 |
| Average Needs | | 68782 | 0.82 | 2.43 | 0.84 | 2.28 |
| Low Needs | | 31436 | 0.80 | 2.24 | 0.82 | 2.12 |
| Charter | | 899 | 0.84 | 2.57 | 0.86 | 2.43 |
| SWD | All codes | 28573 | 0.83 | 2.79 | 0.85 | 2.64 |
| SUA | All codes | 31745 | 0.83 | 2.78 | 0.85 | 2.62 |
| LEP | LEP = Y | 5982 | 0.81 | 2.81 | 0.83 | 2.65 |

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 33. SEMs ranged from 2.18 to 2.63, which is reasonable and small. In other words, the error of measurement from the observed test score ranged from approximately +/-2 to 3 raw score points. SEMs are directly related to reliability: the higher the reliability, the lower the standard error. As discussed, the reliability of these tests is relatively high, so it was expected that the SEMs would be very low.

The SEMs for subpopulations, as computed from Cronbach's Alpha and the Feldt-Raju reliability statistics, are presented in Tables 35a-35f. The SEMs associated with all

reliability estimates for all subpopulations are in the range of 1.99-2.90, which is acceptably close to those for the entire population. This narrow range indicates that across the Grades 3-8 ELA Tests, all students' test scores are reasonably reliable with minimal error.

Performance Level Classification Consistency and Accuracy

This subsection describes the analyses conducted to estimate performance level classification consistency and accuracy for the Grades 3-8 ELA Tests. In other words, this provides statistical information on the classification of students into the four performance categories (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 I.

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

Consistency

The results for classifying students into four performance levels are separated from results based solely on the Level III cut. 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 37 (below) depicts the consistency study results based on the range of performance levels for all grades. Overall, between 68% and 75% 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 0.52 to 0.58.

Table 37. Decision Consistency (All Cuts)

| Grade | N | Agreement | Inconsistency | Kappa |
|-------|--------|-----------|---------------|--------|
| 3 | 185533 | 0.7217 | 0.2783 | 0.5357 |
| 4 | 190847 | 0.7456 | 0.2544 | 0.5807 |
| 5 | 201138 | 0.6889 | 0.3111 | 0.5150 |
| 6 | 204104 | 0.7130 | 0.2870 | 0.5655 |
| 7 | 210518 | 0.7264 | 0.2736 | 0.5781 |
| 8 | 212138 | 0.7288 | 0.2712 | 0.5788 |

Table 38 (below) depicts the consistency study results based on two performance levels (passing and not passing) as defined by the Level III cut. Overall, about 85% to 87% 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 0.69-0.74.

Table 38. Decision Consistency (Level III Cut)

| Grade | N | Agreement | Inconsistency | Kappa |
|-------|--------|-----------|---------------|--------|
| 3 | 185533 | 0.8641 | 0.1359 | 0.6858 |
| 4 | 190847 | 0.8864 | 0.1136 | 0.7389 |
| 5 | 201138 | 0.8560 | 0.1440 | 0.6770 |
| 6 | 204104 | 0.8615 | 0.1385 | 0.7115 |
| 7 | 210518 | 0.8659 | 0.1341 | 0.7274 |
| 8 | 212138 | 0.8503 | 0.1497 | 0.7005 |

Accuracy

The results of classification accuracy are presented in Table 39, below. Included in the table are case counts (N-count), classification accuracy (Accuracy) for all performance

levels (All Cuts) and for the ‘meeting learning standards’ 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 76%-81% of the time across all performance levels, and approximately 90% of the time in regards to the Level III cut score.

Table 39. 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 | 185533 | 0.7900 | 0.1514 | 0.0586 | 0.8988 | 0.0621 | 0.0391 |
| 4 | 190847 | 0.8142 | 0.1244 | 0.0614 | 0.9163 | 0.0529 | 0.0308 |
| 5 | 201138 | 0.7655 | 0.1645 | 0.0700 | 0.8920 | 0.0701 | 0.0379 |
| 6 | 204104 | 0.7882 | 0.1415 | 0.0703 | 0.8992 | 0.0618 | 0.0390 |
| 7 | 210518 | 0.8000 | 0.1306 | 0.0694 | 0.9047 | 0.0506 | 0.0447 |
| 8 | 212138 | 0.8023 | 0.1225 | 0.0751 | 0.8932 | 0.0573 | 0.0495 |

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 ELA Tests. These include the scale score means, standard deviations, percentiles and performance level distributions for each grade's population and specific subgroups. Gender, ethnic identification, Need/Resource Category (NRC), Limited English Proficiency (LEP), Disability, and Accommodation 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 Statistics

Scale Score distribution 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 40. Next, scale score statistics are presented for selected sub-groups in each grade level. Some general observations: Females outperformed Males; Asian and White ethnicities outperformed their peers from other ethnic groups; students from Low Need and Average Need districts (as identified by NRC) outperformed students from other districts (New York City, Big 4 Cities, Urban-Suburban, Rural and Charter); students with LEP, Disabilities and/or Accommodations achieved below the State aggregate (All Students) in every percentile. This pattern of achievement was consistent across all grades.

Table 40. ELA 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 | 185533 | 668.79 | 40.91 | 617 | 642 | 665 | 690 | 717 |
| 4 | 190847 | 665.73 | 40.74 | 616 | 644 | 668 | 689 | 711 |
| 5 | 201138 | 662.69 | 41.17 | 614 | 640 | 661 | 690 | 712 |
| 6 | 204104 | 656.52 | 40.85 | 606 | 632 | 660 | 682 | 706 |
| 7 | 210518 | 652.29 | 40.95 | 605 | 630 | 654 | 675 | 700 |
| 8 | 212138 | 650.14 | 40.78 | 602 | 626 | 647 | 674 | 698 |

Grade 3

Scale score statistics and N-counts of demographic groups for grade 3 are presented in Table 41, below. The population scale score mean was 668.79. By gender subgroup, Females outperformed Males, but the difference was less than a quarter of a standard deviation. Asian and White students' scale score means exceeded the average scale score on the exam, as did students from Low Need and Average Need districts. Students with disabilities, testing accommodations, and limited English proficiency scored, on average, approximately one standard deviation below the mean scale score for the population. The

students with disabilities subgroup, which had a scale score mean about 41 scale score units below the population mean, was the lowest performing group analyzed. At the 50th percentile, the following groups exceeded the population score of 665: Female (672), Asian (680), Native Hawaiian/Other Pacific Islander (672), White (680), Average Need districts (672) and Low Need districts (690).

Table 41. Scale Score Distribution 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 | 185533 | 668.79 | 40.91 | 617 | 642 | 665 | 690 | 717 |
| Gender | Female | 90870 | 673.85 | 40.33 | 625 | 647 | 672 | 701 | 717 |
| | Male | 94663 | 663.93 | 40.88 | 613 | 638 | 665 | 690 | 717 |
| Ethnicity | American Indian or Alaska Native | 1002 | 651.34 | 39.27 | 605 | 629 | 650 | 672 | 701 |
| | Asian | 11810 | 685.69 | 37.75 | 642 | 665 | 680 | 701 | 744 |
| | Black or African-American | 38854 | 650.82 | 38.70 | 605 | 625 | 653 | 672 | 701 |
| | Hispanic or Latino | 30441 | 657.47 | 38.23 | 613 | 633 | 659 | 680 | 701 |
| | Native Hawaiian/Other Pacific Islander | 36 | 667.86 | 54.53 | 597 | 636 | 672 | 701 | 717 |
| | White | 103377 | 677.11 | 39.60 | 629 | 653 | 680 | 701 | 717 |
| Need/ Resource Category | New York City | 61484 | 661.06 | 41.47 | 609 | 638 | 659 | 690 | 717 |
| | Big 4 Cities | 7351 | 647.67 | 38.92 | 601 | 621 | 647 | 672 | 701 |
| | High Need Urban/Suburban | 14570 | 660.74 | 38.40 | 613 | 638 | 659 | 680 | 701 |
| | High Need Rural | 11503 | 663.72 | 37.58 | 617 | 642 | 665 | 690 | 717 |
| | Average Need | 58464 | 674.50 | 38.73 | 629 | 647 | 672 | 701 | 717 |
| | Low Need | 29583 | 686.99 | 37.88 | 642 | 665 | 690 | 701 | 744 |
| | Charter | 1898 | 651.46 | 35.24 | 609 | 629 | 653 | 672 | 701 |
| LEP | LEP = Y | 3639 | 633.54 | 40.77 | 581 | 605 | 633 | 659 | 680 |
| Student With Disability | All Codes | 23259 | 627.38 | 38.93 | 581 | 601 | 629 | 653 | 672 |
| Accommodation | All Codes | 23803 | 629.07 | 38.84 | 581 | 605 | 629 | 653 | 680 |

Grade 4

Scale score statistics and N-counts of demographic groups for grade 4 are presented in Table 42, below. The grade 4 population (All Students) mean was 665.73. By gender subgroup, Females outperformed Males, but the difference was less than a quarter of a

standard deviation. Asian and White students' scale score means exceeded the average scale score on the exam, as did students from Low Need and Average Need districts. Students from the Big 4 Cities achieved a lower scale score mean than their peers from with other NRC designations. Students with disabilities had a scale score mean nearly 44.5 scale score units (more than a standard deviation) below the population mean and were at or below the scale score of any given percentile for any other subgroup. At the 50th percentile, the following groups exceeded the population score of 668: Female (673), Asian (683), Native Hawaiian/Other Pacific Islander (689), White (678), Average Need districts (673) and Low Need districts (683).

Table 42. Scale Score Distribution Summary, by Subgroup, Grade 4

| Demographic Category (Subgroup) | | N | SS Mean | SS SD | 10 th %tile | 25 th %tile | 50 th %tile | 75 th %tile | 90 th %tile |
|---------------------------------|--|--------|---------|-------|------------------------|------------------------|------------------------|------------------------|------------------------|
| State | All Students | 190847 | 665.73 | 40.74 | 616 | 644 | 668 | 689 | 711 |
| Gender | Female | 93374 | 670.33 | 39.49 | 620 | 648 | 673 | 695 | 721 |
| | Male | 97473 | 661.32 | 41.43 | 611 | 640 | 664 | 689 | 711 |
| Ethnicity | American Indian or Alaska Native | 957 | 651.50 | 39.99 | 601 | 628 | 656 | 678 | 695 |
| | Asian | 12595 | 681.50 | 36.48 | 640 | 660 | 683 | 703 | 721 |
| | Black or African-American | 37458 | 649.53 | 40.13 | 601 | 628 | 652 | 673 | 695 |
| | Hispanic or Latino | 33426 | 652.79 | 39.08 | 606 | 632 | 656 | 678 | 695 |
| | Native Hawaiian/Other Pacific Islander | 37 | 678.68 | 58.54 | 611 | 664 | 689 | 711 | 735 |
| | White | 106369 | 673.76 | 38.95 | 628 | 652 | 678 | 695 | 721 |
| Need/Resource Category | New York City | 64468 | 656.69 | 40.69 | 611 | 632 | 660 | 683 | 703 |
| | Big 4 Cities | 7207 | 648.15 | 43.19 | 596 | 620 | 648 | 678 | 703 |
| | High Need Urban/Suburban | 14939 | 658.31 | 40.05 | 611 | 636 | 660 | 683 | 703 |
| | High Need Rural | 11686 | 658.54 | 39.61 | 611 | 636 | 660 | 683 | 703 |
| | Average Need | 59860 | 671.76 | 37.85 | 624 | 652 | 673 | 695 | 711 |
| | Low Need | 30602 | 685.57 | 35.25 | 644 | 664 | 683 | 703 | 735 |
| | Charter | 1463 | 649.98 | 37.21 | 601 | 624 | 652 | 673 | 695 |
| LEP | LEP = Y | 4294 | 623.52 | 43.80 | 569 | 601 | 628 | 652 | 673 |
| Students w/ Disabilities | All Codes | 25698 | 621.26 | 44.50 | 569 | 596 | 624 | 652 | 673 |
| Accommodations | All Codes | 29343 | 625.82 | 44.24 | 569 | 601 | 628 | 656 | 678 |

Grade 5

Scale score summary statistics for grade 5 students are in Table 43, below. Overall, the scale score mean was 662.69, with a standard deviation of 41.17. The difference between mean scale scores by gender groups was very small (about 6 scale score units). Asian and White students' scale score means exceeded the population mean scale score on the

exam, as did students from Low Need and Average Need districts. The students from the Big 4 Cities scored below their peers from schools with other NRC designations, and about a half of standard deviation below the population mean. Students with disabilities, testing accommodations, and limited English proficiency scored approximately one standard deviation below the mean scale score for the population. The LEP subgroup, which had a scale score mean slightly more than 44 scale score units below the population mean, was the lowest performing group analyzed. At the 50th percentile, the following groups exceeded the population score of 661: Female (667), Asian (674), Native Hawaiian/Other Pacific Islander (667), White (674), Average Need districts (667) and Low Need districts (681).

Table 43. Scale Score Distribution Summary, by Subgroup, Grade 5

| Demographic Category (Subgroup) | | N | SS Mean | SS SD | 10 th %tile | 25 th %tile | 50 th %tile | 75 th %tile | 90 th %tile |
|------------------------------------|--|--------|------------|----------|---------------------------|---------------------------|---------------------------|---------------------------|---------------------------|
| State | All Students | 201138 | 662.69 | 41.17 | 614 | 640 | 661 | 690 | 712 |
| Gender | Female | 99056 | 665.89 | 40.11 | 620 | 640 | 667 | 690 | 712 |
| | Male | 102082 | 659.59 | 41.94 | 614 | 635 | 661 | 681 | 712 |
| Ethnicity | American Indian or Alaska Native | 1048 | 650.73 | 39.18 | 609 | 630 | 650 | 674 | 699 |
| | Asian | 13211 | 677.97 | 39.88 | 635 | 655 | 674 | 699 | 729 |
| | Black or African- American | 39898 | 645.65 | 39.14 | 602 | 625 | 645 | 667 | 690 |
| | Hispanic or Latino | 37563 | 648.18 | 38.68 | 602 | 630 | 650 | 674 | 690 |
| | Native Hawaiian/Other Pacific Islander | 49 | 667.24 | 40.48 | 625 | 655 | 667 | 690 | 712 |
| | White | 109365 | 672.16 | 39.24 | 630 | 650 | 674 | 690 | 712 |
| Need/ Resource Category | New York City | 68999 | 653.72 | 41.32 | 609 | 630 | 655 | 674 | 699 |
| | Big 4 Cities | 8004 | 640.79 | 43.10 | 595 | 620 | 640 | 667 | 690 |
| | High Need Urban/Suburban | 15536 | 654.84 | 39.33 | 609 | 635 | 655 | 681 | 699 |
| | High Need Rural | 12130 | 657.65 | 38.18 | 614 | 635 | 661 | 681 | 699 |
| | Average Need | 62193 | 669.00 | 37.76 | 625 | 650 | 667 | 690 | 712 |
| | Low Need | 31165 | 684.01 | 36.15 | 645 | 661 | 681 | 699 | 729 |
| | Charter | 2326 | 648.03 | 43.11 | 602 | 630 | 650 | 674 | 690 |
| LEP | LEP = Y | 6662 | 618.34 | 40.14 | 574 | 602 | 625 | 645 | 661 |
| Students w/ Disability | All Codes | 28262 | 622.03 | 43.10 | 574 | 602 | 625 | 650 | 667 |
| Accommoda- tion | All Codes | 33019 | 625.91 | 42.62 | 574 | 602 | 630 | 650 | 674 |

Grade 6

Scale score summary statistics for grade 6 students are in Table 44, below. The scale score mean of the population (All Students) was 656.52, with a standard deviation of 40.85. The difference between average scale scores by gender groups was about 11 scale

score units. The Asian ethnic group had the highest average scale score mean (673), and White ethnicity students also exceeded the average scale score on the exam (666.5). Average Need and Low Need categories had scale score means that exceeded that of the population and smaller standard deviations, indicating consistently above average performance. The students from Big 4 Cities school districts had a mean score 21 scale score points below the population (about a half of a standard deviation below average), and 10-43 scale score points below their peers from schools with other NRC designations. Students with disabilities, testing accommodations, and/or limited English proficiency scored approximately one standard deviation below the mean scale score for the population. The LEP subgroup, which had a scale score mean 47 scale score units below the population, was the lowest performing group analyzed. At the 50th percentile, the following groups exceeded the population score of 660: Asian (671), Native Hawaiian/Other Pacific Islander (671), White (665), Average Need districts (665) and Low Need districts (676).

Table 44. Scale Score Distribution Summary, by Subgroup, Grade 6

| Demographic Category (Subgroup) | | N | SS Mean | SS SD | 10 th %tile | 25 th %tile | 50 th %tile | 75 th %tile | 90 th %tile |
|---------------------------------|--|--------|---------|-------|------------------------|------------------------|------------------------|------------------------|------------------------|
| State | All Students | 204104 | 656.52 | 40.85 | 606 | 632 | 660 | 682 | 706 |
| Gender | Female | 99692 | 661.04 | 39.86 | 612 | 636 | 660 | 689 | 706 |
| | Male | 104412 | 652.20 | 41.31 | 601 | 627 | 655 | 676 | 697 |
| Ethnicity | American Indian or Alaska Native | 1136 | 642.32 | 40.39 | 589 | 622 | 646 | 665 | 689 |
| | Asian | 12829 | 673.47 | 38.58 | 627 | 650 | 671 | 697 | 717 |
| | Black or African-American | 40882 | 638.78 | 37.88 | 589 | 617 | 641 | 665 | 682 |
| | Hispanic or Latino | 37726 | 640.83 | 37.54 | 595 | 617 | 641 | 665 | 682 |
| | Native Hawaiian/Other Pacific Islander | 38 | 667.03 | 44.46 | 636 | 655 | 671 | 689 | 706 |
| | White | 111492 | 666.52 | 39.10 | 622 | 646 | 665 | 689 | 717 |
| Need/Resource Category | New York City | 69215 | 645.97 | 39.95 | 595 | 622 | 646 | 671 | 697 |
| | Big 4 Cities | 8209 | 635.26 | 39.70 | 589 | 612 | 636 | 660 | 682 |
| | High Need Urban/Suburban | 15964 | 647.52 | 39.27 | 601 | 622 | 650 | 671 | 697 |
| | High Need Rural | 12923 | 651.83 | 38.73 | 606 | 632 | 655 | 676 | 697 |
| | Average Need | 64345 | 663.89 | 37.90 | 617 | 641 | 665 | 689 | 706 |
| | Low Need | 31219 | 678.99 | 36.41 | 636 | 655 | 676 | 697 | 717 |
| | Charter | 1456 | 643.30 | 36.13 | 601 | 622 | 641 | 665 | 689 |
| LEP | LEP = Y | 5695 | 609.33 | 38.26 | 557 | 589 | 612 | 636 | 655 |
| Students w/ Disabilities | All Codes | 28105 | 611.91 | 39.41 | 566 | 589 | 617 | 636 | 660 |
| Accommodation | All Codes | 30674 | 615.37 | 39.58 | 566 | 589 | 617 | 641 | 660 |

Grade 7

Scale score statistics and N-counts of demographic groups for grade 7 are presented in Table 45, below. The population scale score mean was 652.29. By gender subgroup, Females outperformed Males, but the difference was less than a quarter of a standard deviation. Asian and White students' scale score means exceeded the average scale score on the exam, but all ethnic groups' mean scale scores were within a half a standard deviation from the population mean. The mean scale score for students from Low Need and Average Need districts was higher than the population mean, and the Low Need scale score mean was approximately half a standard deviation above the population mean. Students with disabilities, testing accommodations, and limited English proficiency scored approximately one standard deviation below the mean scale score for the population. The LEP subgroup, which had a scale score mean nearly 47 scale score units below the population mean, was the lowest performing group analyzed. At the 50th percentile, the following groups exceeded the population score of 654: Female (659), Asian (669), White (664), Average Need districts (659) and Low Need districts (675).

