

# Wisconsin Forward Exam

## Technical Report 2017



Revised and Submitted to  
Wisconsin Department of Public Instruction  
January 2019



---

## Copyright

---

Developed and published under contract with the Wisconsin Department of Public Instruction by Data Recognition Corporation, 13490 Bass Lake Road, Maple Grove, MN 55311. Copyright © 2019 by the Wisconsin Department of Public Instruction. All rights reserved. Only State of Wisconsin educators and citizens may copy, download and/or print the document, located online at <http://dpi.wi.gov>. Any other use or reproduction of this document, in whole or in part, requires written permission of the Wisconsin Department of Public Instruction.

## Foreword

The technical information herein is intended for use by those who evaluate tests, interpret scores, or use test results in making educational decisions. It is assumed that the reader has technical knowledge of test construction and measurement procedures, as stated in *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014).

# TABLE OF CONTENTS

Copyright .....	ii
Foreword .....	iii
Table of Revisions.....	xii
<b>Part 1: Overview .....</b>	<b>1</b>
1.1 HISTORICAL BACKGROUND .....	1
1.2 USES OF TEST SCORES .....	3
1.2.1 TEST-LEVEL SCORES .....	4
1.2.2 STANDARD-LEVEL SUBSCORES AND PERFORMANCE LEVELS .....	5
1.3 TECHNICAL REPORT STRUCTURE.....	6
<b>Part 2: Test Blueprint and Item Development .....</b>	<b>9</b>
2.1 TEST BLUEPRINTS.....	10
2.2 READING PASSAGE AND ITEM SELECTION FOR SPRING 2016 FIELD TEST .....	10
2.3 FIELD TESTING .....	11
2.4 STATISTICAL ANALYSIS OF SPRING 2016 FIELD TEST DATA .....	11
2.5 REVIEW OF ITEMS WITH DATA.....	12
2.6 SUMMARY .....	13
<b>Part 3: Test Form Development .....</b>	<b>19</b>
<b>3.1 DESIGN OF THE WISCONSIN FORWARD EXAM .....</b>	<b>19</b>
3.1.1 ENGLISH LANGUAGE ARTS .....	19
3.1.2 MATHEMATICS .....	19
3.1.3 SCIENCE.....	20
3.1.4 SOCIAL STUDIES .....	20
3.2 TEST DEVELOPMENT PROCESS .....	20
3.2.1 WISCONSIN FORWARD TEST FORM CREATION .....	21
3.2.2 ITEM SELECTION.....	21
3.2.3 QUALITY REVIEWS .....	23
3.3 DPI APPROVALS .....	24
3.4 SUMMARY .....	24
<b>Part 4: Test Administration.....</b>	<b>28</b>
4.1 ACCESSIBILITY RESOURCES.....	29
4.1.1 UNIVERSAL TOOLS .....	29
4.1.2 DESIGNATED SUPPORTS.....	29
4.1.3 ACCOMMODATIONS .....	30
4.1.4 TRANSLATION.....	31
4.1.5 ADDITIONAL ACCESSIBILITY RESOURCES.....	32
4.2 REPORTING RESULTS OF ASSESSMENTS TAKEN WITH ACCOMMODATIONS .....	32
4.3 TEST SECURITY.....	32
4.3.1 SECURE STUDENT ACCESS.....	33
4.3.2 TEST SECURITY DURING BREAKS.....	34
4.4 TEST ADMINISTRATION TRAINING.....	34
4.5 SUMMARY .....	37
<b>Part 5: Scoring.....</b>	<b>48</b>
5.1 MULTIPLE-CHOICE AND MULTI-SELECT ITEM SCORING PROCESS.....	48
5.2 TECHNOLOGY-ENHANCED, SHORT-ANSWER, AND EVIDENCE-BASED SELECTED RESPONSE ITEM SCORING PROCESS.....	48
5.3 SCORING OF TEXT-DEPENDENT ANALYSIS ITEMS.....	49
5.3.1 DESCRIPTION OF SCORING RUBRICS AND NON-SCORE CODES.....	49
5.3.2 ARTIFICIAL INTELLIGENCE SCORING .....	50