Table 45. Scale Score Distribution 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 | 210518 | 652.29 | 40.95 | 605 | 630 | 654 | 675 | 700 |
| Gender | Female | 102028 | 656.28 | 39.65 | 610 | 633 | 659 | 682 | 700 |
| | Male | 108489 | 648.53 | 41.79 | 601 | 626 | 650 | 675 | 700 |
| Ethnicity | American Indian or Alaska Native | 1105 | 639.80 | 39.74 | 596 | 618 | 641 | 664 | 682 |
| | Asian | 12520 | 667.66 | 38.58 | 622 | 645 | 669 | 691 | 713 |
| | Black or African-American | 43020 | 634.03 | 38.32 | 591 | 614 | 637 | 659 | 675 |
| | Hispanic or Latino | 38298 | 636.49 | 38.11 | 591 | 614 | 637 | 659 | 682 |
| | Native Hawaiian/Other Pacific Islander | 44 | 650.70 | 35.15 | 601 | 628 | 652 | 672 | 691 |
| | White | 115528 | 662.78 | 38.81 | 618 | 641 | 664 | 682 | 713 |
| Need/Resource Category | New York City | 70517 | 641.87 | 40.33 | 596 | 618 | 641 | 669 | 691 |
| | Big 4 Cities | 9186 | 627.23 | 39.67 | 579 | 605 | 630 | 654 | 675 |
| | High Need Urban/Suburban | 16203 | 642.78 | 38.72 | 596 | 622 | 645 | 669 | 691 |
| | High Need Rural | 13676 | 647.56 | 38.51 | 601 | 626 | 650 | 669 | 691 |
| | Average Need | 67355 | 660.47 | 37.47 | 618 | 637 | 659 | 682 | 700 |
| | Low Need | 31333 | 674.85 | 36.09 | 633 | 654 | 675 | 691 | 713 |
| | Charter | 1234 | 638.51 | 38.18 | 596 | 618 | 641 | 664 | 682 |

(Continued on next page)

Table 45. Scale Score Distribution Summary, by Subgroup, Grade 7 (cont.)

| Demographic Category (Subgroup) | | N | SS Mean | SS Std Dev | 10 th %tile | 25 th %tile | 50 th %tile | 75 th %tile | 90 th %tile |
|---------------------------------|-----------|-------|---------|------------|------------------------|------------------------|------------------------|------------------------|------------------------|
| LEP | LEP = Y | 6393 | 605.67 | 38.74 | 563 | 585 | 610 | 630 | 645 |
| Students w/ Disabilities | All Codes | 28509 | 610.60 | 41.56 | 563 | 591 | 614 | 637 | 659 |
| Accommodation | All Codes | 30776 | 613.59 | 40.96 | 563 | 591 | 618 | 641 | 659 |

Grade 8

Scale score statistics and N-counts of demographic groups for grade 8 are presented in Table 46, below. The population scale score mean was 650.14 with a standard deviation of almost 41. The Female gender group scale score mean exceeded that of Males by about 11 scale score units. Asian, Native Hawaiian/Other Pacific Islander and White students' scale score means exceeded the population mean, but all ethnic groups' mean scale scores were within a half a standard deviation from the population mean. The mean scale score for students from Low Need and Average Need districts was higher than the population mean, and the mean of the Big 4 Cities was the lowest for any NRC group. Students with disabilities, testing accommodations, and limited English proficiency scored approximately one standard deviation below the mean scale score for the population. The LEP subgroup, which had a scale score mean of 602.61 (nearly 48 scale score units below the population mean) was the lowest performing group analyzed. At the 50th percentile, the following groups exceeded the population score of 647: Female (652), Asian (668), Native Hawaiian/Other Pacific Islander (662), White (657), Average Need districts (657) and Low Need districts (674).

Table 46. Scale Score Distribution 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 | 212138 | 650.14 | 40.78 | 602 | 626 | 647 | 674 | 698 |
| Gender | Female | 103748 | 655.77 | 40.25 | 606 | 630 | 652 | 681 | 710 |
| | Male | 108390 | 644.76 | 40.56 | 594 | 622 | 643 | 668 | 698 |
| Ethnicity | American Indian or Alaska Native | 1045 | 636.42 | 35.33 | 590 | 614 | 638 | 657 | 681 |
| | Asian | 12405 | 666.33 | 40.44 | 618 | 643 | 668 | 689 | 710 |
| | Black or African-American | 42862 | 630.59 | 36.32 | 586 | 610 | 630 | 652 | 674 |
| | Hispanic or Latino | 37497 | 634.28 | 36.35 | 590 | 614 | 634 | 657 | 681 |
| | Native Hawaiian/Other Pacific Islander | 29 | 654.97 | 35.39 | 610 | 634 | 662 | 674 | 698 |
| | White | 118298 | 660.68 | 39.30 | 614 | 638 | 657 | 681 | 710 |

(Continued on next page)

Table 46. Scale Score Distribution Summary, by Subgroup, Grade 8 (cont.)

| Demographic Category (Subgroup) | | N | SS Mean | SS Std Dev | 10 th %tile | 25 th %tile | 50 th %tile | 75 th %tile | 90 th %tile |
|------------------------------------|-----------------------------|-------|------------|------------------|---------------------------|---------------------------|---------------------------|---------------------------|---------------------------|
| Need/ Resource Category | New York City | 70763 | 638.82 | 39.38 | 594 | 614 | 638 | 662 | 689 |
| | Big 4 Cities | 9462 | 626.92 | 37.05 | 582 | 606 | 626 | 647 | 674 |
| | High Need Urban/Suburban | 15827 | 641.11 | 38.09 | 598 | 618 | 638 | 662 | 689 |
| | High Need Rural | 13640 | 647.18 | 37.79 | 602 | 626 | 647 | 668 | 698 |
| | Average Need | 68782 | 658.06 | 37.61 | 614 | 634 | 657 | 681 | 710 |
| | Low Need | 31436 | 673.79 | 37.94 | 630 | 652 | 674 | 698 | 728 |
| | Charter | 899 | 638.36 | 35.46 | 594 | 618 | 638 | 662 | 681 |
| LEP | LEP = Y | 5982 | 602.61 | 35.05 | 560 | 582 | 606 | 626 | 643 |
| Students w/ Disabilities | All Codes | 28573 | 608.46 | 35.90 | 566 | 586 | 610 | 630 | 652 |
| Accommoda- tion | All Codes | 31745 | 611.54 | 36.39 | 566 | 590 | 614 | 634 | 652 |

Performance Level Distributions

Tables 47-53 show the performance level distribution for all examinees from public and charter school with valid scores. Table 47 presents performance level data for total populations of students in grades 3-8. Tables 48 to 53 contain performance level data for selected subgroups of students. In general, these distributions reflect the same achievement trends in the scale score summary discussion. More female students were classified in Level III and above categories as compared to male students. Similarly more White and Asian students were classified in Level III and above categories as compared to their peers from other ethnic groups. Consistently with the scale score distribution across groups pattern, students from Low and Average Needs districts outperformed students from High Need districts (New York City, Big 4 Cities, Urban –Suburban, and Rural). The Level III and above rates for Limited English Proficiency students, students with disabilities, and students using testing accommodations 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. Please note that the case counts for the Native Hawaiian/Other Pacific Islander subgroup are very low and are heavily influenced by very high and/or very low achieving individual students.

Table 47. ELA Grades 3-8 Test Performance Level Distributions

| Grade | N | Level I | Level II | Level III | Level IV | Levels III & IV |
|-------|--------|---------|----------|-----------|----------|-----------------|
| 3 | 185533 | 8.53% | 22.47% | 61.92% | 7.07% | 69.00% |
| 4 | 190847 | 8.92% | 22.40% | 59.94% | 8.74% | 68.68% |
| 5 | 201138 | 6.38% | 26.45% | 54.86% | 12.31% | 67.17% |
| 6 | 204104 | 7.28% | 32.24% | 48.88% | 11.60% | 60.48% |
| 7 | 210518 | 8.03% | 35.55% | 48.66% | 7.76% | 56.42% |
| 8 | 212138 | 9.42% | 41.20% | 44.53% | 4.84% | 49.38% |

Grade 3

Performance level distributions and N-counts of demographic groups for grade 3 are presented in Table 48, below. Statewide, 69% of 3rd graders are Meeting Learning Standards or Meeting Learning Standards with Distinction (Levels III and IV). Over 10% of Male students are Not Meeting Standards (Level I), as compared to only 6% of Female students. The percent of students in Levels III and IV varies widely by ethnicity and Need/Resource Category subgroup. Over 85% of Asian students and/or students from Low Need districts are meeting or exceeding the Standards; whereas, about 50% of American Indian or Alaskan Native, Black or African-American, Charter, and/or Big 4 Cities students are in Level I (Not Meeting Standards) or II (Partially Meeting Learning Standards). About a third of students with limited English proficiency, disability status or testing accommodations are in Level I and only about 1% are in Level IV. The following groups had pass rates (percent of students in Levels III & IV) above the State average: Female, Asian, White, Average Need districts, and Low Need districts.

Table 48. Performance Level Distributions, By Subgroup, Grade 3

| Demographic Category (Subgroup) | | N | Level I | Level II | Level III | Level IV | Levels III & IV |
|---------------------------------|--|--------|---------|----------|-----------|----------|-----------------|
| State | All Students | 185533 | 8.53% | 22.47% | 61.92% | 7.07% | 69.00% |
| Gender | Female | 90870 | 6.36% | 20.18% | 64.86% | 8.60% | 73.45% |
| | Male | 94663 | 10.61% | 24.67% | 59.11% | 5.61% | 64.72% |
| Ethnicity | American Indian or Alaska Native | 1002 | 15.47% | 34.53% | 47.41% | 2.59% | 50.00% |
| | Asian | 11810 | 2.46% | 12.42% | 72.97% | 12.14% | 85.11% |
| | Black or African-American | 38854 | 16.09% | 33.20% | 48.12% | 2.58% | 50.70% |
| | Hispanic or Latino | 30441 | 11.86% | 29.24% | 55.67% | 3.23% | 58.90% |
| | Native Hawaiian/Other Pacific Islander | 36 | 16.67% | 16.67% | 58.33% | 8.33% | 66.67% |
| | White | 103377 | 5.33% | 17.48% | 67.83% | 9.36% | 77.19% |
| Need/Resource Category | New York City | 61484 | 11.93% | 26.60% | 56.30% | 5.17% | 61.47% |
| | Big 4 Cities | 7351 | 18.56% | 35.53% | 43.42% | 2.49% | 45.91% |
| | High Need Urban/Suburban | 14570 | 10.76% | 27.56% | 57.17% | 4.51% | 61.68% |
| | High Need Rural | 11503 | 8.96% | 25.24% | 61.17% | 4.63% | 65.80% |
| | Average Need | 58464 | 5.46% | 19.54% | 66.92% | 8.07% | 74.99% |
| | Low Need | 29583 | 2.48% | 11.88% | 72.75% | 12.89% | 85.64% |
| | Charter | 1898 | 14.59% | 35.19% | 48.05% | 2.16% | 50.21% |
| LEP | LEP = Y | 3639 | 31.27% | 34.13% | 33.42% | 1.18% | 34.60% |
| Students w/ Disabilities | All Codes | 23259 | 37.47% | 36.12% | 25.61% | 0.81% | 26.42% |
| Accommodation | All Codes | 23803 | 35.63% | 36.08% | 27.48% | 0.82% | 28.29% |

Grade 4

Performance level distributions and N-counts of demographic groups for grade 4 are presented in Table 49, below. Across New York, approximately 69% of 4th grade students are in Levels III and IV (also considered meeting standards). As was seen in grade 3, The Low Need subgroup had the highest percent of students in Levels III and IV (87.17%), and the Student with Disability subgroup had the lowest (26.15%). Students in the American Indian or Alaska Native, Black or African-American, and Hispanic or Latino subgroups had percent meeting standards slightly above 50% which was over 20% below the other ethnic subgroups. Over twice as many Big 4 City students are in Level I than the population. Over a third of students with limited English proficiency, disability status or testing accommodations are in Level I (approximately four times as many as the statewide rate of 8.92%) and only about 1% is in Level IV. The following groups had

percent of students meeting standards above the State average: Female, Asian, Native Hawaiian/Other Pacific Islander, White, Average Need districts, and Low Need districts.

Table 49. Performance Level Distributions, By Subgroup, Grade 4

| Demographic Category (Subgroup) | | N | Level I | Level II | Level III | Level IV | Levels III & IV |
|---------------------------------|--|--------|---------|----------|-----------|----------|-----------------|
| State | All Students | 190847 | 8.92% | 22.40% | 59.94% | 8.74% | 68.68% |
| Gender | Female | 93374 | 6.91% | 21.02% | 61.35% | 10.72% | 72.07% |
| | Male | 97473 | 10.84% | 23.71% | 58.60% | 6.84% | 65.44% |
| Ethnicity | American Indian or Alaska Native | 957 | 15.78% | 28.94% | 50.99% | 4.28% | 55.28% |
| | Asian | 12595 | 2.90% | 13.62% | 67.70% | 15.78% | 83.49% |
| | Black or African-American | 37458 | 15.52% | 32.65% | 48.10% | 3.72% | 51.83% |
| | Hispanic or Latino | 33426 | 13.10% | 31.42% | 51.46% | 4.02% | 55.48% |
| | Native Hawaiian/Other Pacific Islander | 37 | 10.81% | 8.11% | 59.46% | 21.62% | 81.08% |
| | White | 106369 | 5.93% | 16.94% | 65.94% | 11.19% | 77.13% |
| Need/Resource Category | New York City | 64468 | 11.79% | 29.28% | 53.04% | 5.88% | 58.93% |
| | Big 4 Cities | 7207 | 19.54% | 31.01% | 44.30% | 5.15% | 49.45% |
| | High Need Urban/Suburban | 14939 | 11.82% | 26.00% | 56.49% | 5.69% | 62.18% |
| | High Need Rural | 11686 | 10.72% | 26.30% | 57.56% | 5.41% | 62.97% |
| | Average Need | 59860 | 6.05% | 18.30% | 65.87% | 9.79% | 75.65% |
| | Low Need | 30602 | 2.59% | 10.24% | 70.44% | 16.73% | 87.17% |
| | Charter | 1463 | 16.68% | 30.49% | 49.56% | 3.28% | 52.84% |
| LEP | LEP = Y | 4294 | 36.17% | 37.19% | 25.52% | 1.12% | 26.64% |
| Students w/ Disabilities | All Codes | 25698 | 38.63% | 35.22% | 25.46% | 0.68% | 26.15% |
| Accommodation | All Codes | 29343 | 34.52% | 35.28% | 29.22% | 0.97% | 30.19% |

Grade 5

Performance level distributions and N-counts of demographic groups for grade 5 are presented in Table 50, below. About 68% of the grade 5 population is in Levels III and IV. As was seen in grades 3 and 4, The Low Need subgroup had the highest percent of students (meeting standards) in Levels III and IV (87.36%). The grade 5 Student with Disability subgroup has 26.24% of students meeting standards, second only to the LEP subgroup (21%). Fewer male students were in the Level I category than was observed with grades 3 and 4, by a few percentage points. Students in the American Indian or Alaska Native, Black or African-American, and Hispanic or Latino subgroups had rates around 50% of students meeting standards, approximately 25% percent less than other ethnic subgroups. Over twice as many Big 4 City students are in Level I than the

population's rate. Over a third of students with limited English proficiency, disability status or testing accommodations are in Level I (approximately four times as many as the statewide rate of 6.38%), yet less than a third are in Levels III and IV (combined) and a very low percent (less than 2) is Meeting Learning Standards with Distinction (Level IV). The following groups had percent of students meeting standards above the State average: Female, Asian, Native Hawaiian/Other Pacific Islander, White, Average Need districts, and Low Need districts.

Table 50. Performance Level Distributions, By Subgroup, Grade 5

| Demographic Category (Subgroup) | | N | Level I | Level II | Level III | Level IV | Levels III & IV |
|---------------------------------|--|---------------|---------|----------|-----------|----------|-----------------|
| State | All Students | 201138 | 6.38% | 26.45% | 54.86% | 12.31% | 67.17% |
| Gender | Female | 99056 | 5.00% | 25.05% | 56.38% | 13.57% | 69.95% |
| | Male | 102082 | 7.72% | 27.80% | 53.39% | 11.09% | 64.48% |
| Ethnicity | American Indian or Alaska Native | 1048 | 9.92% | 35.97% | 47.71% | 6.39% | 54.10% |
| | Asian | 13211 | 2.57% | 16.27% | 60.62% | 20.53% | 81.15% |
| | Black or African-American | 39898 | 11.12% | 39.93% | 44.04% | 4.90% | 48.94% |
| | Hispanic or Latino | 37563 | 10.15% | 37.42% | 47.11% | 5.32% | 52.43% |
| | Native Hawaiian/Other Pacific Islander | 49 | 4.08% | 16.33% | 67.35% | 12.24% | 79.59% |
| | White | 109365 | 3.78% | 18.90% | 60.84% | 16.48% | 77.32% |
| | Need/Resource Category | New York City | 68999 | 8.87% | 34.41% | 47.98% | 8.75% |
| | Big 4 Cities | 8004 | 15.85% | 40.18% | 38.98% | 4.99% | 43.97% |
| | High Need Urban/Suburban | 15536 | 7.96% | 32.65% | 51.36% | 8.02% | 59.38% |
| | High Need Rural | 12130 | 6.65% | 29.36% | 55.78% | 8.21% | 63.99% |
| | Average Need | 62193 | 3.86% | 20.92% | 61.37% | 13.85% | 75.22% |
| | Low Need | 31165 | 1.25% | 11.38% | 63.88% | 23.48% | 87.36% |
| | Charter | 2326 | 10.40% | 34.35% | 49.31% | 5.93% | 55.25% |
| LEP | LEP = Y | 6662 | 30.25% | 48.75% | 20.19% | 0.81% | 21.00% |
| Students w/ Disabilities | All Codes | 28262 | 28.12% | 45.64% | 24.82% | 1.43% | 26.24% |
| Accommodation | All Codes | 33019 | 25.00% | 45.42% | 27.75% | 1.83% | 29.58% |

Grade 6

Performance level distributions and N-counts of demographic groups for grade 6 are presented in Table 51, below. Statewide, 60.48% of grade 6 students were meeting standards (percent of students in Levels III and IV). As was seen in other grades, The Low Need subgroup had the most students meeting standards (nearly 83%), and the LEP, Student with Disability, and Accommodation subgroups had the fewest (13.63%, 16.55%

and 19.47%, respectively). Students in the American Indian or Alaska Native, Black or African-American, and Hispanic or Latino subgroups had below 50% of students meeting standards - over 25% below the other ethnic subgroups. Students from Low Need districts outperformed students in all other subgroups, across demographic categories. Over twice as many Big 4 City students were placed in in Level I than the general population and only about 38% of students from those districts were meeting standards (with 3.76% Meeting Learning Standards with Distinction). More than 50% of New York City and Charter school students are in Levels I and II, and High Need Urban/Suburban and Rural are only doing slightly better (with 48% and 43%, respectively, in Levels I and II). The majority of students with limited English proficiency, disability status and/or testing accommodations were in Level II (Partially Meeting Learning Standards), but less than 1% were in Level IV. The following groups had pass rates above the State average: Female, Asian, Native Hawaiian/Other Pacific Islander, White, Average Need districts, and Low Need districts.

Table 51. Performance Level Distributions, By Subgroup, Grade 6

| Demographic Category (Subgroup) | | N | Level I | Level II | Level III | Level IV | Levels III & IV |
|---------------------------------|--|--------|---------|----------|-----------|----------|-----------------|
| State | All Students | 204104 | 7.28% | 32.24% | 48.88% | 11.60% | 60.48% |
| Gender | Female | 99692 | 5.39% | 30.28% | 50.51% | 13.82% | 64.33% |
| | Male | 104412 | 9.08% | 34.11% | 47.32% | 9.48% | 56.81% |
| Ethnicity | American Indian or Alaska Native | 1136 | 13.03% | 40.85% | 41.11% | 5.02% | 46.13% |
| | Asian | 12829 | 2.68% | 20.19% | 56.09% | 21.04% | 77.13% |
| | Black or African-American | 40882 | 12.71% | 45.99% | 37.44% | 3.86% | 41.30% |
| | Hispanic or Latino | 37726 | 11.46% | 45.22% | 39.07% | 4.24% | 43.31% |
| | Native Hawaiian/Other Pacific Islander | 38 | 5.26% | 15.79% | 65.79% | 13.16% | 78.95% |
| | White | 111492 | 4.34% | 24.11% | 55.64% | 15.91% | 71.55% |
| Need/Resource Category | New York City | 69215 | 10.14% | 41.26% | 41.52% | 7.08% | 48.60% |
| | Big 4 Cities | 8209 | 15.71% | 46.58% | 33.94% | 3.76% | 37.70% |
| | High Need Urban/Suburban | 15964 | 9.95% | 38.49% | 44.53% | 7.03% | 51.56% |
| | High Need Rural | 12923 | 7.75% | 35.69% | 48.50% | 8.06% | 56.56% |
| | Average Need | 64345 | 4.49% | 26.33% | 55.57% | 13.62% | 69.19% |
| | Low Need | 31219 | 1.74% | 15.30% | 59.06% | 23.91% | 82.97% |
| | Charter | 1456 | 9.41% | 46.50% | 39.22% | 4.88% | 44.09% |
| LEP | LEP = Y | 5695 | 34.75% | 51.62% | 12.92% | 0.70% | 13.63% |
| Students w/ Disabilities | All Codes | 28105 | 32.84% | 50.61% | 15.86% | 0.69% | 16.55% |
| Accommodation | All Codes | 30674 | 30.06% | 50.47% | 18.50% | 0.96% | 19.47% |

Grade 7

Performance level distributions and N-counts of demographic groups for grade 7 are presented in Table 52, below. 56.42% of the grade 7 population is Meeting Learning Standards or Meeting Learning Standards with Distinction (Levels III and IV). Over 6% more Female than Male students are meeting standards. The percent of students in Levels III and IV varies widely by ethnicity and Need/Resource Category subgroup. 72% of Asian students and 68% of White students are in Levels III and IV, whereas all other ethnic subgroups are below the population average. Over 60% of Black or African-American and Hispanic or Latino students are in Levels I and II, and over 70% of Big 4 Cities students are in those Levels. Yet, over 80% of Low Need students are in Levels III and IV. Average Need schools outperformed the State average, with 66.12% of students passing. Less than 10% of LEP students are in Levels III and IV. The LEP, Student with Disability, and Accommodations subgroups are well below the performance achievement of the general population, with over 82% of those students in Levels I and II. The following subgroups have percent of students meeting standards exceeding the general population: Asian, White, Average Need districts, and Low Need districts.

Table 52. Performance Level Distributions, By Subgroup, Grade 7

| Demographic Category (Subgroup) | | N | Level I | Level II | Level III | Level IV | Levels III & IV |
|---------------------------------|--|--------|---------|----------|-----------|----------|-----------------|
| State | All Students | 210518 | 8.03% | 35.55% | 48.66% | 7.76% | 56.42% |
| Gender | Female | 102028 | 6.04% | 34.05% | 50.83% | 9.08% | 59.91% |
| | Male | 108489 | 9.90% | 36.96% | 46.61% | 6.53% | 53.14% |
| Ethnicity | American Indian or Alaska Native | 1105 | 12.04% | 44.89% | 39.10% | 3.98% | 43.08% |
| | Asian | 12520 | 3.33% | 24.61% | 58.14% | 13.92% | 72.06% |
| | Black or African-American | 43020 | 14.37% | 49.63% | 33.74% | 2.25% | 36.00% |
| | Hispanic or Latino | 38298 | 13.01% | 48.48% | 35.98% | 2.53% | 38.51% |
| | Native Hawaiian/Other Pacific Islander | 44 | 9.09% | 36.36% | 50.00% | 4.55% | 54.55% |
| | White | 115528 | 4.49% | 27.12% | 57.48% | 10.92% | 68.40% |
| Need/Resource Category | New York City | 70517 | 11.42% | 44.36% | 39.56% | 4.67% | 44.22% |
| | Big 4 Cities | 9186 | 18.99% | 51.25% | 28.00% | 1.76% | 29.76% |
| | High Need Urban/Suburban | 16203 | 10.66% | 44.10% | 41.05% | 4.18% | 45.24% |
| | High Need Rural | 13676 | 8.58% | 39.56% | 46.90% | 4.96% | 51.86% |
| | Average Need | 67355 | 4.51% | 29.36% | 56.77% | 9.35% | 66.12% |
| | Low Need | 31333 | 1.75% | 17.73% | 63.96% | 16.56% | 80.52% |
| | Charter | 1234 | 12.24% | 46.92% | 37.76% | 3.08% | 40.84% |
| LEP | LEP = Y | 6393 | 35.77% | 54.36% | 9.54% | 0.33% | 9.87% |

(Continued on next page)

Table 52. Performance Level Distributions, By Subgroup, Grade 7 (cont.)

| Demographic Category (Subgroup) | | N | Level I | Level II | Level III | Level IV | Levels III & IV |
|---------------------------------|-----------|-------|---------|----------|-----------|----------|-----------------|
| Students w/ Disabilities | All Codes | 28509 | 32.60% | 51.38% | 15.47% | 0.55% | 16.02% |
| Accommodation | All Codes | 30776 | 30.01% | 52.08% | 17.22% | 0.69% | 17.92% |

Grade 8

Performance level distributions and N-counts of demographic groups for grade 8 are presented in Table 53, below. Currently, 49.38% of the grade 8 population is Meeting Learning Standards or Meeting Learning Standards with Distinction (Levels III and IV). About 10% more Female than Male students are in Levels III or IV. The percent of students in Levels III and IV varies widely by ethnicity and Need/Resource Category subgroup. Over 65% of American Indian or Alaska Native, Black or African-American, and Hispanic or Latino students are in Levels I and II, and over 75% of Big 4 Cities students are in those Levels. Yet, 75% of Low Need students are in Levels III and IV. Average Need schools outperformed the State average, with 58.41% of students meeting standards. Less than 7% of LEP students are in Levels III and IV. The LEP, Student with Disability, and Accommodation subgroups are well below the performance achievement of the general population, with over 87% of those students in Levels I and II. The following subgroups have a higher percent of students meeting standards than the general population: Asian, Native Hawaiian/Pacific Islander, White, Average Need districts, and Low Need districts.

Table 53. Performance Level Distributions, By Subgroup, Grade 8

| Demographic Category (Subgroup) | | N | Level I | Level II | Level III | Level IV | Levels III & IV |
|---------------------------------|--|--------|---------|----------|-----------|----------|-----------------|
| State | All Students | 212138 | 9.42% | 41.20% | 44.53% | 4.84% | 49.38% |
| Gender | Female | 103748 | 6.92% | 38.14% | 48.93% | 6.00% | 54.94% |
| | Male | 108390 | 11.82% | 44.12% | 40.32% | 3.73% | 44.06% |
| Ethnicity | American Indian or Alaska Native | 1045 | 14.35% | 51.10% | 33.59% | 0.96% | 34.55% |
| | Asian | 12405 | 4.61% | 28.44% | 58.03% | 8.92% | 66.95% |
| | Black or African-American | 42862 | 17.93% | 53.99% | 26.92% | 1.16% | 28.08% |
| | Hispanic or Latino | 37497 | 15.01% | 53.56% | 30.07% | 1.35% | 31.43% |
| | Native Hawaiian/Other Pacific Islander | 29 | 6.90% | 31.03% | 58.62% | 3.45% | 62.07% |
| | White | 118298 | 5.03% | 33.90% | 54.18% | 6.89% | 61.07% |

(Continued on next page)

Table 53. Performance Level Distributions, 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 | 70763 | 14.07% | 49.36% | 33.85% | 2.72% | 36.57% |
| | Big 4 Cities | 9462 | 21.42% | 54.16% | 23.22% | 1.19% | 24.41% |
| | High Need Urban/Suburban | 15827 | 12.12% | 48.94% | 36.14% | 2.80% | 38.94% |
| | High Need Rural | 13640 | 8.86% | 45.50% | 42.26% | 3.38% | 45.64% |
| | Average Need | 68782 | 4.97% | 36.62% | 52.78% | 5.64% | 58.41% |
| | Low Need | 31436 | 2.06% | 22.89% | 64.12% | 10.92% | 75.04% |
| | Charter | 899 | 13.46% | 50.83% | 33.26% | 2.45% | 35.71% |
| LEP | LEP = Y | 5982 | 44.65% | 48.73% | 6.47% | 0.15% | 6.62% |
| Students w/ Disabilities | All Codes | 28573 | 38.15% | 51.56% | 10.06% | 0.23% | 10.29% |
| Accommodation | All Codes | 31745 | 35.18% | 52.30% | 12.14% | 0.38% | 12.52% |

Section X: Special Studies

Linking ELA grades 4 and 8 2006 to 2005 assessments

This section provides a description and results of the linking study conducted to provide a crosswalk between the 2005 grades 4 and 8 ELA assessments and 2006 grades 4 and 8 ELA assessments. The 2005 and 2006 ELA assessments for grades 4 and 8 are on different scales and were not equated to each other. Although equating is considered to be the strongest linking procedure it was not appropriate for linking the two years of New York State ELA assessments. In order to conduct direct test equating two major conditions must be satisfied (Linn, 1991):

- Linked test forms must measure the same construct with an equal degree of reliability (the forms are then interchangeable), and
- Common items must be administered during both administrations or common examinees must take at least some items from both assessments

Although the first condition was met for the 2005 and 2006 ELA grades 4 and 8 assessments, the second one was not. Because New York State test items are released after each operational administration, the inclusion of 2005 operational items in 2006 ELA assessments was not feasible. For the same reason, it was not appropriate to administer the 2005 assessments to the same examinees that participated in 2006 assessments. In the previous 4 and 8 testing program a pre-equating design was employed to equate field test items to the operational scale. Using the same design to equate 2006 to 2005 tests was neither recommended (for psychometric reasons) nor feasible (for content reasons). The New York State field test design has a number of limitations. First, field test items are not embedded in operational forms, nor are they administered during the operational testing window. Consequently, examinees know that they are taking a field test and may be less motivated to perform well on field test items. Second, the New York State field test is administered in a format of several mini-forms to randomly equivalent groups of people. As a result, the field test format, test length, testing time and item order do not resemble those of the operational tests. In addition, one of the important components of the test, Writing Mechanics, was not field tested and no field test item parameters for these items were available. The pre-equating design only allows for development of scoring tables based on field test item parameters. Because the Writing Mechanics component is an important part of the test content and construct, not having parameters for this component prevented scoring table generation at this stage of analysis. A post-equating design was also considered to link the 2005 and 2006 grade 4 and 8 ELA assessments. In this design, the 2005 operational test items are re-calibrated after operational test administration using field test parameters as anchors. This approach provides more reliable estimates of operational parameters, but nevertheless relies on field test parameters as anchor input. This approach will be used in 2007 and beyond to equate the new ELA test forms to the baseline 2006 year. The post-equating design was not employed in linking the 2005 and 2006 assessment because its outcome was not in alignment with qualitative results of the standard setting.

New performance standards for grades 4 and 8 were set during the 2006 Standard Setting process. The new passing score (Level III cut) was set for these grades based on test content and various educational policies. The impact data associated with the new cut scores also satisfied the notion of vertical moderation of standards for grades 3 through 8 (more details on setting standards and vertical moderation is provided in Section VII: Standard Setting, the *Bookmark Standard Setting Technical Report 2006 for Grades 3, 4, 5, 6, 7, and 8 English Language Arts* and *New York State ELA 2006 Measurement Review Technical Report 2006 for English Language Arts*). In a process of vertical moderation of the new standards, the 2006 impact data were established to be similar to that of 2005 for grades 4 and 8. Such moderation provided desired consistency between the 2005 and 2006 proficiency trends for grades 4 and 8. It can be said that because proportions of students in Levels III and IV proficiency categories in 2005 and 2006 were taken into consideration while moderating standards for grades 4 and 8 (as well as remaining grades), this procedure was equipercentile-like.

In summary, limitations of the Grades 3-8 operational tests (no common items administered every year), limitations of the field test design, and the method used to set and smooth new performance cut scores led to a decision to conduct an equipercentile linking of the 2005 and 2006 grades 4 and 8 assessments and to develop concordance tables. This method is appropriate to use if the two tests are comparable in terms of test content, format, and are administered to randomly equivalent groups. Student populations in 2005 and 2006 are comparable in terms of population characteristics and the NYSTP 2005 and 2006 ELA tests for grades 4 and 8 are comparable in test content and format. The equipercentile linking of the 2005 test to the 2006 test was conducted by identifying scores on the 2005 test that have the same percentile rank as the scores on the 2006 test.

The percentile rank of an integer score as defined by Kolen and Brennan (1995) is the percentile rank at the midpoint of the interval that contains that score and is computed as follows:

$$\%y = CFDy + 1/2(FDy)$$

where

$\%y$ is the percentile rank of a score y

$CFDy$ is the cumulative % frequency distribution below the score y

FDy is the % frequency distribution at the score y

It should be noted that the New York State ELA scales span over 300 possible scale score points and the number of raw score points is 42 to 44 depending on grade level and administration year. Therefore, the percentile ranks for scale scores that were not in 2005 and 2006 scoring tables were interpolated. This resulted in assignment of percentile rank to consecutive scale scores in 2005 and 2006 scale scores. Next, the 2006 scale scores were linked to the 2005 scale scores via same percentile rank. This procedure resulted in an accurate link between the new and old test forms. The concordance tables including scale scores from 2005 and 2006 assessments and associated percentile ranks are presented in Tables 54 and 55, below.

Table 54. Grade 4 ELA Linking of 2006 to 2005 Scale Scores

| 2006 Scale | Percentile rank | 2005 Scale | 2006 Scale Scores |
|------------|-----------------|------------|-------------------|
| 430 | 0.00 | 455 | ***** |
| 431 | 0.00 | 455 | |
| 432 | 0.01 | 456 | |
| 433 | 0.01 | 457 | |
| 434 | 0.02 | 457 | |
| 435 | 0.02 | 458 | |
| 436 | 0.02 | 458 | |
| 437 | 0.03 | 459 | |
| 438 | 0.03 | 459 | |
| 439 | 0.03 | 460 | |
| 440 | 0.04 | 461 | |
| 441 | 0.04 | 461 | |
| 442 | 0.05 | 462 | |
| 443 | 0.05 | 462 | |
| 444 | 0.05 | 463 | |
| 445 | 0.06 | 463 | |
| 446 | 0.06 | 464 | |
| 447 | 0.06 | 465 | |
| 448 | 0.07 | 465 | |
| 449 | 0.07 | 466 | |
| 450 | 0.08 | 466 | |
| 451 | 0.08 | 467 | |
| 452 | 0.08 | 467 | |
| 453 | 0.09 | 468 | |
| 454 | 0.09 | 468 | |
| 455 | 0.09 | 469 | |
| 456 | 0.10 | 470 | |
| 457 | 0.10 | 470 | |
| 458 | 0.10 | 471 | |
| 459 | 0.11 | 471 | |
| 460 | 0.11 | 472 | |
| 461 | 0.12 | 472 | ***** |
| 462 | 0.12 | 473 | |
| 463 | 0.12 | 473 | |
| 464 | 0.12 | 474 | |
| 465 | 0.13 | 474 | |
| 466 | 0.13 | 474 | |
| 467 | 0.13 | 475 | |
| 468 | 0.13 | 475 | |
| 469 | 0.14 | 475 | |

(Continued on next page)

Table 54. Grade 4 ELA Linking of 2006 to 2005 Scale Scores (cont.)

| 2006 Scale | Percentile rank | 2005 Scale | 2006 Scale Scores |
|------------|-----------------|------------|-------------------|
| 470 | 0.14 | 476 | |
| 471 | 0.14 | 476 | |
| 472 | 0.14 | 477 | |
| 473 | 0.15 | 477 | |
| 474 | 0.15 | 477 | |
| 475 | 0.15 | 478 | |
| 476 | 0.15 | 478 | |
| 477 | 0.16 | 479 | |
| 478 | 0.16 | 479 | |
| 479 | 0.16 | 479 | |
| 480 | 0.16 | 480 | |
| 481 | 0.17 | 480 | |
| 482 | 0.17 | 480 | |
| 483 | 0.17 | 481 | |
| 484 | 0.17 | 481 | |
| 485 | 0.18 | 482 | |
| 486 | 0.18 | 482 | |
| 487 | 0.18 | 482 | |
| 488 | 0.18 | 483 | |
| 489 | 0.19 | 483 | |
| 490 | 0.19 | 484 | |
| 491 | 0.19 | 484 | |
| 492 | 0.19 | 484 | |
| 493 | 0.20 | 485 | |
| 494 | 0.20 | 485 | |
| 495 | 0.20 | 485 | |
| 496 | 0.20 | 486 | |
| 497 | 0.21 | 486 | |
| 498 | 0.21 | 487 | |
| 499 | 0.21 | 487 | |
| 500 | 0.21 | 487 | |
| 501 | 0.22 | 488 | |
| 502 | 0.22 | 488 | |
| 503 | 0.22 | 489 | |
| 504 | 0.22 | 489 | ***** |
| 505 | 0.23 | 490 | |
| 506 | 0.24 | 491 | |
| 507 | 0.25 | 493 | |
| 508 | 0.26 | 494 | |
| 509 | 0.27 | 495 | |

(Continued on next page)

Table 54. Grade 4 ELA Linking of 2006 to 2005 Scale Scores (cont.)

| 2006 Scale | Percentile rank | 2005 Scale | 2006 Scale Scores |
|------------|-----------------|------------|-------------------|
| 510 | 0.28 | 497 | |
| 511 | 0.28 | 498 | |
| 512 | 0.29 | 499 | |
| 513 | 0.30 | 500 | |
| 514 | 0.31 | 502 | |
| 515 | 0.32 | 503 | |
| 516 | 0.33 | 504 | |
| 517 | 0.34 | 505 | |
| 518 | 0.34 | 507 | |
| 519 | 0.35 | 508 | |
| 520 | 0.36 | 509 | |
| 521 | 0.37 | 510 | |
| 522 | 0.38 | 512 | |
| 523 | 0.39 | 513 | |
| 524 | 0.40 | 514 | ***** |
| 525 | 0.41 | 517 | |
| 526 | 0.43 | 519 | |
| 527 | 0.45 | 521 | |
| 528 | 0.46 | 524 | |
| 529 | 0.48 | 526 | |
| 530 | 0.50 | 529 | |
| 531 | 0.51 | 530 | |
| 532 | 0.53 | 531 | |
| 533 | 0.55 | 532 | |
| 534 | 0.56 | 533 | |
| 535 | 0.58 | 534 | |
| 536 | 0.59 | 535 | |
| 537 | 0.61 | 536 | |
| 538 | 0.63 | 536 | ***** |
| 539 | 0.65 | 538 | |
| 540 | 0.68 | 539 | |
| 541 | 0.70 | 541 | |
| 542 | 0.73 | 542 | |
| 543 | 0.75 | 544 | |
| 544 | 0.77 | 545 | |
| 545 | 0.80 | 546 | |
| 546 | 0.82 | 547 | |
| 547 | 0.85 | 548 | |
| 548 | 0.87 | 548 | |
| 549 | 0.90 | 549 | |

(Continued on next page)

Table 54. Grade 4 ELA Linking of 2006 to 2005 Scale Scores (cont.)

| 2006 Scale | Percentile rank | 2005 Scale | 2006 Scale Scores |
|------------|-----------------|------------|-------------------|
| 550 | 0.92 | 550 | ***** |
| 551 | 0.96 | 551 | |
| 552 | 0.99 | 552 | |
| 553 | 1.03 | 553 | |
| 554 | 1.07 | 554 | |
| 555 | 1.10 | 556 | |
| 556 | 1.14 | 557 | |
| 557 | 1.18 | 558 | |
| 558 | 1.21 | 559 | |
| 559 | 1.25 | 559 | |
| 560 | 1.28 | 560 | ***** |
| 561 | 1.34 | 561 | |
| 562 | 1.39 | 562 | |
| 563 | 1.44 | 563 | |
| 564 | 1.49 | 564 | |
| 565 | 1.54 | 565 | |
| 566 | 1.59 | 566 | |
| 567 | 1.64 | 567 | |
| 568 | 1.69 | 568 | |
| 569 | 1.74 | 569 | ***** |
| 570 | 1.82 | 570 | |
| 571 | 1.89 | 571 | |
| 572 | 1.96 | 572 | |
| 573 | 2.04 | 573 | |
| 574 | 2.11 | 574 | |
| 575 | 2.18 | 575 | |
| 576 | 2.25 | 576 | |
| 577 | 2.33 | 577 | ***** |
| 578 | 2.43 | 578 | |
| 579 | 2.53 | 579 | |
| 580 | 2.62 | 581 | |
| 581 | 2.72 | 582 | |
| 582 | 2.82 | 583 | |
| 583 | 2.92 | 584 | |
| 584 | 3.02 | 585 | ***** |
| 585 | 3.15 | 586 | |
| 586 | 3.28 | 587 | |
| 587 | 3.41 | 588 | |
| 588 | 3.55 | 589 | |

(Continued on next page)

Table 54. Grade 4 ELA Linking of 2006 to 2005 Scale Scores (cont.)

| 2006 Scale | Percentile rank | 2005 Scale | 2006 Scale Scores |
|------------|-----------------|------------|-------------------|
| 589 | 3.68 | 590 | |
| 590 | 3.81 | 591 | ***** |
| 591 | 3.96 | 592 | |
| 592 | 4.11 | 593 | |
| 593 | 4.26 | 594 | |
| 594 | 4.41 | 595 | |
| 595 | 4.56 | 596 | |
| 596 | 4.71 | 596 | ***** |
| 597 | 4.92 | 598 | |
| 598 | 5.12 | 599 | |
| 599 | 5.33 | 600 | |
| 600 | 5.53 | 601 | |
| 601 | 5.74 | 602 | ***** |
| 602 | 5.97 | 603 | |
| 603 | 6.20 | 604 | |
| 604 | 6.43 | 605 | |
| 605 | 6.66 | 606 | |
| 606 | 6.89 | 607 | ***** |
| 607 | 7.16 | 607 | |
| 608 | 7.42 | 608 | |
| 609 | 7.69 | 609 | |
| 610 | 7.95 | 609 | |
| 611 | 8.21 | 610 | ***** |
| 612 | 8.51 | 611 | |
| 613 | 8.82 | 612 | |
| 614 | 9.12 | 613 | |
| 615 | 9.42 | 614 | |
| 616 | 9.72 | 614 | ***** |
| 617 | 10.14 | 615 | |
| 618 | 10.57 | 616 | |
| 619 | 11.00 | 617 | |
| 620 | 11.43 | 618 | ***** |
| 621 | 11.91 | 619 | |
| 622 | 12.40 | 620 | |
| 623 | 12.88 | 621 | |
| 624 | 13.37 | 622 | ***** |
| 625 | 13.91 | 623 | |
| 626 | 14.44 | 624 | |
| 627 | 14.98 | 625 | |
| 628 | 15.52 | 625 | ***** |

(Continued on next page)

Table 54. Grade 4 ELA Linking of 2006 to 2005 Scale Scores (cont.)

| 2006 Scale | Percentile rank | 2005 Scale | 2006 Scale Scores |
|------------|-----------------|------------|-------------------|
| 629 | 16.12 | 626 | |
| 630 | 16.71 | 627 | |
| 631 | 17.30 | 628 | |
| 632 | 17.90 | 629 | ***** |
| 633 | 18.55 | 630 | |
| 634 | 19.20 | 631 | |
| 635 | 19.85 | 632 | |
| 636 | 20.50 | 633 | ***** |
| 637 | 21.20 | 633 | |
| 638 | 21.91 | 634 | |
| 639 | 22.62 | 635 | |
| 640 | 23.33 | 636 | ***** |
| 641 | 24.09 | 637 | |
| 642 | 24.86 | 638 | |
| 643 | 25.62 | 639 | |
| 644 | 26.39 | 640 | ***** |
| 645 | 27.20 | 640 | |
| 646 | 28.01 | 641 | |
| 647 | 28.83 | 642 | |
| 648 | 29.64 | 643 | ***** |
| 649 | 30.52 | 644 | |
| 650 | 31.40 | 645 | |
| 651 | 32.28 | 646 | |
| 652 | 33.17 | 647 | ***** |
| 653 | 34.14 | 648 | |
| 654 | 35.12 | 649 | |
| 655 | 36.10 | 649 | |
| 656 | 37.08 | 650 | ***** |
| 657 | 38.13 | 651 | |
| 658 | 39.19 | 652 | |
| 659 | 40.24 | 653 | |
| 660 | 41.30 | 654 | ***** |
| 661 | 42.43 | 655 | |
| 662 | 43.56 | 656 | |
| 663 | 44.69 | 657 | |
| 664 | 45.82 | 659 | ***** |
| 665 | 47.04 | 660 | |
| 666 | 48.26 | 661 | |
| 667 | 49.48 | 662 | |