5.3.3 HANDSCORING PROCESS.....	50
5.3.4 HANDSCORING SYSTEM.....	50
5.3.5 ANCHOR PAPERS AND TRAINING PAPERS .....	51
5.3.6 SCORING PERSONNEL AND QUALIFICATIONS .....	51
5.3.7 SCORER TRAINING .....	52
5.3.8 MONITORING THE SCORING PROCESS .....	52
5.3.9 FINAL SCORES .....	53
5.4 INTER-RATER RELIABILITY .....	53
5.4.1 DISTRIBUTION OF TDA ITEM SCORES.....	53
5.5 SUMMARY .....	54
<b>Part 6: Calibration, Equating, and Deriving Scale Scores.....</b>	<b>60</b>
6.1 ITEM CALIBRATION .....	60
6.1.1 CALIBRATION MODELS.....	60
6.1.2 CALIBRATION SAMPLE.....	61
6.1.3 CALIBRATION PROCEDURE .....	62
6.1.4 CALIBRATION SOFTWARE.....	62
6.1.5 CALIBRATION RESULTS .....	63
6.2 TEST EQUATING.....	65
6.2.1 EVALUATION OF ANCHOR ITEMS .....	66
6.2.2 REMOVAL OF ANCHOR ITEMS.....	67
6.2.3 EVALUATION OF EQUATING RESULTS.....	68
6.2.4 TEST SCALES .....	69
6.3 DERIVING SCALE SCORES IN THE WISCONSIN FORWARD EXAM.....	72
6.3.1 CONDITIONAL STANDARD ERROR OF MEASUREMENT .....	76
6.3.2 LOSS AND HOSS .....	77
6.4 SUMMARY .....	78
<b>Part 7: Standard Setting .....</b>	<b>125</b>
7.1 BACKGROUND INFORMATION .....	125
7.2 STANDARD SETTING METHODOLOGY .....	126
7.3 PERFORMANCE LEVEL DESCRIPTORS.....	126
7.4 CUT SCORES .....	126
7.5 SUMMARY .....	127
<b>Part 8: Test Results .....</b>	<b>129</b>
8.1 CLASSICAL ITEM ANALYSIS: ITEM LEVEL STATISTICS.....	129
8.1.1 FLAGGING FOR A POSITIVE DISTRACTOR CORRELATION .....	132
8.1.2 FLAGGING FOR THE ITEM-TOTAL CORRELATION .....	132
8.1.3 FLAGGING FOR P-VALUE.....	132
8.1.4 FLAGGING FOR OMIT RATE.....	132
8.1.5 SPEEDEDNESS .....	132
8.1.6 SUPPLEMENTAL TABLES ON CLASSICAL ITEM ANALYSIS .....	133
8.2 RAW SCORE RESULTS .....	133
8.3 SUMMARY STATISTICS FOR SCALE SCORES .....	136
8.4 CUT SCORES AND PERFORMANCE LEVEL CLASSIFICATIONS.....	139
8.5 STANDARD PERFORMANCE INDEX FOR CONTENT STANDARDS.....	141
8.6 LONGITUDINAL COMPARISONS OF TEST SCORES .....	144
8.7 SUMMARY .....	145
<b>Part 9: Reliability .....</b>	<b>218</b>
9.1 MEASURES OF INTERNAL CONSISTENCY AND STANDARD ERROR OF MEASUREMENT .....	220
9.1.1 CONDITIONAL STANDARD ERROR OF MEASUREMENT .....	222
9.2 CLASSIFICATION CONSISTENCY AND ACCURACY .....	223
9.2.1 KOLEN AND KIM’S METHOD FOR PATTERN SCORING .....	224
9.3 INTER-RATER RELIABILITY FOR TDA ITEMS.....	227
9.4 SUMMARY .....	229
<b>Part 10: Validity .....</b>	<b>244</b>

10.1 DIFFERENTIAL ITEM FUNCTIONING.....	248
10.2 VALIDITY EVIDENCE BASED ON INTERNAL TEST STRUCTURE.....	252
10.2.1 CORRELATIONS BETWEEN CONTENT STANDARDS .....	252
10.2.2 PRINCIPAL COMPONENT ANALYSIS .....	253
10.3 VALIDITY EVIDENCE BASED ON RELATIONSHIP WITH OTHER VARIABLES .....	254
10.3.1 CORRELATIONS BETWEEN CONTENT AREA TEST SCORES .....	254
10.3.2 COMPARISON OF THE WISCONSIN FORWARD EXAM AND WISCONSIN NAEP IMPACT DATA .....	255
10.4 TEST INTEGRITY: DATA FORENSIC ANALYSES .....	257
10.5 STANDARDIZED TEST ADMINISTRATION.....	257
10.6 SUMMARY.....	258
<b>Part 11: Summary Recommendations .....</b>	<b>275</b>
<b>References .....</b>	<b>276</b>

## APPENDICES

<b>Appendix A: Spring 2016 Field Test Data Review Training Slides .....</b>	<b>280</b>
<b>Appendix B: Spring 2016 Field Test Data Review Results.....</b>	<b>293</b>
<b>Appendix C: Spring 2017 English Language Arts Operational Test Maps.....</b>	<b>295</b>
<b>Appendix D: Spring 2017 Mathematics Operational Test Maps.....</b>	<b>302</b>
<b>Appendix E: Spring 2017 Science Operational Test Maps.....</b>	<b>315</b>
<b>Appendix F: Spring 2017 Social Studies Operational Test Maps.....</b>	<b>320</b>
<b>Appendix G: Classical Item Analysis Results.....</b>	<b>327</b>
<b>Appendix H: Wisconsin Standard Performance Index Score Computation.....</b>	<b>363</b>
<b>Appendix I: Conditional Standard Error of Measurement with Cut Scores.....</b>	<b>371</b>
<b>Appendix J: Classification Consistency and Accuracy Indices by Subgroup.....</b>	<b>389</b>
<b>Appendix K: Glossary.....</b>	<b>419</b>