(Continued on next page)

Table 54. Grade 4 ELA Linking of 2006 to 2005 Scale Scores (cont.)

| 2006 Scale | Percentile rank | 2005 Scale | 2006 Scale Scores |
|------------|-----------------|------------|-------------------|
| 668 | 50.70 | 663 | ***** |
| 669 | 51.73 | 664 | |
| 670 | 52.76 | 665 | |
| 671 | 53.78 | 666 | |
| 672 | 54.81 | 667 | |
| 673 | 55.84 | 668 | ***** |
| 674 | 56.92 | 669 | |
| 675 | 58.01 | 670 | |
| 676 | 59.09 | 671 | |
| 677 | 60.17 | 672 | |
| 678 | 61.26 | 673 | ***** |
| 679 | 62.41 | 674 | |
| 680 | 63.56 | 676 | |
| 681 | 64.71 | 677 | |
| 682 | 65.86 | 679 | |
| 683 | 67.01 | 680 | ***** |
| 684 | 68.00 | 681 | |
| 685 | 68.99 | 683 | |
| 686 | 69.98 | 684 | |
| 687 | 70.97 | 685 | |
| 688 | 71.96 | 686 | |
| 689 | 72.95 | 687 | ***** |
| 690 | 73.93 | 689 | |
| 691 | 74.91 | 690 | |
| 692 | 75.88 | 691 | |
| 693 | 76.86 | 693 | |
| 694 | 77.83 | 694 | |
| 695 | 78.81 | 696 | ***** |
| 696 | 79.50 | 697 | |
| 697 | 80.18 | 698 | |
| 698 | 80.86 | 699 | |
| 699 | 81.55 | 700 | |
| 700 | 82.23 | 702 | |
| 701 | 82.92 | 703 | |
| 702 | 83.60 | 704 | |
| 703 | 84.29 | 706 | ***** |
| 704 | 84.88 | 707 | |
| 705 | 85.48 | 708 | |
| 706 | 86.08 | 709 | |

(Continued on next page)

Table 54. Grade 4 ELA Linking of 2006 to 2005 Scale Scores (cont.)

| 2006 Scale | Percentile rank | 2005 Scale | 2006 Scale Scores |
|------------|-----------------|------------|-------------------|
| 707 | 86.68 | 710 | |
| 708 | 87.28 | 712 | |
| 709 | 87.88 | 713 | |
| 710 | 88.48 | 715 | |
| 711 | 89.08 | 717 | ***** |
| 712 | 89.47 | 718 | |
| 713 | 89.86 | 720 | |
| 714 | 90.24 | 721 | |
| 715 | 90.63 | 722 | |
| 716 | 91.02 | 723 | |
| 717 | 91.41 | 724 | |
| 718 | 91.79 | 726 | |
| 719 | 92.18 | 727 | |
| 720 | 92.57 | 729 | |
| 721 | 92.96 | 731 | ***** |
| 722 | 93.17 | 732 | |
| 723 | 93.39 | 733 | |
| 724 | 93.61 | 735 | |
| 725 | 93.83 | 736 | |
| 726 | 94.05 | 737 | |
| 727 | 94.27 | 738 | |
| 728 | 94.49 | 739 | |
| 729 | 94.71 | 740 | |
| 730 | 94.93 | 742 | |
| 731 | 95.15 | 743 | |
| 732 | 95.37 | 744 | |
| 733 | 95.59 | 745 | |
| 734 | 95.81 | 746 | |
| 735 | 96.02 | 748 | ***** |
| 736 | 96.14 | 750 | |
| 737 | 96.25 | 751 | |
| 738 | 96.36 | 752 | |
| 739 | 96.47 | 753 | |
| 740 | 96.58 | 754 | |
| 741 | 96.69 | 756 | |
| 742 | 96.80 | 757 | |
| 743 | 96.92 | 758 | |
| 744 | 97.03 | 759 | |
| 745 | 97.14 | 761 | |

(Continued on next page)

Table 54. Grade 4 ELA Linking of 2006 to 2005 Scale Scores (cont.)

| 2006 Scale | Percentile rank | 2005 Scale | 2006 Scale Scores |
|------------|-----------------|------------|-------------------|
| 746 | 97.25 | 762 | |
| 747 | 97.36 | 763 | |
| 748 | 97.47 | 764 | |
| 749 | 97.58 | 766 | |
| 750 | 97.70 | 767 | |
| 751 | 97.81 | 768 | |
| 752 | 97.92 | 769 | |
| 753 | 98.03 | 771 | |
| 754 | 98.14 | 772 | |
| 755 | 98.25 | 773 | |
| 756 | 98.36 | 774 | ***** |
| 757 | 98.43 | 776 | |
| 758 | 98.50 | 777 | |
| 759 | 98.57 | 779 | |
| 760 | 98.64 | 780 | |
| 761 | 98.71 | 781 | |
| 762 | 98.77 | 783 | |
| 763 | 98.84 | 784 | |
| 764 | 98.91 | 786 | |
| 765 | 98.98 | 787 | |
| 766 | 99.05 | 788 | |
| 767 | 99.12 | 790 | |
| 768 | 99.18 | 791 | |
| 769 | 99.25 | 792 | |
| 770 | 99.32 | 794 | |
| 771 | 99.39 | 795 | |
| 772 | 99.46 | 797 | |
| 773 | 99.53 | 798 | |
| 774 | 99.60 | 799 | |
| 775 | 99.66 | 800 | ***** |

Note: ***** denotes 2006 scale scores that are in actual 2006 scoring table.

Table 55. Grade 8 ELA Linking of 2006 to 2005 Scale Scores

| 2006 Scale | Percentile Rank | 2005 Scale | 2005 Scale |
|------------|-----------------|------------|------------|
| 430 | 0.00 | 527 | ***** |
| 431 | 0.00 | 527 | |
| 432 | 0.01 | 528 | |
| 433 | 0.01 | 528 | |
| 434 | 0.01 | 529 | |
| 435 | 0.01 | 529 | |
| 436 | 0.02 | 530 | |
| 437 | 0.02 | 530 | |
| 438 | 0.02 | 531 | |
| 439 | 0.02 | 531 | |
| 440 | 0.03 | 531 | |
| 441 | 0.03 | 532 | |
| 442 | 0.03 | 532 | |
| 443 | 0.03 | 533 | |
| 444 | 0.03 | 533 | |
| 445 | 0.04 | 534 | |
| 446 | 0.04 | 534 | |
| 447 | 0.04 | 535 | |
| 448 | 0.04 | 535 | |
| 449 | 0.05 | 536 | |
| 450 | 0.05 | 536 | |
| 451 | 0.05 | 537 | |
| 452 | 0.05 | 537 | |
| 453 | 0.06 | 538 | |
| 454 | 0.06 | 538 | |
| 455 | 0.06 | 538 | |
| 456 | 0.06 | 539 | |
| 457 | 0.07 | 539 | |
| 458 | 0.07 | 540 | ***** |
| 459 | 0.07 | 540 | |
| 460 | 0.07 | 540 | |
| 461 | 0.07 | 541 | |
| 462 | 0.08 | 541 | |
| 463 | 0.08 | 541 | |
| 464 | 0.08 | 541 | |
| 465 | 0.08 | 542 | |
| 466 | 0.08 | 542 | |
| 467 | 0.08 | 542 | |
| 468 | 0.08 | 543 | |
| 469 | 0.09 | 543 | |
| 470 | 0.09 | 543 | |

(Continued on next page)

Table 55. Grade 8 ELA Linking of 2006 to 2005 Scale Scores (cont.)

| 2006 Scale | Percentile Rank | 2005 Scale | 2005 Scale |
|------------|-----------------|------------|------------|
| 471 | 0.09 | 543 | |
| 472 | 0.09 | 544 | |
| 473 | 0.09 | 544 | |
| 474 | 0.09 | 544 | |
| 475 | 0.09 | 544 | |
| 476 | 0.09 | 545 | |
| 477 | 0.10 | 545 | |
| 478 | 0.10 | 545 | |
| 479 | 0.10 | 545 | |
| 480 | 0.10 | 546 | |
| 481 | 0.10 | 546 | |
| 482 | 0.10 | 546 | |
| 483 | 0.10 | 546 | |
| 484 | 0.11 | 547 | |
| 485 | 0.11 | 547 | |
| 486 | 0.11 | 547 | |
| 487 | 0.11 | 548 | |
| 488 | 0.11 | 548 | |
| 489 | 0.11 | 548 | |
| 490 | 0.11 | 548 | |
| 491 | 0.12 | 549 | |
| 492 | 0.12 | 549 | |
| 493 | 0.12 | 549 | |
| 494 | 0.12 | 549 | |
| 495 | 0.12 | 550 | |
| 496 | 0.12 | 550 | |
| 497 | 0.12 | 550 | |
| 498 | 0.13 | 550 | |
| 499 | 0.13 | 551 | ***** |
| 500 | 0.13 | 552 | |
| 501 | 0.14 | 553 | |
| 502 | 0.14 | 554 | |
| 503 | 0.15 | 555 | |
| 504 | 0.15 | 555 | |
| 505 | 0.16 | 556 | |
| 506 | 0.16 | 557 | |
| 507 | 0.17 | 558 | |
| 508 | 0.17 | 559 | |
| 509 | 0.18 | 560 | |
| 510 | 0.18 | 561 | |

(Continued on next page)

Table 55. Grade 8 ELA Linking of 2006 to 2005 Scale Scores (cont.)

| 2006 Scale | Percentile rank | 2005 Scale | 2006 Scale Scores |
|------------|-----------------|------------|-------------------|
| 511 | 0.19 | 562 | |
| 512 | 0.19 | 563 | |
| 513 | 0.20 | 564 | |
| 514 | 0.20 | 565 | |
| 515 | 0.21 | 566 | |
| 516 | 0.21 | 567 | ***** |
| 517 | 0.22 | 569 | |
| 518 | 0.23 | 571 | |
| 519 | 0.24 | 573 | |
| 520 | 0.25 | 574 | |
| 521 | 0.26 | 576 | |
| 522 | 0.27 | 578 | |
| 523 | 0.28 | 580 | |
| 524 | 0.29 | 582 | |
| 525 | 0.30 | 584 | |
| 526 | 0.31 | 585 | |
| 527 | 0.32 | 586 | |
| 528 | 0.33 | 587 | |
| 529 | 0.34 | 588 | ***** |
| 530 | 0.36 | 590 | |
| 531 | 0.38 | 593 | |
| 532 | 0.40 | 595 | |
| 533 | 0.42 | 597 | |
| 534 | 0.44 | 599 | |
| 535 | 0.46 | 601 | |
| 536 | 0.48 | 604 | |
| 537 | 0.50 | 606 | |
| 538 | 0.52 | 608 | ***** |
| 539 | 0.54 | 609 | |
| 540 | 0.57 | 610 | |
| 541 | 0.59 | 611 | |
| 542 | 0.62 | 612 | |
| 543 | 0.65 | 613 | |
| 544 | 0.67 | 614 | |
| 545 | 0.70 | 615 | |
| 546 | 0.72 | 617 | |
| 547 | 0.75 | 618 | ***** |
| 548 | 0.79 | 619 | |
| 549 | 0.84 | 621 | |

(Continued on next page)

Table 55. Grade 8 ELA Linking of 2006 to 2005 Scale Scores (cont.)

| 2006 Scale | Percentile rank | 2005 Scale | 2006 Scale Scores |
|------------|-----------------|------------|-------------------|
| 550 | 0.88 | 621 | |
| 551 | 0.92 | 622 | |
| 552 | 0.97 | 623 | |
| 553 | 1.01 | 624 | |
| 554 | 1.05 | 625 | ***** |
| 555 | 1.12 | 626 | |
| 556 | 1.18 | 628 | |
| 557 | 1.24 | 629 | |
| 558 | 1.30 | 629 | |
| 559 | 1.37 | 630 | |
| 560 | 1.43 | 631 | ***** |
| 561 | 1.50 | 632 | |
| 562 | 1.58 | 633 | |
| 563 | 1.66 | 634 | |
| 564 | 1.73 | 634 | |
| 565 | 1.81 | 635 | |
| 566 | 1.89 | 636 | ***** |
| 567 | 1.98 | 637 | |
| 568 | 2.08 | 638 | |
| 569 | 2.17 | 639 | |
| 570 | 2.27 | 640 | |
| 571 | 2.36 | 640 | |
| 572 | 2.45 | 641 | ***** |
| 573 | 2.59 | 642 | |
| 574 | 2.72 | 642 | |
| 575 | 2.86 | 643 | |
| 576 | 2.99 | 644 | |
| 577 | 3.13 | 644 | ***** |
| 578 | 3.29 | 645 | |
| 579 | 3.45 | 646 | |
| 580 | 3.61 | 647 | |
| 581 | 3.77 | 647 | |
| 582 | 3.93 | 648 | ***** |
| 583 | 4.16 | 649 | |
| 584 | 4.40 | 649 | |
| 585 | 4.63 | 650 | |
| 586 | 4.87 | 651 | ***** |
| 587 | 5.14 | 652 | |
| 588 | 5.41 | 653 | |

(Continued on next page)

Table 55. Grade 8 ELA Linking of 2006 to 2005 Scale Scores (cont.)

| 2006 Scale | Percentile rank | 2005 Scale | 2006 Scale Scores |
|------------|-----------------|------------|-------------------|
| 589 | 5.68 | 654 | |
| 590 | 5.95 | 655 | ***** |
| 591 | 6.26 | 656 | |
| 592 | 6.57 | 656 | |
| 593 | 6.89 | 657 | |
| 594 | 7.20 | 658 | ***** |
| 595 | 7.56 | 658 | |
| 596 | 7.92 | 659 | |
| 597 | 8.29 | 659 | |
| 598 | 8.65 | 660 | ***** |
| 599 | 9.06 | 661 | |
| 600 | 9.47 | 662 | |
| 601 | 9.89 | 662 | |
| 602 | 10.30 | 663 | ***** |
| 603 | 10.77 | 664 | |
| 604 | 11.24 | 664 | |
| 605 | 11.71 | 665 | |
| 606 | 12.18 | 665 | ***** |
| 607 | 12.72 | 666 | |
| 608 | 13.25 | 667 | |
| 609 | 13.79 | 668 | |
| 610 | 14.32 | 668 | ***** |
| 611 | 14.94 | 669 | |
| 612 | 15.55 | 670 | |
| 613 | 16.16 | 670 | |
| 614 | 16.77 | 671 | ***** |
| 615 | 17.45 | 672 | |
| 616 | 18.13 | 673 | |
| 617 | 18.81 | 673 | |
| 618 | 19.49 | 674 | ***** |
| 619 | 20.26 | 674 | |
| 620 | 21.03 | 675 | |
| 621 | 21.79 | 676 | |
| 622 | 22.56 | 676 | ***** |
| 623 | 23.43 | 677 | |
| 624 | 24.30 | 678 | |
| 625 | 25.17 | 678 | |
| 626 | 26.03 | 679 | ***** |
| 627 | 26.98 | 680 | |

(Continued on next page)

Table 55. Grade 8 ELA Linking of 2006 to 2005 Scale Scores (cont.)

| 2006 Scale | Percentile rank | 2005 Scale | 2006 Scale Scores |
|------------|-----------------|------------|-------------------|
| 628 | 27.93 | 681 | |
| 629 | 28.88 | 681 | |
| 630 | 29.83 | 682 | ***** |
| 631 | 30.87 | 682 | |
| 632 | 31.90 | 683 | |
| 633 | 32.94 | 684 | |
| 634 | 33.98 | 684 | ***** |
| 635 | 35.09 | 685 | |
| 636 | 36.20 | 686 | |
| 637 | 37.31 | 687 | |
| 638 | 38.42 | 688 | ***** |
| 639 | 39.36 | 688 | |
| 640 | 40.31 | 689 | |
| 641 | 41.25 | 689 | |
| 642 | 42.19 | 690 | |
| 643 | 43.14 | 690 | ***** |
| 644 | 44.38 | 691 | |
| 645 | 45.62 | 692 | |
| 646 | 46.86 | 693 | |
| 647 | 48.10 | 694 | ***** |
| 648 | 49.13 | 695 | |
| 649 | 50.16 | 695 | |
| 650 | 51.18 | 696 | |
| 651 | 52.21 | 697 | |
| 652 | 53.24 | 697 | ***** |
| 653 | 54.31 | 698 | |
| 654 | 55.37 | 699 | |
| 655 | 56.43 | 699 | |
| 656 | 57.50 | 700 | |
| 657 | 58.56 | 701 | ***** |
| 658 | 59.63 | 701 | |
| 659 | 60.71 | 702 | |
| 660 | 61.78 | 703 | |
| 661 | 62.85 | 704 | |
| 662 | 63.93 | 705 | ***** |
| 663 | 64.82 | 705 | |
| 664 | 65.71 | 706 | |
| 665 | 66.60 | 707 | |
| 666 | 67.49 | 707 | |

(Continued on next page)

Table 55. Grade 8 ELA Linking of 2006 to 2005 Scale Scores (cont.)

| 2006 Scale | Percentile rank | 2005 Scale | 2006 Scale Scores |
|------------|-----------------|------------|-------------------|
| 667 | 68.38 | 708 | |
| 668 | 69.27 | 709 | ***** |
| 669 | 70.15 | 709 | |
| 670 | 71.03 | 710 | |
| 671 | 71.92 | 711 | |
| 672 | 72.80 | 712 | |
| 673 | 73.69 | 713 | |
| 674 | 74.57 | 714 | ***** |
| 675 | 75.30 | 715 | |
| 676 | 76.04 | 715 | |
| 677 | 76.77 | 716 | |
| 678 | 77.51 | 717 | |
| 679 | 78.25 | 718 | |
| 680 | 78.98 | 719 | |
| 681 | 79.72 | 720 | ***** |
| 682 | 80.33 | 721 | |
| 683 | 80.95 | 722 | |
| 684 | 81.56 | 723 | |
| 685 | 82.18 | 723 | |
| 686 | 82.79 | 724 | |
| 687 | 83.41 | 725 | |
| 688 | 84.03 | 726 | |
| 689 | 84.64 | 727 | ***** |
| 690 | 85.15 | 728 | |
| 691 | 85.66 | 729 | |
| 692 | 86.17 | 730 | |
| 693 | 86.68 | 731 | |
| 694 | 87.19 | 732 | |
| 695 | 87.70 | 733 | |
| 696 | 88.22 | 734 | |
| 697 | 88.73 | 735 | |
| 698 | 89.24 | 736 | ***** |
| 699 | 89.57 | 737 | |
| 700 | 89.91 | 738 | |
| 701 | 90.25 | 739 | |
| 702 | 90.59 | 740 | |
| 703 | 90.92 | 740 | |
| 704 | 91.26 | 741 | |
| 705 | 91.60 | 742 | |

(Continued on next page)

Table 55. Grade 8 ELA Linking of 2006 to 2005 Scale Scores (cont.)

| 2006 Scale | Percentile rank | 2005 Scale | 2006 Scale Scores |
|------------|-----------------|------------|-------------------|
| 706 | 91.94 | 743 | |
| 707 | 92.27 | 744 | |
| 708 | 92.61 | 745 | |
| 709 | 92.95 | 746 | |
| 710 | 93.28 | 747 | ***** |
| 711 | 93.47 | 748 | |
| 712 | 93.65 | 749 | |
| 713 | 93.83 | 749 | |
| 714 | 94.01 | 750 | |
| 715 | 94.19 | 751 | |
| 716 | 94.38 | 751 | |
| 717 | 94.56 | 752 | |
| 718 | 94.74 | 753 | |
| 719 | 94.92 | 754 | |
| 720 | 95.10 | 754 | |
| 721 | 95.28 | 755 | |
| 722 | 95.47 | 756 | |
| 723 | 95.65 | 756 | |
| 724 | 95.83 | 757 | |
| 725 | 96.01 | 758 | |
| 726 | 96.19 | 760 | |
| 727 | 96.37 | 761 | |
| 728 | 96.56 | 762 | ***** |
| 729 | 96.62 | 763 | |
| 730 | 96.69 | 763 | |
| 731 | 96.76 | 764 | |
| 732 | 96.83 | 764 | |
| 733 | 96.90 | 765 | |
| 734 | 96.97 | 765 | |
| 735 | 97.03 | 766 | |
| 736 | 97.10 | 766 | |
| 737 | 97.17 | 767 | |
| 738 | 97.24 | 767 | |
| 739 | 97.31 | 768 | |
| 740 | 97.38 | 768 | |
| 741 | 97.45 | 769 | |
| 742 | 97.51 | 769 | |
| 743 | 97.58 | 770 | |
| 744 | 97.65 | 770 | |

(Continued on next page)

Table 55. Grade 8 ELA Linking of 2006 to 2005 Scale Scores (cont.)

| 2006 Scale | Percentile rank | 2005 Scale | 2006 Scale Scores |
|------------|-----------------|------------|-------------------|
| 745 | 97.72 | 771 | |
| 746 | 97.79 | 771 | |
| 747 | 97.86 | 772 | |
| 748 | 97.93 | 772 | |
| 749 | 97.99 | 773 | |
| 750 | 98.06 | 773 | |
| 751 | 98.13 | 774 | |
| 752 | 98.20 | 775 | |
| 753 | 98.27 | 778 | |
| 754 | 98.34 | 781 | |
| 755 | 98.40 | 783 | |
| 756 | 98.47 | 786 | |
| 757 | 98.54 | 789 | |
| 758 | 98.61 | 792 | |
| 759 | 98.68 | 794 | |
| 760 | 98.75 | 797 | ***** |
| 761 | 98.78 | 799 | |
| 762 | 98.82 | 800 | |
| 763 | 98.85 | 801 | |
| 764 | 98.88 | 803 | |
| 765 | 98.92 | 804 | |
| 766 | 98.95 | 805 | |
| 767 | 98.99 | 807 | |
| 768 | 99.02 | 808 | |
| 769 | 99.05 | 810 | |
| 770 | 99.09 | 811 | |
| 771 | 99.12 | 812 | |
| 772 | 99.16 | 814 | |
| 773 | 99.19 | 815 | |
| 774 | 99.22 | 816 | |
| 775 | 99.26 | 818 | |
| 776 | 99.29 | 819 | |
| 777 | 99.33 | 821 | |
| 778 | 99.36 | 822 | |
| 779 | 99.39 | 823 | |
| 780 | 99.43 | 825 | |
| 781 | 99.46 | 826 | |
| 782 | 99.50 | 828 | |
| 783 | 99.53 | 829 | |
| 784-790 | 99.56 | 830 | ***** |

Note: ***** denotes 2006 scale scores that are in actual 2006 scoring table.

After linking the 2006 to 2005 tests, the statistical properties of the 2005 and linked 2006 tests were evaluated in terms of test mean, standard deviation, skewness and kurtosis. The results for both grades are presented below.

Table 56. ELA Grade 4 Equipercentile Linking Summary

| ELA 4 Test | Mean | Standard Deviation | Skewness | Kurtosis |
|-------------------------|--------|--------------------|----------|----------|
| 2005 Test | 664.57 | 44.09 | -0.01 | 4.59 |
| 2006 Test on 2006 Scale | 665.73 | 40.75 | -0.50 | 4.86 |
| 2006 Test on 2005 Scale | 664.66 | 44.14 | 0.01 | 4.40 |

Table 57. ELA Grade 8 Equipercentile Linking Summary

| ELA 8 Test | Mean | Standard Deviation | Skewness | Kurtosis |
|-------------------------|--------|--------------------|----------|----------|
| 2005 Test | 697.57 | 32.56 | 0.35 | 6.30 |
| 2006 Test on 2006 Scale | 650.14 | 40.78 | 0.10 | 4.23 |
| 2006 Test on 2005 Scale | 697.76 | 32.60 | 0.42 | 5.52 |

As shown in Tables 56 and 57, the statistical properties of 2005 tests and 2006 tests on the 2005 scale were in expected alignment.

A shortcoming of an equipercentile procedure is an assumption of no growth. However, using the equipercentile procedure for linking New York State 2005 and 2006 ELA grades 4 and 8 tests was justified to a large degree by the New York State NAEP testing results from 2003-2005 that indicated very little or no growth in Reading and Mathematics areas for New York State grade 4 and 8 students. For more details on New York State student performance on NAEP assessment please follow the link: <http://nces.ed.gov/nationsreportcard/states/>.

Section XI: 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.
- 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.
- 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 – ELA Passage Specifications

General Guidelines

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

Table A1. Number, Type, and Length of Passages

| Grade | # of Listening Passages | Approximate Word Length | # of Reading Passages | Passage Types | Approximate Word Length | Passage Types |
|--------------|--------------------------------|--------------------------------|--|---|--------------------------------|------------------------------------|
| 3 | 8 | 200-400 | 20 (includes 5 sets of short paired passages) | Literary | 200-600 | 50% Literary; 50% Informational |
| 4 | 5 | 250-450 | 20 (includes 8 sets of short paired passages) | Literary | 250-600 | 50% Literary; 50% Informational |
| 5 | 12 | 300-500 | 20 (includes 5 sets of short paired passages) | Literary | 250-600 | 50% Literary; 50% Informational |
| 6 | 8 | 350-550 | 24 (includes 5 sets of short paired passages) | Informational | 300-650 | 50% Literary; 50% Informational |
| 7 | 8 | 400-600 | 24 (includes 5 sets of short paired passages) | Informational (May be 2 short paired pieces) | 350-700 | 50% Literary; 50% Informational |
| 8 | 5 | 450-650 | 20 (includes 8 sets of short paired passages) | Informational (May be 2 short paired pieces) | 350-800 | 50% Literary; 50% Informational |

Appendices: Appendix B – Criteria for Item Acceptability

For Multiple-Choice Items:

Check that the content of each item

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

Check that the format of each item

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

Also check that

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

For Constructed-Response Items:**Check that the content of each item is**

- designed to assess the targeted performance indicator
- appropriate for the grade level being tested
- presented at a reading level suitable for the grade level being tested
- appropriate in context
- written so that a student possessing knowledge or skill being tested can construct a response that 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 C – 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.

Here are general guidelines for item selection:

- Satisfy the content specifications in terms of objective coverage and the number and percent of MC and CR items on the test. An often used criterion for objective coverage is within 5% 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 D – Factor Analysis Results

As described in Section III (Validity) a Principal Component factor analysis was conducted on the Grades 3-8 ELA Tests data. The analyses were conducted for the total population of students and selected subpopulations: Limited English Proficiency (LEP), Students with Disabilities (SWD), and students using accommodations (SUA). This Appendix contains figures of scree plots obtained from the analysis of the total population and subpopulation data ELA data and a table of eigenvalues and proportion of variance accounted for by extracted factors for subgroups.

Figure D1. Grade 3 ELA Scree Plot (Total Population)

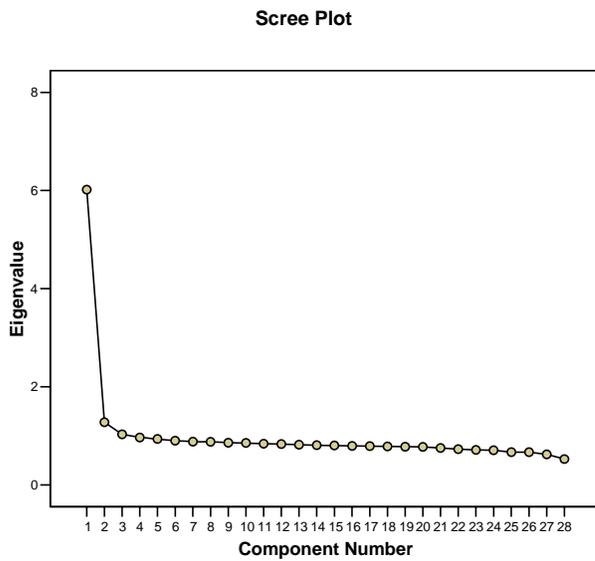


Figure D2. Grade 3 Scree Plot (LEP Students)

Scree Plot

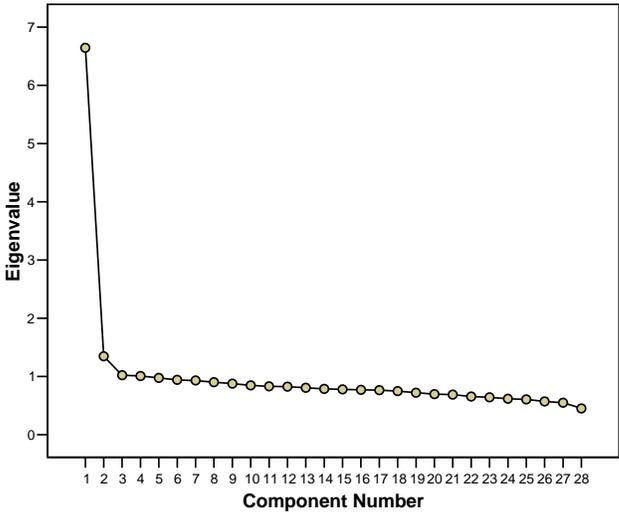


Figure D3. Grade 3 Scree Plot (Students with Disabilities)

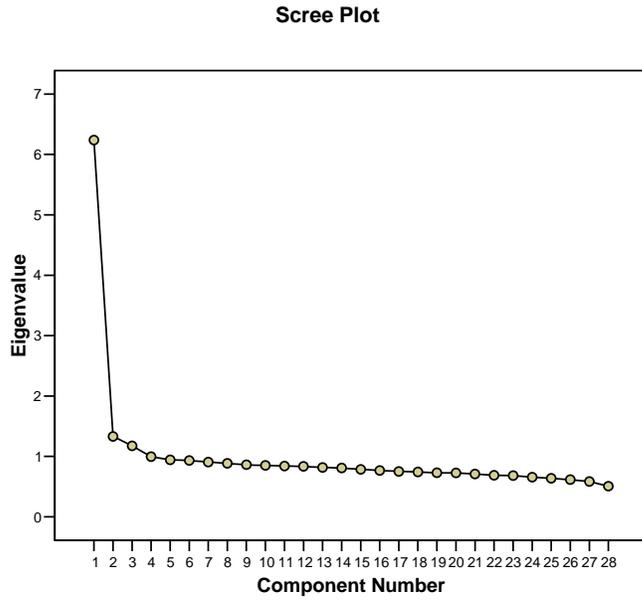


Figure D4. Grade 3 Scree Plot (Students using Accommodations)

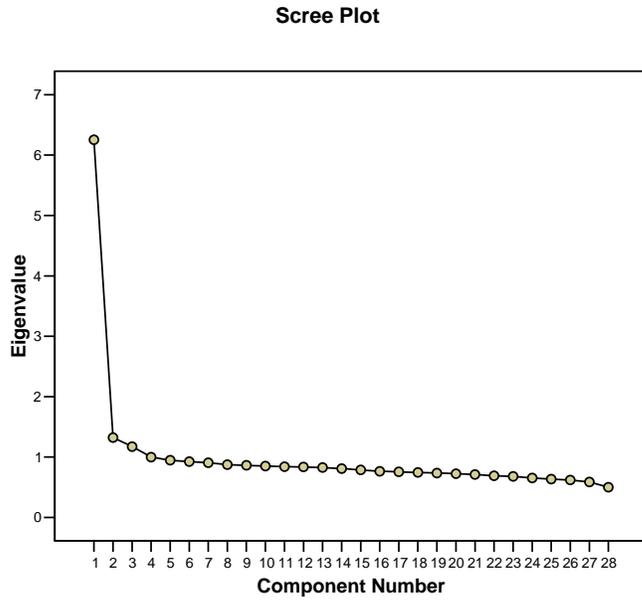


Figure D5. Grade 4 Scree Plot (Total Population)

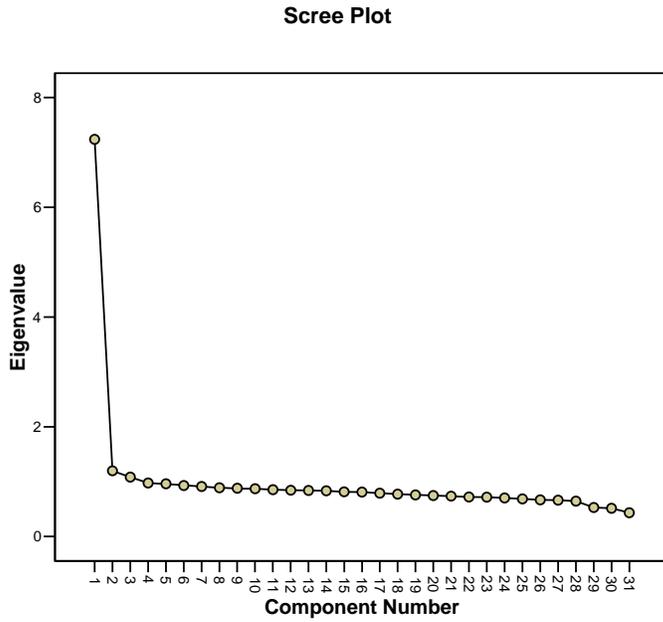


Figure D6. Grade 4 Scree Plot (LEP Students)

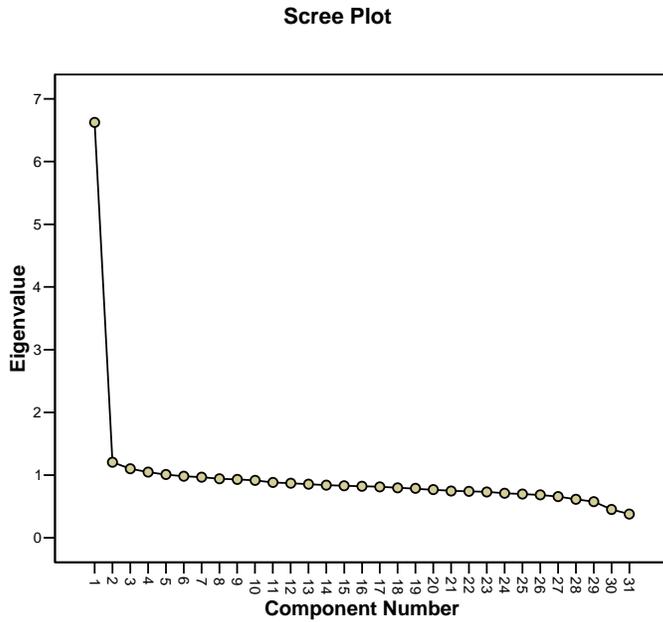


Figure D7. Grade 4 Scree Plot (Students with Disabilities)

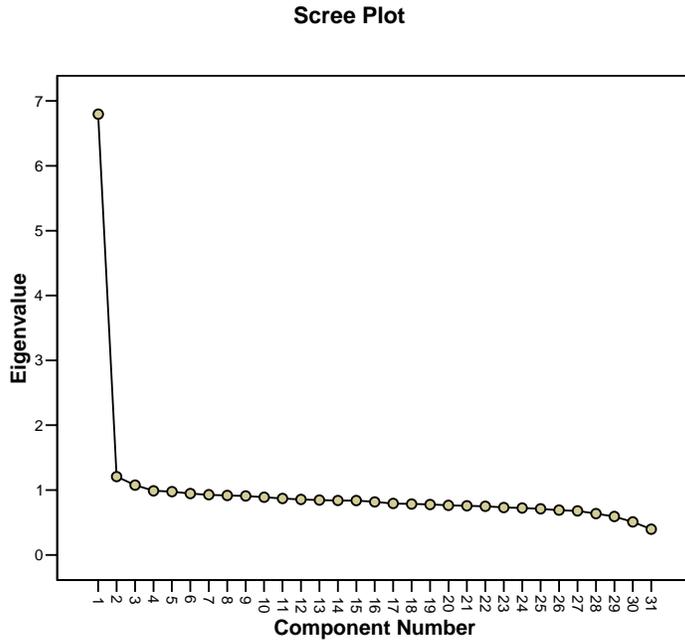


Figure D8. Grade 4 Scree Plot (Students using Accommodations)

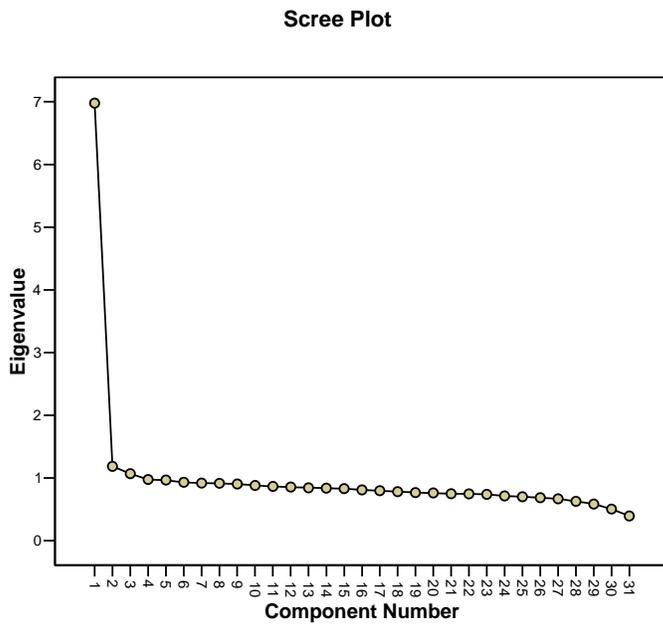


Figure D9. Grade 5 Scree Plot (Total Population)

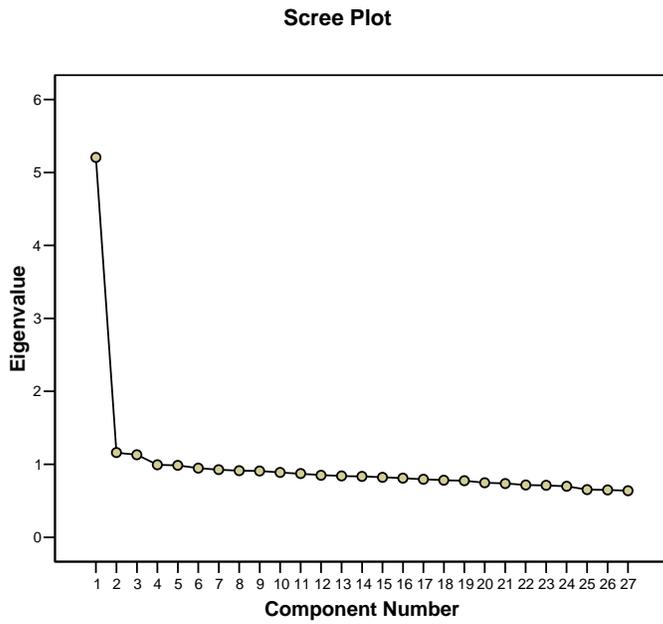


Figure D10. Grade 5 Scree Plot (LEP Students)

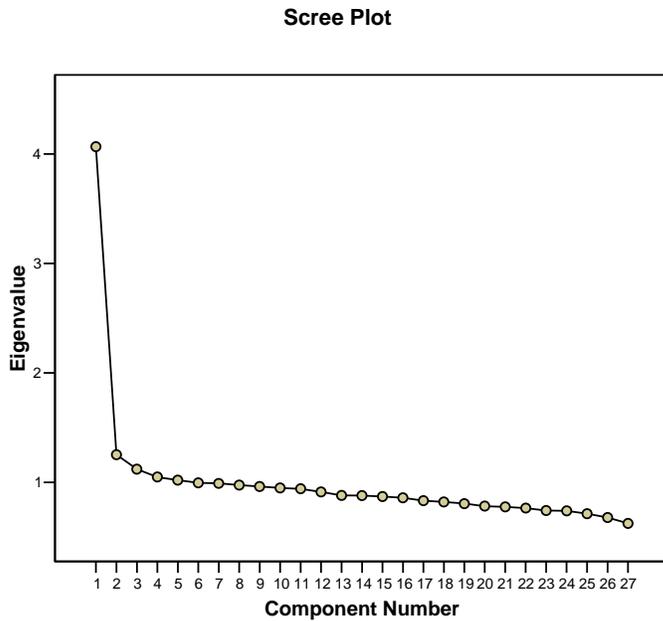


Figure D11. Grade 5 Scree Plot (Students with Disabilities)

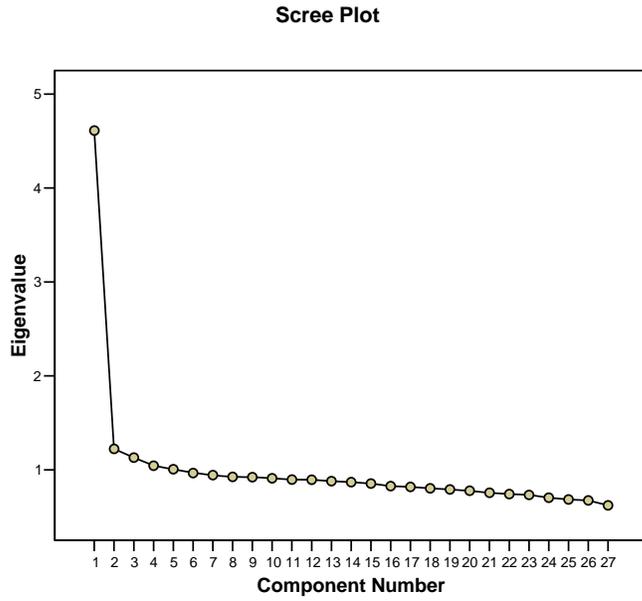


Figure D12. Grade 5 Scree Plot (Students using Accommodations)

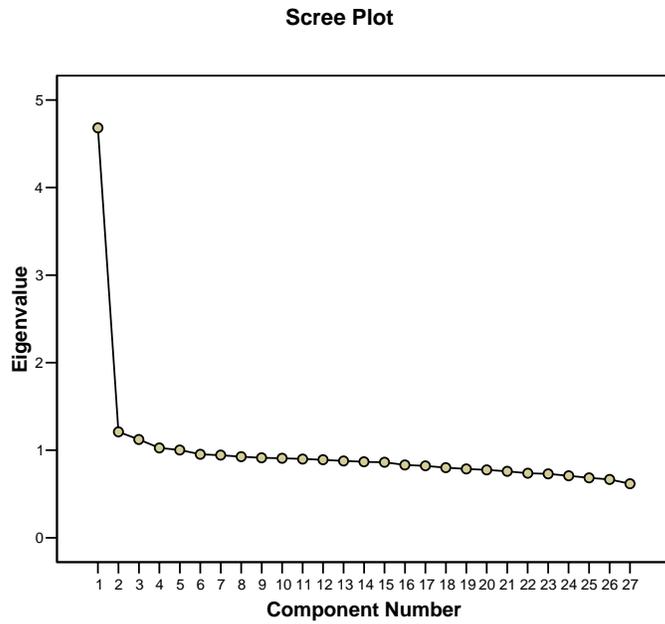


Figure D13. Grade 6 Scree Plot (Total Population)

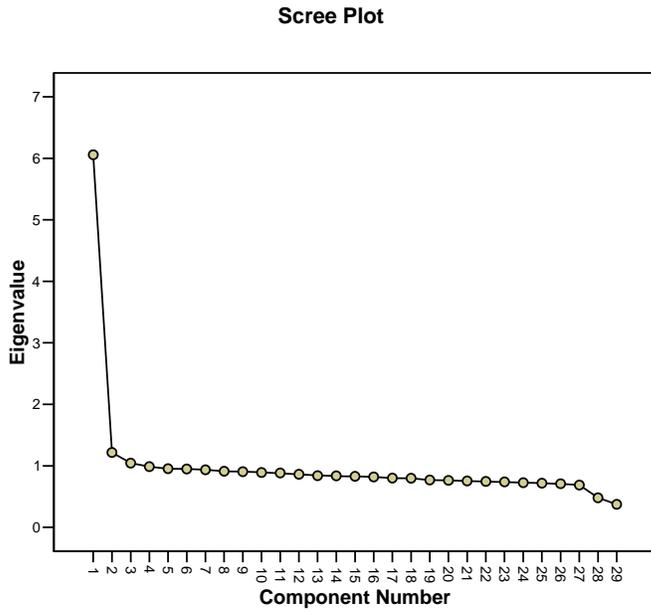


Figure D14. Grade 6 Scree Plot (LEP Students)

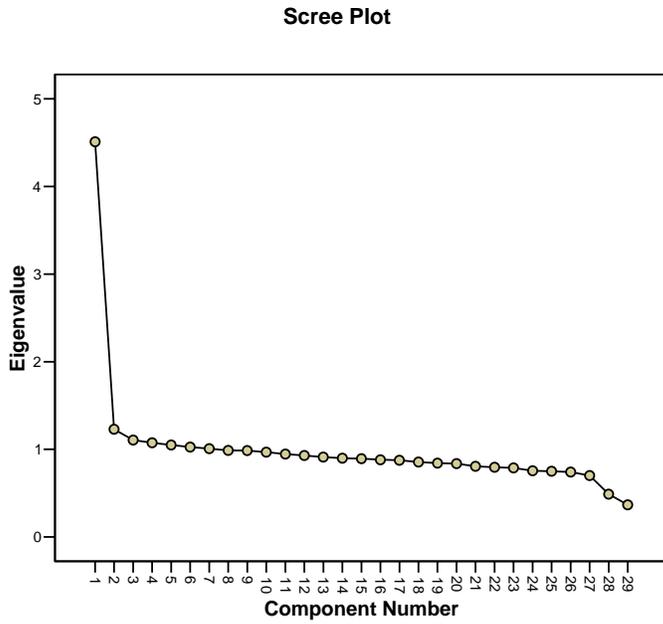


Figure D15. Grade 6 Scree Plot (Students with Disabilities)

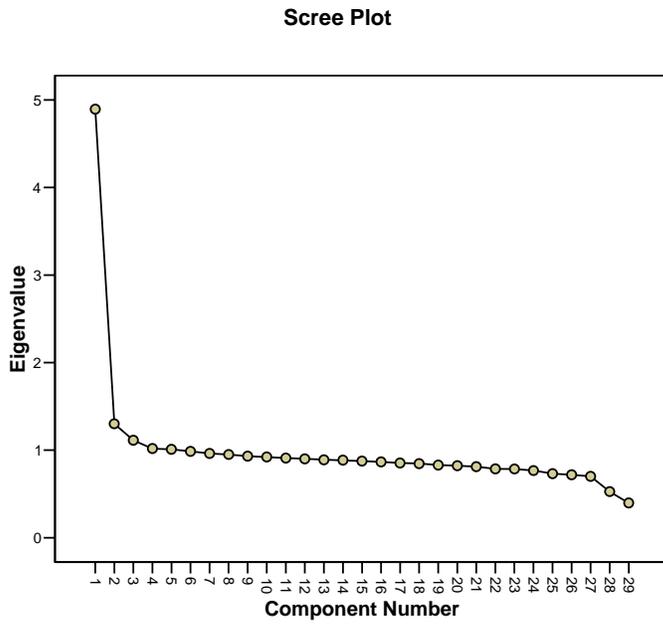


Figure D16. Grade 6 Scree Plot (Students using Accommodations)

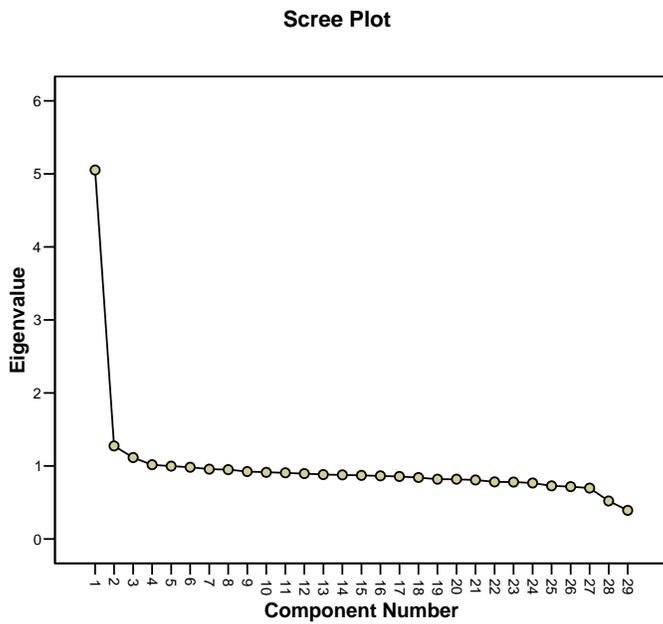


Figure D17. Grade 7 Scree Plot (Total Population)

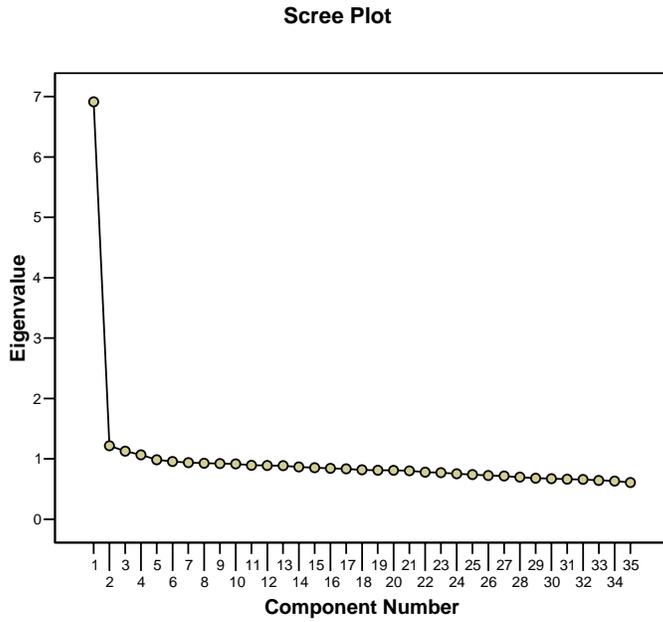


Figure D18. Grade 7 Scree Plot (LEP Students)

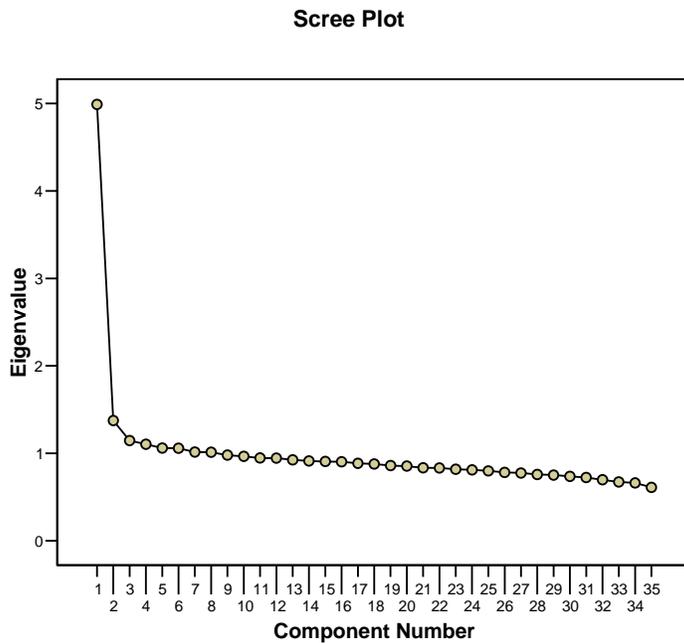


Figure D19. Grade 7 Scree Plot (Students with Disabilities)

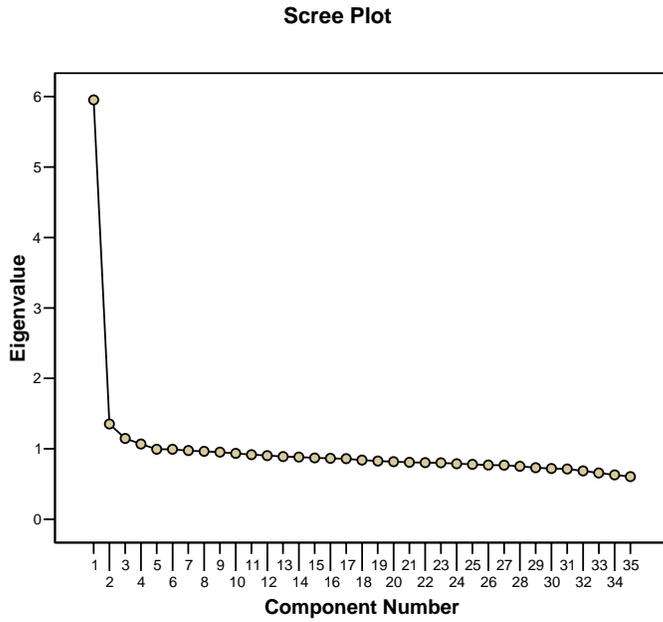


Figure D20. Grade 7 Scree Plot (Students using Accommodations)

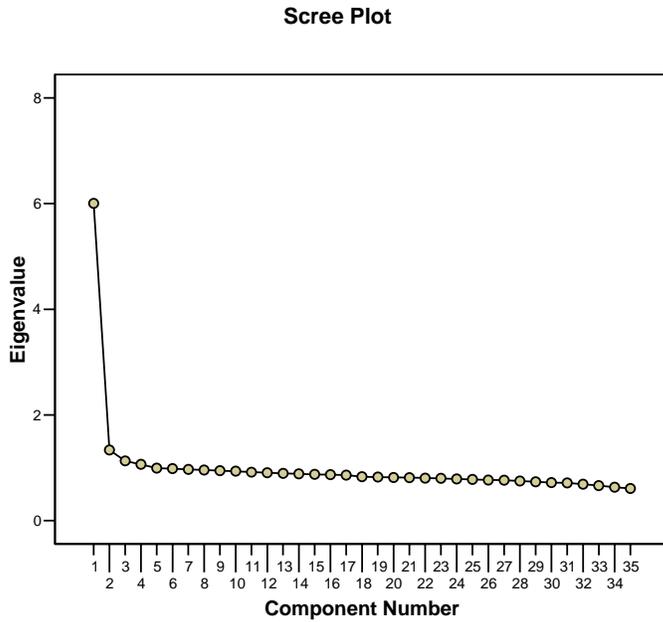


Figure D21. Grade 8 Scree Plot (Total Population)

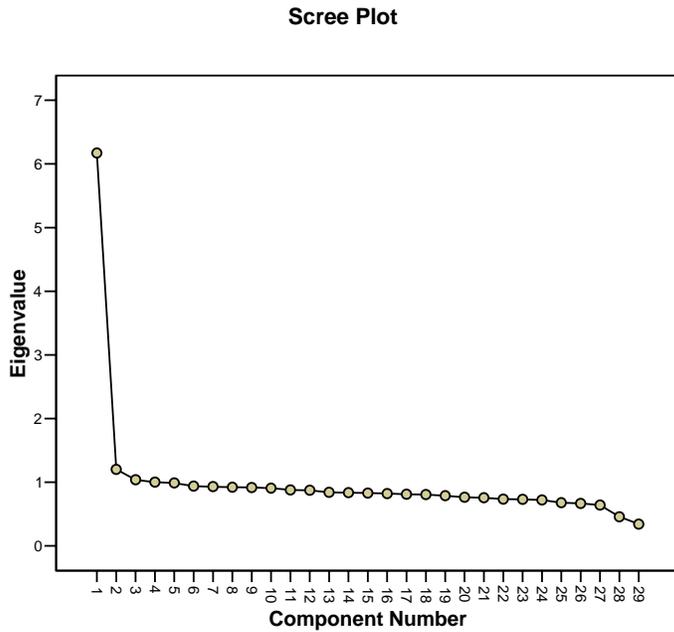


Figure D22. Grade 8 Scree Plot (LEP Students)

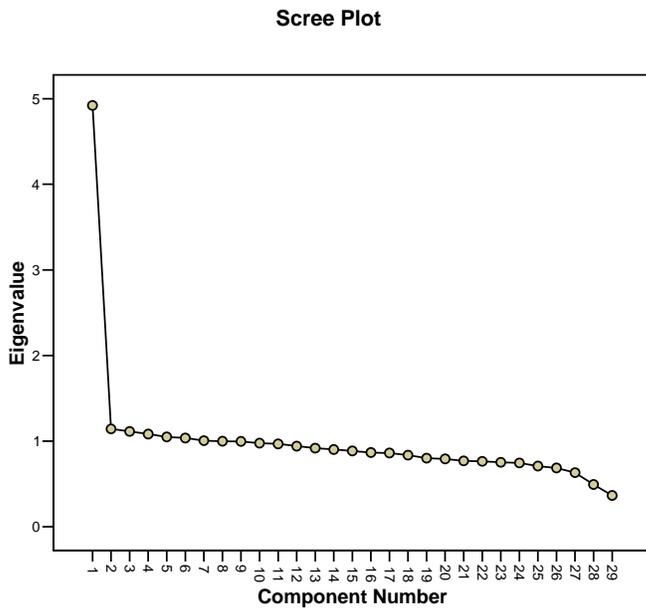


Figure D23. Grade 8 Scree Plot (Students with Disabilities)

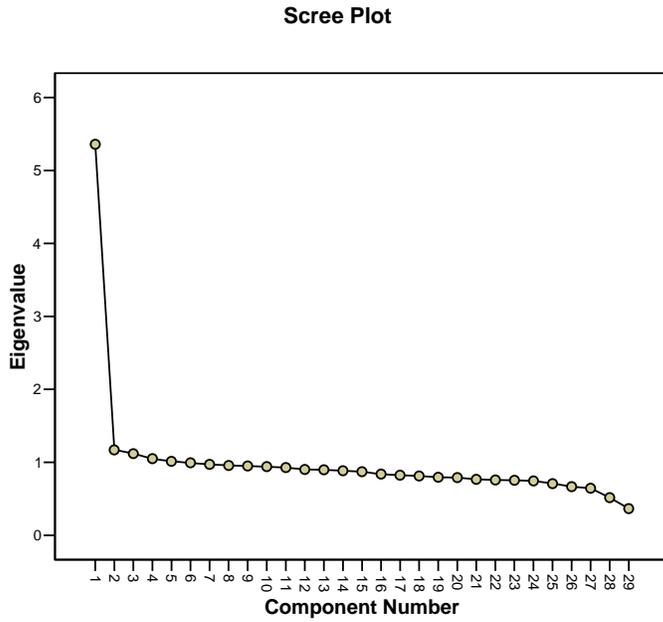


Figure D24. Grade 8 Scree Plot (Students using Accommodations)

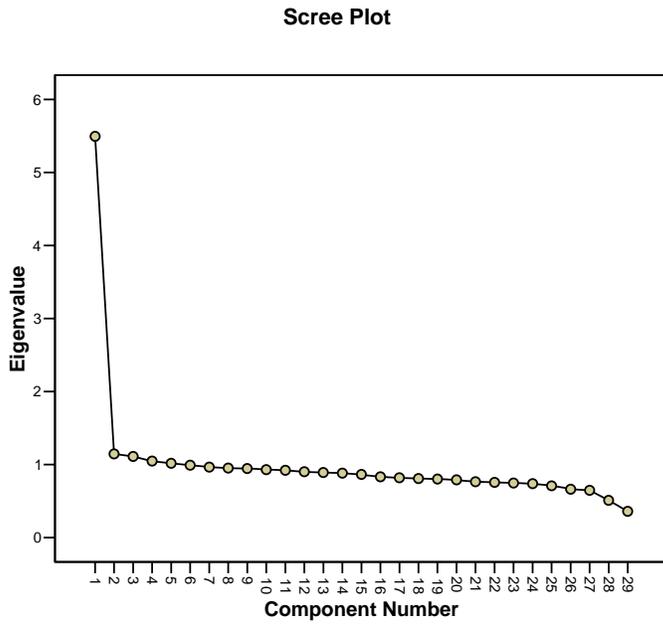


Table D1. Factor Analysis Results for ELA tests (Selected Sub-Populations)

| Grade | Subgroup | Initial Eigenvalues | | | |
|-------|----------|---------------------|-------------|---------------|--------------|
| | | Component | Total | % of Variance | Cumulative % |
| 3 | LEP | 1 | 6.64 | 23.73 | 23.73 |
| | | 2 | 1.35 | 4.81 | 28.54 |
| | | 3 | 1.02 | 3.65 | 32.19 |
| | | 4 | 1.01 | 3.60 | 35.79 |
| | SWD | 1 | 6.24 | 22.28 | 22.28 |
| | | 2 | 1.33 | 4.76 | 27.04 |
| | | 3 | 1.18 | 4.20 | 31.23 |
| | SUA | 1 | 6.25 | 22.33 | 22.33 |
| | | 2 | 1.32 | 4.72 | 27.05 |
| 3 | | 1.17 | 4.19 | 31.24 | |
| 4 | LEP | 1 | 6.63 | 21.37 | 21.37 |
| | | 2 | 1.21 | 3.89 | 25.27 |
| | | 3 | 1.10 | 3.56 | 28.82 |
| | | 4 | 1.05 | 3.38 | 32.21 |
| | | 5 | 1.01 | 3.26 | 35.46 |
| | SWD | 1 | 6.80 | 21.92 | 21.92 |
| | | 2 | 1.21 | 3.90 | 25.82 |
| | | 3 | 1.08 | 3.47 | 29.29 |
| | SUA | 1 | 6.98 | 22.51 | 22.51 |
| 2 | | 1.19 | 3.82 | 26.33 | |
| 3 | | 1.07 | 3.45 | 29.78 | |
| 5 | LEP | 1 | 4.07 | 15.06 | 15.06 |
| | | 2 | 1.25 | 4.64 | 19.70 |
| | | 3 | 1.12 | 4.15 | 23.85 |
| | | 4 | 1.05 | 3.88 | 27.73 |
| | | 5 | 1.02 | 3.78 | 31.51 |
| | SWD | 1 | 4.61 | 17.08 | 17.08 |
| | | 2 | 1.22 | 4.53 | 21.61 |
| | | 3 | 1.13 | 4.19 | 25.79 |
| | | 4 | 1.04 | 3.87 | 29.66 |
| | | 5 | 1.01 | 3.73 | 33.39 |
| | SUA | 1 | 4.68 | 17.34 | 17.34 |
| | | 2 | 1.21 | 4.48 | 21.82 |
| | | 3 | 1.12 | 4.16 | 25.98 |
| | | 4 | 1.03 | 3.80 | 29.77 |
| | | 5 | 1.00 | 3.71 | 33.49 |

(Continued on next page)

Table D1. Factor Analysis Results for ELA tests (Selected Sub-Populations) (cont.)

| Grade | Subgroup | Initial Eigenvalues | | | |
|-------|----------|---------------------|--------------|---------------|--------------|
| | | Component | Total | % of Variance | Cumulative % |
| 6 | LEP | 1 | 4.51 | 15.55 | 15.55 |
| | | 2 | 1.23 | 4.24 | 19.79 |
| | | 3 | 1.11 | 3.81 | 23.61 |
| | | 4 | 1.08 | 3.71 | 27.31 |
| | | 5 | 1.05 | 3.62 | 30.93 |
| | | 6 | 1.03 | 3.54 | 34.47 |
| | | 7 | 1.01 | 3.47 | 37.94 |
| | SWD | 1 | 4.89 | 16.88 | 16.88 |
| | | 2 | 1.30 | 4.49 | 21.36 |
| | | 3 | 1.11 | 3.84 | 25.20 |
| | | 4 | 1.02 | 3.51 | 28.71 |
| | | 5 | 1.01 | 3.48 | 32.19 |
| SUA | 1 | 5.05 | 17.43 | 17.43 | |
| | 2 | 1.28 | 4.40 | 21.82 | |
| | 3 | 1.12 | 3.84 | 25.66 | |
| | 4 | 1.02 | 3.51 | 29.17 | |
| 7 | LEP | 1 | 4.99 | 14.25 | 14.25 |
| | | 2 | 1.38 | 3.93 | 18.18 |
| | | 3 | 1.15 | 3.28 | 21.46 |
| | | 4 | 1.10 | 3.16 | 24.61 |
| | | 5 | 1.06 | 3.03 | 27.65 |
| | | 6 | 1.06 | 3.03 | 30.67 |
| | | 7 | 1.02 | 2.90 | 33.57 |
| | | 8 | 1.01 | 2.89 | 36.46 |
| | SWD | 1 | 5.95 | 17.01 | 17.01 |
| | | 2 | 1.35 | 3.87 | 20.87 |
| | | 3 | 1.15 | 3.28 | 24.15 |
| | | 4 | 1.07 | 3.05 | 27.20 |
| | SUA | 1 | 6.00 | 17.15 | 17.15 |
| | | 2 | 1.34 | 3.82 | 20.97 |
| | | 3 | 1.13 | 3.23 | 24.20 |
| | | 4 | 1.06 | 3.04 | 27.24 |

(Continued on next page)

Table D1. Factor Analysis Results for ELA tests (Selected Sub-Populations) (cont.)

| Grade | Subgroup | Initial Eigenvalues | | | |
|-------|----------|---------------------|-------------|---------------|--------------|
| | | Component | Total | % of Variance | Cumulative % |
| 8 | LEP | 1 | 4.92 | 16.97 | 16.97 |
| | | 2 | 1.14 | 3.94 | 20.91 |
| | | 3 | 1.11 | 3.84 | 24.75 |
| | | 4 | 1.08 | 3.73 | 28.48 |
| | | 5 | 1.05 | 3.62 | 32.10 |
| | | 6 | 1.04 | 3.57 | 35.67 |
| | | 7 | 1.01 | 3.47 | 39.14 |
| | SWD | 1 | 5.36 | 18.48 | 18.48 |
| | | 2 | 1.17 | 4.04 | 22.52 |
| | | 3 | 1.12 | 3.86 | 26.38 |
| | | 4 | 1.05 | 3.62 | 30.00 |
| | | 5 | 1.01 | 3.49 | 33.49 |
| | SUA | 1 | 5.50 | 18.95 | 18.95 |
| | | 2 | 1.15 | 3.95 | 22.90 |
| | | 3 | 1.11 | 3.83 | 26.72 |
| | | 4 | 1.05 | 3.61 | 30.33 |
| | | 5 | 1.02 | 3.51 | 33.84 |

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

Appendices: Appendix E – Items Flagged for DIF

These tables support the DIF information in Section V (Operational Test Data Collection and Classical Analyses) and Section VI (IRT Scaling). 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 E1 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 E1. NYSTP ELA 2006 Classical DIF Item Flags

| Grade | Item # | Subgroup | DIF | SMD | Mantel-Haenszel | Delta |
|-------|--------|----------|----------|---------|-----------------|---------|
| 3 | 8 | Asian | Against | -0.13 | 65.28 | -2.27 |
| 3 | 10 | Asian | In Favor | No flag | 7.13 | 2.49 |
| 3 | 25 | Hispanic | In Favor | 0.10 | n/a | n/a |
| 3 | 27 | Asian | In Favor | No flag | 20.29 | 1.51 |
| 3 | 28 | Asian | In Favor | 0.13 | No flag | No flag |
| 4 | 13 | Asian | Against | -0.12 | No flag | No flag |
| 4 | 30 | Asian | In Favor | 0.17 | n/a | n/a |
| 4 | 30 | Female | In Favor | 0.11 | n/a | n/a |
| 5 | 1 | Asian | Against | No flag | 35.33 | -1.93 |
| 5 | 2 | Asian | Against | No flag | 23.64 | -1.64 |
| 5 | 26 | Hispanic | In Favor | 0.12 | n/a | n/a |
| 5 | 26 | Asian | In Favor | 0.14 | n/a | n/a |
| 5 | 27 | Black | Against | -0.13 | n/a | n/a |
| 6 | 14 | Asian | In Favor | 0.10 | No flag | No flag |
| 6 | 15 | Hispanic | Against | -0.14 | 87.74 | -1.6 |
| 6 | 15 | Asian | Against | -0.10 | No flag | No flag |
| 6 | 26 | Asian | In Favor | No flag | 15.01 | 1.53 |
| 6 | 27 | Black | In Favor | 0.11 | n/a | n/a |
| 6 | 27 | Hispanic | In Favor | 0.16 | n/a | n/a |
| 6 | 27 | Asian | In Favor | 0.12 | n/a | n/a |
| 6 | 27 | Female | In Favor | 0.10 | n/a | n/a |
| 6 | 27 | High NRC | In Favor | 0.10 | n/a | n/a |
| 6 | 28 | Female | In Favor | 0.14 | n/a | n/a |
| 6 | 29 | Hispanic | In Favor | 0.14 | n/a | n/a |
| 6 | 29 | Asian | In Favor | 0.11 | n/a | n/a |
| 6 | 29 | Female | In Favor | 0.16 | n/a | n/a |
| 7 | 1 | Female | Against | -0.10 | No flag | No flag |
| 7 | 11 | Black | In Favor | No flag | 33.02 | 1.60 |
| 7 | 11 | Hispanic | In Favor | No flag | 26.63 | 1.59 |
| 7 | 15 | Female | Against | -0.15 | 259.14 | -2.29 |

(Continued on next page)

Table E1. NYSTP ELA 2006 Classical DIF Item Flags (cont.)

| Grade | Item # | Subgroup | DIF | SMD | Mantel-Haenszel | Delta |
|-------|--------|----------|----------|---------|-----------------|---------|
| 7 | 31 | Black | Against | No flag | 31.60 | -1.64 |
| 7 | 31 | High NRC | Against | No flag | 29.60 | -1.52 |
| 7 | 35 | High NRC | Against | -0.11 | n/a | n/a |
| 8 | 7 | Female | Against | No flag | 80.91 | -1.93 |
| 8 | 10 | Female | Against | -0.11 | No flag | No flag |
| 8 | 13 | Black | Against | No flag | 56.32 | -1.65 |
| 8 | 13 | Hispanic | Against | No flag | 41.62 | -1.62 |
| 8 | 13 | Asian | Against | No flag | 23.37 | -1.86 |
| 8 | 27 | Female | In Favor | 0.14 | n/a | n/a |
| 8 | 28 | Female | In Favor | 0.14 | n/a | n/a |

Note that positive values of D_{ig} in Table E2 indicate differential item functioning in favor of a focal group and negative values of D_{ig} indicate differential item functioning against a focal group.

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

| Grade | Item | Focal Group | Direction | Magnitude (D_{ig}) |
|-------|------|-------------|-----------|------------------------|
| 3 | 28 | Asian | In Favor | 0.132 |
| 4 | 30 | Asian | In Favor | 0.101 |
| 5 | 26 | Asian | In Favor | 0.100 |
| 5 | 27 | Black | Against | -0.101 |
| 6 | 9 | Asian | Against | -0.109 |
| 6 | 27 | Asian | In Favor | 0.103 |
| 6 | 27 | Hispanic | In Favor | 0.101 |
| 6 | 29 | Hispanic | In Favor | 0.106 |

Appendices: Appendix F – 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 all items in the Grades 3-8 ELA Tests was ok ($Z_{crit} > Z$).

Table F1. Q1 Fit Statistics, Grade 3

| Item Number | Model | Chi Sqr | DF | N | Z | Z_crit | Fit Ok? |
|-------------|-------|---------|----|--------|--------|--------|---------|
| 1 | 3PL | 32.66 | 7 | 175664 | 6.86 | 468.44 | Yes |
| 2 | 3PL | 434.47 | 7 | 175664 | 114.25 | 468.44 | Yes |
| 3 | 3PL | 61.69 | 7 | 175664 | 14.62 | 468.44 | Yes |
| 4 | 3PL | 108.15 | 7 | 175664 | 27.03 | 468.44 | Yes |
| 5 | 3PL | 202.91 | 7 | 175664 | 52.36 | 468.44 | Yes |
| 6 | 3PL | 117.67 | 7 | 175664 | 29.58 | 468.44 | Yes |
| 7 | 3PL | 136.73 | 7 | 175664 | 34.67 | 468.44 | Yes |
| 8 | 3PL | 22.49 | 7 | 175664 | 4.14 | 468.44 | Yes |
| 9 | 3PL | 300.81 | 7 | 175664 | 78.52 | 468.44 | Yes |
| 10 | 3PL | 74.01 | 7 | 175664 | 17.91 | 468.44 | Yes |
| 11 | 3PL | 226.23 | 7 | 175664 | 58.59 | 468.44 | Yes |
| 12 | 3PL | 368.20 | 7 | 175664 | 96.53 | 468.44 | Yes |
| 13 | 3PL | 82.45 | 7 | 175664 | 20.17 | 468.44 | Yes |
| 14 | 3PL | 300.83 | 7 | 175664 | 78.53 | 468.44 | Yes |
| 15 | 3PL | 200.00 | 7 | 175664 | 51.58 | 468.44 | Yes |
| 16 | 3PL | 93.00 | 7 | 175664 | 22.99 | 468.44 | Yes |
| 17 | 3PL | 214.62 | 7 | 175664 | 55.49 | 468.44 | Yes |
| 18 | 2PPC | 727.10 | 17 | 173901 | 121.78 | 463.74 | Yes |
| 19 | 3PL | 403.02 | 7 | 175664 | 105.84 | 468.44 | Yes |
| 20 | 3PL | 357.60 | 7 | 175664 | 93.70 | 468.44 | Yes |
| 21 | 3PL | 293.37 | 7 | 175664 | 76.54 | 468.44 | Yes |
| 22 | 3PL | 91.29 | 7 | 175664 | 22.53 | 468.44 | Yes |
| 23 | 3PL | 267.06 | 7 | 175664 | 69.50 | 468.44 | Yes |
| 24 | 3PL | 78.82 | 7 | 175664 | 19.19 | 468.44 | Yes |
| 25 | 2PPC | 523.16 | 17 | 174841 | 86.81 | 466.24 | Yes |
| 26 | 2PPC | 801.90 | 17 | 173270 | 134.61 | 462.05 | Yes |
| 27 | 3PL | 418.61 | 7 | 175664 | 110.01 | 468.44 | Yes |
| 28 | 2PPC | 797.34 | 26 | 175253 | 106.96 | 467.34 | Yes |

Table F2. Q1 Fit Statistics, Grade 4

| ItemNo | Model | Chi Sqr | DF | N | Z | Z_crit | Fit Ok? |
|--------|-------|---------|----|--------|--------|--------|---------|
| 1 | 3PL | 867.99 | 7 | 182093 | 230.11 | 485.58 | Yes |
| 2 | 3PL | 21.26 | 7 | 182093 | 3.81 | 485.58 | Yes |
| 3 | 3PL | 74.31 | 7 | 182093 | 17.99 | 485.58 | Yes |
| 4 | 3PL | 67.89 | 7 | 182093 | 16.27 | 485.58 | Yes |
| 5 | 3PL | 199.38 | 7 | 182093 | 51.42 | 485.58 | Yes |
| 6 | 3PL | 101.06 | 7 | 182093 | 25.14 | 485.58 | Yes |
| 7 | 3PL | 151.94 | 7 | 182093 | 38.74 | 485.58 | Yes |
| 8 | 3PL | 455.44 | 7 | 182093 | 119.85 | 485.58 | Yes |
| 9 | 3PL | 411.84 | 7 | 182093 | 108.20 | 485.58 | Yes |
| 10 | 3PL | 72.89 | 7 | 182093 | 17.61 | 485.58 | Yes |
| 11 | 3PL | 26.55 | 7 | 182093 | 5.23 | 485.58 | Yes |
| 12 | 3PL | 107.04 | 7 | 182093 | 26.74 | 485.58 | Yes |
| 13 | 3PL | 316.21 | 7 | 182093 | 82.64 | 485.58 | Yes |
| 14 | 3PL | 46.71 | 7 | 182093 | 10.61 | 485.58 | Yes |
| 15 | 3PL | 27.12 | 7 | 182093 | 5.38 | 485.58 | Yes |
| 16 | 3PL | 248.71 | 7 | 182093 | 64.60 | 485.58 | Yes |
| 17 | 3PL | 119.35 | 7 | 182093 | 30.03 | 485.58 | Yes |
| 18 | 3PL | 311.83 | 7 | 182093 | 81.47 | 485.58 | Yes |
| 19 | 3PL | 171.89 | 7 | 182093 | 44.07 | 485.58 | Yes |
| 20 | 3PL | 541.80 | 7 | 182093 | 142.93 | 485.58 | Yes |
| 21 | 3PL | 90.88 | 7 | 182093 | 22.42 | 485.58 | Yes |
| 22 | 3PL | 54.26 | 7 | 182093 | 12.63 | 485.58 | Yes |
| 23 | 3PL | 225.19 | 7 | 182093 | 58.31 | 485.58 | Yes |
| 24 | 3PL | 391.43 | 7 | 182093 | 102.74 | 485.58 | Yes |
| 25 | 3PL | 454.87 | 7 | 182093 | 119.70 | 485.58 | Yes |
| 26 | 3PL | 214.56 | 7 | 182093 | 55.47 | 485.58 | Yes |
| 27 | 3PL | 104.59 | 7 | 182093 | 26.08 | 485.58 | Yes |
| 28 | 3PL | 110.78 | 7 | 182093 | 27.74 | 485.58 | Yes |
| 29 | 2PPC | 1214.70 | 35 | 181991 | 141.00 | 485.31 | Yes |
| 30 | 2PPC | 890.99 | 26 | 182007 | 119.95 | 485.35 | Yes |
| 31 | 2PPC | 1984.40 | 35 | 182093 | 233.00 | 485.58 | Yes |

Table F3. Q1 Fit Statistics, Grade 5

| ItemNo | Model | Chi Sqr | DF | N | Z | Z_crit | Fit Ok? |
|--------|-------|---------|----|--------|--------|--------|---------|
| 1 | 3PL | 87.27 | 7 | 189898 | 21.45 | 506.39 | Yes |
| 2 | 3PL | 210.14 | 7 | 189898 | 54.29 | 506.39 | Yes |
| 3 | 3PL | 787.99 | 7 | 189898 | 208.73 | 506.39 | Yes |
| 4 | 3PL | 93.89 | 7 | 189898 | 23.22 | 506.39 | Yes |
| 5 | 3PL | 131.26 | 7 | 189898 | 33.21 | 506.39 | Yes |
| 6 | 3PL | 59.59 | 7 | 189898 | 14.05 | 506.39 | Yes |
| 7 | 3PL | 204.26 | 7 | 189898 | 52.72 | 506.39 | Yes |
| 8 | 3PL | 48.76 | 7 | 189898 | 11.16 | 506.39 | Yes |
| 9 | 3PL | 215.41 | 7 | 189898 | 55.7 | 506.39 | Yes |
| 10 | 3PL | 293.35 | 7 | 189898 | 76.53 | 506.39 | Yes |
| 11 | 3PL | 353.38 | 7 | 189898 | 92.57 | 506.39 | Yes |
| 12 | 2PPC | 788.67 | 17 | 188634 | 132.34 | 503.02 | Yes |
| 13 | 3PL | 74.61 | 7 | 189898 | 18.07 | 506.39 | Yes |
| 14 | 3PL | 54.20 | 7 | 189898 | 12.61 | 506.39 | Yes |
| 15 | 3PL | 194.01 | 7 | 189898 | 49.98 | 506.39 | Yes |
| 16 | 3PL | 174.47 | 7 | 189898 | 44.76 | 506.39 | Yes |
| 17 | 3PL | 239.39 | 7 | 189898 | 62.11 | 506.39 | Yes |
| 18 | 3PL | 370.39 | 7 | 189898 | 97.12 | 506.39 | Yes |
| 19 | 3PL | 84.35 | 7 | 189898 | 20.67 | 506.39 | Yes |
| 20 | 3PL | 261.69 | 7 | 189898 | 68.07 | 506.39 | Yes |
| 21 | 3PL | 279.34 | 7 | 189898 | 72.79 | 506.39 | Yes |
| 22 | 3PL | 180.28 | 7 | 189898 | 46.31 | 506.39 | Yes |
| 23 | 3PL | 250.44 | 7 | 189898 | 65.06 | 506.39 | Yes |
| 24 | 3PL | 359.55 | 7 | 189898 | 94.22 | 506.39 | Yes |
| 25 | 3PL | 245.97 | 7 | 189898 | 63.87 | 506.39 | Yes |
| 26 | 2PPC | 404.56 | 17 | 189669 | 66.47 | 505.78 | Yes |
| 27 | 2PPC | 592.53 | 26 | 189393 | 78.56 | 505.05 | Yes |

Table F4. Q1 Fit Statistics, Grade 6

| ItemNo | Model | Chi Sqr | DF | N | Z | Z_crit | Fit Ok? |
|--------|-------|---------|----|--------|--------|--------|---------|
| 1 | 3PL | 125.88 | 7 | 192500 | 31.77 | 513.33 | Yes |
| 2 | 3PL | 77.86 | 7 | 192500 | 18.94 | 513.33 | Yes |
| 3 | 3PL | 108.21 | 7 | 192500 | 27.05 | 513.33 | Yes |
| 4 | 3PL | 82.34 | 7 | 192500 | 20.14 | 513.33 | Yes |
| 5 | 3PL | 89.19 | 7 | 192500 | 21.97 | 513.33 | Yes |
| 6 | 3PL | 170.59 | 7 | 192500 | 43.72 | 513.33 | Yes |
| 7 | 3PL | 473.45 | 7 | 192500 | 124.66 | 513.33 | Yes |
| 8 | 3PL | 204.84 | 7 | 192500 | 52.88 | 513.33 | Yes |
| 9 | 3PL | 647.58 | 7 | 192500 | 171.20 | 513.33 | Yes |
| 10 | 3PL | 113.73 | 7 | 192500 | 28.52 | 513.33 | Yes |
| 11 | 3PL | 19.42 | 7 | 192500 | 3.32 | 513.33 | Yes |
| 12 | 3PL | 48.92 | 7 | 192500 | 11.20 | 513.33 | Yes |
| 13 | 3PL | 26.19 | 7 | 192500 | 5.13 | 513.33 | Yes |
| 14 | 3PL | 1764.67 | 7 | 192500 | 469.76 | 513.33 | Yes |
| 15 | 3PL | 101.42 | 7 | 192500 | 25.23 | 513.33 | Yes |
| 16 | 3PL | 179.08 | 7 | 192500 | 45.99 | 513.33 | Yes |
| 17 | 3PL | 97.70 | 7 | 192500 | 24.24 | 513.33 | Yes |
| 18 | 3PL | 326.04 | 7 | 192500 | 85.27 | 513.33 | Yes |
| 19 | 3PL | 166.97 | 7 | 192500 | 42.75 | 513.33 | Yes |
| 20 | 3PL | 164.67 | 7 | 192500 | 42.14 | 513.33 | Yes |
| 21 | 3PL | 169.22 | 7 | 192500 | 43.36 | 513.33 | Yes |
| 22 | 3PL | 115.44 | 7 | 192500 | 28.98 | 513.33 | Yes |
| 23 | 3PL | 106.19 | 7 | 192500 | 26.51 | 513.33 | Yes |
| 24 | 3PL | 175.97 | 7 | 192500 | 45.16 | 513.33 | Yes |
| 25 | 3PL | 305.58 | 7 | 192500 | 79.80 | 513.33 | Yes |
| 26 | 3PL | 265.97 | 7 | 192500 | 69.21 | 513.33 | Yes |
| 27 | 2PPC | 1617.33 | 44 | 192388 | 167.72 | 513.03 | Yes |
| 28 | 2PPC | 1082.47 | 26 | 192302 | 146.51 | 512.81 | Yes |
| 29 | 2PPC | 1823.21 | 44 | 192500 | 189.66 | 513.33 | Yes |

Table F5. Q1 Fit Statistics, Grade 7

| ItemNo | Model | Chi Sqr | DF | N | Z | Z_crit | Fit Ok? |
|--------|-------|---------|----|--------|--------|--------|---------|
| 1 | 3PL | 120.82 | 7 | 199458 | 30.42 | 531.89 | Yes |
| 2 | 3PL | 251.66 | 7 | 199458 | 65.39 | 531.89 | Yes |
| 3 | 3PL | 202.84 | 7 | 199458 | 52.34 | 531.89 | Yes |
| 4 | 3PL | 126.69 | 7 | 199458 | 31.99 | 531.89 | Yes |
| 5 | 3PL | 9.29 | 7 | 199458 | 0.61 | 531.89 | Yes |
| 6 | 3PL | 40.61 | 7 | 199458 | 8.98 | 531.89 | Yes |
| 7 | 3PL | 44.39 | 7 | 199458 | 9.99 | 531.89 | Yes |
| 8 | 3PL | 337.48 | 7 | 199458 | 88.33 | 531.89 | Yes |
| 9 | 3PL | 181.92 | 7 | 199458 | 46.75 | 531.89 | Yes |
| 10 | 3PL | 173.28 | 7 | 199458 | 44.44 | 531.89 | Yes |
| 11 | 3PL | 114.76 | 7 | 199458 | 28.80 | 531.89 | Yes |
| 12 | 3PL | 86.52 | 7 | 199458 | 21.25 | 531.89 | Yes |
| 13 | 3PL | 165.02 | 7 | 199458 | 42.23 | 531.89 | Yes |
| 14 | 3PL | 1352.60 | 7 | 199458 | 359.63 | 531.89 | Yes |
| 15 | 3PL | 97.64 | 7 | 199458 | 24.22 | 531.89 | Yes |
| 16 | 3PL | 215.25 | 7 | 199458 | 55.66 | 531.89 | Yes |
| 17 | 2PPC | 412.43 | 17 | 194739 | 67.82 | 519.30 | Yes |
| 18 | 3PL | 359.08 | 7 | 199458 | 94.10 | 531.89 | Yes |
| 19 | 3PL | 109.32 | 7 | 199458 | 27.35 | 531.89 | Yes |
| 20 | 3PL | 760.73 | 7 | 199458 | 201.44 | 531.89 | Yes |
| 21 | 3PL | 50.25 | 7 | 199458 | 11.56 | 531.89 | Yes |
| 22 | 3PL | 565.32 | 7 | 199458 | 149.22 | 531.89 | Yes |
| 23 | 2PPC | 200.25 | 17 | 194583 | 31.43 | 518.89 | Yes |
| 24 | 3PL | 207.49 | 7 | 199458 | 53.58 | 531.89 | Yes |
| 25 | 3PL | 49.55 | 7 | 199458 | 11.37 | 531.89 | Yes |
| 26 | 3PL | 128.91 | 7 | 199458 | 32.58 | 531.89 | Yes |
| 27 | 3PL | 92.05 | 7 | 199458 | 22.73 | 531.89 | Yes |
| 28 | 3PL | 372.67 | 7 | 199458 | 97.73 | 531.89 | Yes |
| 29 | 3PL | 59.88 | 7 | 199458 | 14.13 | 531.89 | Yes |
| 30 | 3PL | 453.43 | 7 | 199458 | 119.31 | 531.89 | Yes |
| 31 | 3PL | 162.05 | 7 | 199458 | 41.44 | 531.89 | Yes |
| 32 | 2PPC | 264.58 | 17 | 198409 | 42.46 | 529.09 | Yes |
| 33 | 2PPC | 213.40 | 17 | 198509 | 33.68 | 529.36 | Yes |
| 34 | 3PL | 113.91 | 7 | 199458 | 28.57 | 531.89 | Yes |
| 35 | 2PPC | 1337.02 | 26 | 198509 | 181.81 | 529.36 | Yes |

Table F6. Q1 Fit Statistics, Grade 8

| ItemNo | Model | Chi Sqr | DF | N | Z | Z_crit | Fit Ok? |
|--------|-------|---------|----|--------|--------|--------|---------|
| 1 | 3PL | 46.36 | 7 | 201438 | 10.52 | 537.17 | Yes |
| 2 | 3PL | 66.58 | 7 | 201438 | 15.92 | 537.17 | Yes |
| 3 | 3PL | 57.18 | 7 | 201438 | 13.41 | 537.17 | Yes |
| 4 | 3PL | 1238.64 | 7 | 201438 | 329.17 | 537.17 | Yes |
| 5 | 3PL | 364.40 | 7 | 201438 | 95.52 | 537.17 | Yes |
| 6 | 3PL | 134.60 | 7 | 201438 | 34.10 | 537.17 | Yes |
| 7 | 3PL | 118.20 | 7 | 201438 | 29.72 | 537.17 | Yes |
| 8 | 3PL | 107.91 | 7 | 201438 | 26.97 | 537.17 | Yes |
| 9 | 3PL | 119.81 | 7 | 201438 | 30.15 | 537.17 | Yes |
| 10 | 3PL | 300.35 | 7 | 201438 | 78.40 | 537.17 | Yes |
| 11 | 3PL | 168.38 | 7 | 201438 | 43.13 | 537.17 | Yes |
| 12 | 3PL | 71.29 | 7 | 201438 | 17.18 | 537.17 | Yes |
| 13 | 3PL | 54.58 | 7 | 201438 | 12.72 | 537.17 | Yes |
| 14 | 3PL | 56.44 | 7 | 201438 | 13.21 | 537.17 | Yes |
| 15 | 3PL | 51.37 | 7 | 201438 | 11.86 | 537.17 | Yes |
| 16 | 3PL | 894.31 | 7 | 201438 | 237.14 | 537.17 | Yes |
| 17 | 3PL | 31.73 | 7 | 201438 | 6.61 | 537.17 | Yes |
| 18 | 3PL | 45.00 | 7 | 201438 | 10.15 | 537.17 | Yes |
| 19 | 3PL | 423.64 | 7 | 201438 | 111.35 | 537.17 | Yes |
| 20 | 3PL | 165.02 | 7 | 201438 | 42.23 | 537.17 | Yes |
| 21 | 3PL | 859.23 | 7 | 201438 | 227.77 | 537.17 | Yes |
| 22 | 3PL | 444.32 | 7 | 201438 | 116.88 | 537.17 | Yes |
| 23 | 3PL | 53.98 | 7 | 201438 | 12.56 | 537.17 | Yes |
| 24 | 3PL | 163.07 | 7 | 201438 | 41.71 | 537.17 | Yes |
| 25 | 3PL | 102.57 | 7 | 201438 | 25.54 | 537.17 | Yes |
| 26 | 3PL | 54.34 | 7 | 201438 | 12.65 | 537.17 | Yes |
| 27 | 2PPC | 2600.77 | 44 | 201195 | 272.55 | 536.52 | Yes |
| 28 | 2PPC | 890.66 | 26 | 201213 | 119.91 | 536.57 | Yes |
| 29 | 2PPC | 2386.75 | 44 | 201438 | 249.74 | 537.17 | Yes |

Appendices: Appendix G – Derivation of the Generalized SPI Procedure

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

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

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

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

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

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

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

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 H – 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 I – Scale Score Frequency Distributions

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

Table I1. Grade 3 ELA 2006 Scale Score FD, State

| SS | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|-----|-----------|---------|----------------------|--------------------|
| 475 | 93 | 0.05 | 93 | 0.05 |
| 503 | 143 | 0.08 | 236 | 0.13 |
| 538 | 292 | 0.16 | 528 | 0.28 |
| 554 | 489 | 0.26 | 1017 | 0.55 |
| 565 | 639 | 0.34 | 1656 | 0.89 |
| 574 | 911 | 0.49 | 2567 | 1.38 |
| 581 | 1044 | 0.56 | 3611 | 1.95 |
| 587 | 1173 | 0.63 | 4784 | 2.58 |
| 592 | 1369 | 0.74 | 6153 | 3.32 |
| 597 | 1504 | 0.81 | 7657 | 4.13 |
| 601 | 1723 | 0.93 | 9380 | 5.06 |
| 605 | 1862 | 1.00 | 11242 | 6.06 |
| 609 | 2113 | 1.14 | 13355 | 7.20 |
| 613 | 2469 | 1.33 | 15824 | 8.53 |
| 617 | 2813 | 1.52 | 18637 | 10.05 |
| 621 | 3201 | 1.73 | 21838 | 11.77 |
| 625 | 3690 | 1.99 | 25528 | 13.76 |
| 629 | 4444 | 2.40 | 29972 | 16.15 |
| 633 | 5176 | 2.79 | 35148 | 18.94 |
| 638 | 6207 | 3.35 | 41355 | 22.29 |
| 642 | 7397 | 3.99 | 48752 | 26.28 |
| 647 | 8764 | 4.72 | 57516 | 31.00 |
| 653 | 10310 | 5.56 | 67826 | 36.56 |
| 659 | 11805 | 6.36 | 79631 | 42.92 |
| 665 | 13530 | 7.29 | 93161 | 50.21 |
| 672 | 15084 | 8.13 | 108245 | 58.34 |
| 680 | 16621 | 8.96 | 124866 | 67.30 |

(Continued on next page)

Table I1. Grade 3 ELA 2006 Scale Score FD, State (cont.)

| SS | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|-----|-----------|---------|----------------------|--------------------|
| 690 | 17462 | 9.41 | 142328 | 76.71 |
| 701 | 16325 | 8.80 | 158653 | 85.51 |
| 717 | 13754 | 7.41 | 172407 | 92.93 |
| 744 | 9243 | 4.98 | 181650 | 97.91 |
| 780 | 3883 | 2.09 | 185533 | 100.00 |

Table I2. Grade 4 ELA 2006 Scale Score FD, State

| SS | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|-----|-----------|---------|----------------------|--------------------|
| 430 | 152 | 0.08 | 152 | 0.08 |
| 461 | 139 | 0.07 | 291 | 0.15 |
| 504 | 276 | 0.14 | 567 | 0.30 |
| 524 | 377 | 0.20 | 944 | 0.49 |
| 538 | 509 | 0.27 | 1453 | 0.76 |
| 550 | 612 | 0.32 | 2065 | 1.08 |
| 560 | 770 | 0.40 | 2835 | 1.49 |
| 569 | 989 | 0.52 | 3824 | 2.00 |
| 577 | 1236 | 0.65 | 5060 | 2.65 |
| 584 | 1411 | 0.74 | 6471 | 3.39 |
| 590 | 1593 | 0.83 | 8064 | 4.23 |
| 596 | 1858 | 0.97 | 9922 | 5.20 |
| 601 | 2050 | 1.07 | 11972 | 6.27 |
| 606 | 2359 | 1.24 | 14331 | 7.51 |
| 611 | 2691 | 1.41 | 17022 | 8.92 |
| 616 | 3043 | 1.59 | 20065 | 10.51 |
| 620 | 3483 | 1.83 | 23548 | 12.34 |
| 624 | 3919 | 2.05 | 27467 | 14.39 |
| 628 | 4320 | 2.26 | 31787 | 16.66 |
| 632 | 4741 | 2.48 | 36528 | 19.14 |
| 636 | 5172 | 2.71 | 41700 | 21.85 |
| 640 | 5637 | 2.95 | 47337 | 24.80 |
| 644 | 6039 | 3.16 | 53376 | 27.97 |
| 648 | 6390 | 3.35 | 59766 | 31.32 |
| 652 | 7060 | 3.70 | 66826 | 35.02 |

(Continued on next page)

Table I2. Grade 4 ELA 2006 Scale Score FD, State (cont.)

| SS | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|-----|-----------|---------|----------------------|--------------------|
| 656 | 7862 | 4.12 | 74688 | 39.14 |
| 660 | 8258 | 4.33 | 82946 | 43.46 |
| 664 | 8994 | 4.71 | 91940 | 48.17 |
| 668 | 9638 | 5.05 | 101578 | 53.22 |
| 673 | 9985 | 5.23 | 111563 | 58.46 |
| 678 | 10688 | 5.60 | 122251 | 64.06 |
| 683 | 11258 | 5.90 | 133509 | 69.96 |
| 689 | 11444 | 6.00 | 144953 | 75.95 |
| 695 | 10910 | 5.72 | 155863 | 81.67 |
| 703 | 9986 | 5.23 | 165849 | 86.90 |
| 711 | 8320 | 4.36 | 174169 | 91.26 |
| 721 | 6468 | 3.39 | 180637 | 94.65 |
| 735 | 5245 | 2.75 | 185882 | 97.40 |
| 756 | 3683 | 1.93 | 189565 | 99.33 |
| 775 | 1282 | 0.67 | 190847 | 100.00 |

Table I3. Grade 5 ELA 2006 Scale Score FD, State

| SS | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|-----|-----------|---------|----------------------|--------------------|
| 495 | 1146 | 0.57 | 1146 | 0.57 |
| 524 | 899 | 0.45 | 2045 | 1.02 |
| 557 | 1296 | 0.64 | 3341 | 1.66 |
| 574 | 1700 | 0.85 | 5041 | 2.51 |
| 586 | 2156 | 1.07 | 7197 | 3.58 |
| 595 | 2578 | 1.28 | 9775 | 4.86 |
| 602 | 3058 | 1.52 | 12833 | 6.38 |
| 609 | 3657 | 1.82 | 16490 | 8.20 |
| 614 | 4480 | 2.23 | 20970 | 10.43 |
| 620 | 5258 | 2.61 | 26228 | 13.04 |
| 625 | 5878 | 2.92 | 32106 | 15.96 |
| 630 | 6866 | 3.41 | 38972 | 19.38 |
| 635 | 7973 | 3.96 | 46945 | 23.34 |
| 640 | 9019 | 4.48 | 55964 | 27.82 |
| 645 | 10062 | 5.00 | 66026 | 32.83 |

(Continued on next page)

Table I3. Grade 5 ELA 2006 Scale Score FD, State (cont.)

| SS | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|-----|-----------|---------|----------------------|--------------------|
| 650 | 11349 | 5.64 | 77375 | 38.47 |
| 655 | 12589 | 6.26 | 89964 | 44.73 |
| 661 | 13606 | 6.76 | 103570 | 51.49 |
| 667 | 14549 | 7.23 | 118119 | 58.73 |
| 674 | 15101 | 7.51 | 133220 | 66.23 |
| 681 | 15246 | 7.58 | 148466 | 73.81 |
| 690 | 14661 | 7.29 | 163127 | 81.10 |
| 699 | 13248 | 6.59 | 176375 | 87.69 |
| 712 | 11189 | 5.56 | 187564 | 93.25 |
| 729 | 7942 | 3.95 | 195506 | 97.20 |
| 764 | 4359 | 2.17 | 199865 | 99.37 |
| 795 | 1273 | 0.63 | 201138 | 100.00 |

Table I4. Grade 6 ELA 2006 Scale Score FD, State

| SS | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|-----|-----------|---------|----------------------|--------------------|
| 480 | 258 | 0.13 | 258 | 0.13 |
| 495 | 236 | 0.12 | 494 | 0.24 |
| 518 | 402 | 0.20 | 896 | 0.44 |
| 534 | 601 | 0.29 | 1497 | 0.73 |
| 546 | 833 | 0.41 | 2330 | 1.14 |
| 557 | 1152 | 0.56 | 3482 | 1.71 |
| 566 | 1474 | 0.72 | 4956 | 2.43 |
| 574 | 1834 | 0.90 | 6790 | 3.33 |
| 582 | 2283 | 1.12 | 9073 | 4.45 |
| 589 | 2661 | 1.30 | 11734 | 5.75 |
| 595 | 3124 | 1.53 | 14858 | 7.28 |
| 601 | 3625 | 1.78 | 18483 | 9.06 |
| 606 | 4237 | 2.08 | 22720 | 11.13 |
| 612 | 4935 | 2.42 | 27655 | 13.55 |
| 617 | 5471 | 2.68 | 33126 | 16.23 |
| 622 | 6278 | 3.08 | 39404 | 19.31 |
| 627 | 6911 | 3.39 | 46315 | 22.69 |

(Continued on next page)

Table I4. Grade 6 ELA 2006 Scale Score FD, State (cont.)

| SS | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|-----|-----------|---------|----------------------|--------------------|
| 632 | 7594 | 3.72 | 53909 | 26.41 |
| 636 | 8168 | 4.00 | 62077 | 30.41 |
| 641 | 8925 | 4.37 | 71002 | 34.79 |
| 646 | 9658 | 4.73 | 80660 | 39.52 |
| 650 | 9998 | 4.90 | 90658 | 44.42 |
| 655 | 10750 | 5.27 | 101408 | 49.68 |
| 660 | 10959 | 5.37 | 112367 | 55.05 |
| 665 | 11816 | 5.79 | 124183 | 60.84 |
| 671 | 11762 | 5.76 | 135945 | 66.61 |
| 676 | 11691 | 5.73 | 147636 | 72.33 |
| 682 | 11585 | 5.68 | 159221 | 78.01 |
| 689 | 11134 | 5.46 | 170355 | 83.46 |
| 697 | 10069 | 4.93 | 180424 | 88.40 |
| 706 | 8618 | 4.22 | 189042 | 92.62 |
| 717 | 6763 | 3.31 | 195805 | 95.93 |
| 733 | 4765 | 2.33 | 200570 | 98.27 |
| 762 | 2667 | 1.31 | 203237 | 99.58 |
| 785 | 867 | 0.42 | 204104 | 100.00 |

Table I5. Grade 7 ELA 2006 Scale Score FD, State

| SS | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|-----|-----------|---------|----------------------|--------------------|
| 470 | 569 | 0.27 | 569 | 0.27 |
| 489 | 444 | 0.21 | 1013 | 0.48 |
| 522 | 642 | 0.30 | 1655 | 0.79 |
| 540 | 922 | 0.44 | 2577 | 1.22 |
| 553 | 1109 | 0.53 | 3686 | 1.75 |
| 563 | 1404 | 0.67 | 5090 | 2.42 |
| 572 | 1635 | 0.78 | 6725 | 3.19 |
| 579 | 2028 | 0.96 | 8753 | 4.16 |
| 585 | 2303 | 1.09 | 11056 | 5.25 |
| 591 | 2695 | 1.28 | 13751 | 6.53 |
| 596 | 3150 | 1.50 | 16901 | 8.03 |

(Continued on next page)

Table I5. Grade 7 ELA 2006 Scale Score FD, State (cont.)

| SS | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|-----|-----------|---------|----------------------|--------------------|
| 601 | 3560 | 1.69 | 20461 | 9.72 |
| 605 | 4019 | 1.91 | 24480 | 11.63 |
| 610 | 4265 | 2.03 | 28745 | 13.65 |
| 614 | 4972 | 2.36 | 33717 | 16.02 |
| 618 | 5510 | 2.62 | 39227 | 18.63 |
| 622 | 5959 | 2.83 | 45186 | 21.46 |
| 626 | 6412 | 3.05 | 51598 | 24.51 |
| 630 | 6956 | 3.30 | 58554 | 27.81 |
| 633 | 7301 | 3.47 | 65855 | 31.28 |
| 637 | 8011 | 3.81 | 73866 | 35.09 |
| 641 | 8652 | 4.11 | 82518 | 39.20 |
| 645 | 9226 | 4.38 | 91744 | 43.58 |
| 650 | 9707 | 4.61 | 101451 | 48.19 |
| 654 | 10622 | 5.05 | 112073 | 53.24 |
| 659 | 11385 | 5.41 | 123458 | 58.64 |
| 664 | 12170 | 5.78 | 135628 | 64.43 |
| 669 | 12360 | 5.87 | 147988 | 70.30 |
| 675 | 12693 | 6.03 | 160681 | 76.33 |
| 682 | 12245 | 5.82 | 172926 | 82.14 |
| 691 | 11562 | 5.49 | 184488 | 87.64 |
| 700 | 9689 | 4.60 | 194177 | 92.24 |
| 713 | 7820 | 3.71 | 201997 | 95.95 |
| 730 | 5169 | 2.46 | 207166 | 98.41 |
| 759 | 2602 | 1.24 | 209768 | 99.64 |
| 790 | 750 | 0.36 | 210518 | 100.00 |

Table I6. Grade 8 ELA 2006 Scale Score FD, State

| SS | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|-----|-----------|---------|----------------------|--------------------|
| 430 | 96 | 0.05 | 96 | 0.05 |
| 458 | 103 | 0.05 | 199 | 0.09 |
| 499 | 141 | 0.07 | 340 | 0.16 |
| 516 | 227 | 0.11 | 567 | 0.27 |

(Continued on next page)

Table I6. Grade 8 ELA 2006 Scale Score FD, State (cont.)

| SS | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|-----|-----------|---------|----------------------|--------------------|
| 529 | 313 | 0.15 | 880 | 0.41 |
| 538 | 430 | 0.20 | 1310 | 0.62 |
| 547 | 563 | 0.27 | 1873 | 0.88 |
| 554 | 727 | 0.34 | 2600 | 1.23 |
| 560 | 855 | 0.40 | 3455 | 1.63 |
| 566 | 1100 | 0.52 | 4555 | 2.15 |
| 572 | 1300 | 0.61 | 5855 | 2.76 |
| 577 | 1567 | 0.74 | 7422 | 3.50 |
| 582 | 1827 | 0.86 | 9249 | 4.36 |
| 586 | 2154 | 1.02 | 11403 | 5.38 |
| 590 | 2431 | 1.15 | 13834 | 6.52 |
| 594 | 2868 | 1.35 | 16702 | 7.87 |
| 598 | 3291 | 1.55 | 19993 | 9.42 |
| 602 | 3713 | 1.75 | 23706 | 11.17 |
| 606 | 4273 | 2.01 | 27979 | 13.19 |
| 610 | 4816 | 2.27 | 32795 | 15.46 |
| 614 | 5567 | 2.62 | 38362 | 18.08 |
| 618 | 5974 | 2.82 | 44336 | 20.90 |
| 622 | 7048 | 3.32 | 51384 | 24.22 |
| 626 | 7691 | 3.63 | 59075 | 27.85 |
| 630 | 8416 | 3.97 | 67491 | 31.81 |
| 634 | 9169 | 4.32 | 76660 | 36.14 |
| 638 | 9686 | 4.57 | 86346 | 40.70 |
| 643 | 10329 | 4.87 | 96675 | 45.57 |
| 647 | 10712 | 5.05 | 107387 | 50.62 |
| 652 | 11122 | 5.24 | 118509 | 55.86 |
| 657 | 11432 | 5.39 | 129941 | 61.25 |
| 662 | 11338 | 5.34 | 141279 | 66.60 |
| 668 | 11318 | 5.34 | 152597 | 71.93 |
| 674 | 11185 | 5.27 | 163782 | 77.21 |
| 681 | 10651 | 5.02 | 174433 | 82.23 |
| 689 | 10250 | 4.83 | 184683 | 87.06 |
| 698 | 9241 | 4.36 | 193924 | 91.41 |
| 710 | 7937 | 3.74 | 201861 | 95.16 |

(Continued on next page)

Table I6. Grade 8 ELA 2006 Scale Score FD, State (cont.)

| SS | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|-----|-----------|---------|----------------------|--------------------|
| 728 | 5939 | 2.80 | 207800 | 97.96 |
| 760 | 3359 | 1.58 | 211159 | 99.54 |
| 790 | 979 | 0.46 | 212138 | 100.00 |