## LIST OF TABLES

### PART 2

Table 2-1 College- and Career-Ready Item Bank Development Activities .....	14
Table 2-2 CCR Item Bank Item Type Descriptions.....	15
Table 2-3 English Language Arts Test Blueprints for Grades 3–8.....	16
Table 2-4 Mathematics Test Blueprints for Grades 3–8.....	17
Table 2-5 Science Test Blueprints for Grades 4 and 8 .....	18
Table 2-6 Social Studies Test Blueprints for Grades 4, 8, and 10 .....	18

### PART 3

Table 3-1 English Language Arts Test Design .....	24
Table 3-2 Mathematics Test Design .....	26
Table 3-3 Science Test Design .....	27
Table 3-4 Social Studies Test Design .....	27

### PART 4

Table 4-1 Number and Percentage of Students Using Accommodations or Designated Supports: Grade 3 .....	38
Table 4-2 Number and Percentage of Students Using Accommodations or Designated Supports: Grade 4 .....	39
Table 4-3 Number and Percentage of Students Using Accommodations or Designated Supports: Grade 5 .....	40
Table 4-4 Number and Percentage of Students Using Accommodations or Designated Supports: Grade 6 .....	41
Table 4-5 Number and Percentage of Students Using Accommodations or Designated Supports: Grade 7 .....	42
Table 4-6 Number and Percentage of Students Using Accommodations or Designated Supports: Grade 8 .....	43
Table 4-7 Number and Percentage of Students Using Accommodations or Designated Supports: Grade 10 .....	44
Table 4-8 Summary Table of Manual Materials.....	45

### PART 5

Table 5-1 TDA Item Scoring Guidelines, Grades 3–8.....	55
Table 5-2 TDA Item Non-scorable Codes, Grades 3–8.....	57
Table 5-3 TDA Item Score Distribution .....	58
Table 5-4 TDA Item Score Distribution: AI Engine vs. Human Scorer .....	58
Table 5-5 TDA Item Percentage Score Distribution: AI Engine vs. Human Scorer.....	59

### PART 6

Table 6-A Example of Item Parameters for a Test .....	73
Table 6-B Example of Item Response Pattern .....	74
Table 6-1 English Language Arts Calibration Sample Demographics Compared to Population .....	79
Table 6-2 Mathematics Calibration Sample Demographics Compared to Population .....	82
Table 6-3 Science Calibration Sample Demographics Compared to Population.....	85
Table 6-4 Social Studies Calibration Sample Demographics Compared to Population .....	86
Table 6-5 Item Flagged Based on Yen’s Q1.....	88
Table 6-6 Equating Evaluation Results, Stocking and Lord Method .....	88
Table 6-7 Statistics Comparing IRT Item-Ability Regression Curves for Flagged Anchor Items .....	89
Table 6-8 Scale Transformation Constants.....	89
Table 6-9 Scoring Table for English Language Arts Grade 3 .....	90
Table 6-10 Scoring Table for English Language Arts Grade 4 .....	91
Table 6-11 Scoring Table for English Language Arts Grade 5 .....	92
Table 6-12 Scoring Table for English Language Arts Grade 6 .....	93
Table 6-13 Scoring Table for English Language Arts Grade 7 .....	94
Table 6-14 Scoring Table for English Language Arts Grade 8 .....	95

Table 6-15 Scoring Table for Mathematics Grade 3.....	96
Table 6-16 Scoring Table for Mathematics Grade 4.....	97
Table 6-17 Scoring Table for Mathematics Grade 5.....	98
Table 6-18 Scoring Table for Mathematics Grade 6.....	99
Table 6-19 Scoring Table for Mathematics Grade 7.....	100
Table 6-20 Scoring Table for Mathematics Grade 8.....	101
Table 6-21 Scoring Table for Science Grade 4.....	102
Table 6-22 Scoring Table for Science Grade 8.....	103
Table 6-23 Scoring Table for Social Studies Grade 4.....	104
Table 6-24 Scoring Table for Social Studies Grade 8.....	105
Table 6-25 Scoring Table for Social Studies Grade 10.....	106
Table 6-26 The Number and Percentage of Students at LOSS and HOSS.....	107

**PART 7**

Table 7-1 Policy Performance Level Descriptors for the Wisconsin Forward Exam.....	128
Table 7-2 Wisconsin Forward Exam Cut Scores.....	128

**PART 8**

Table 8-A Summary of Flagged Operational Items on the Spring 2017 Wisconsin Forward Exam.....	146
Table 8-B English Language Arts Items Flagged for Classical Item Analysis Statistics.....	147
Table 8-C Mathematics Items Flagged for Classical Item Analysis Statistics.....	148
Table 8-D Science and Social Studies Items Flagged for Classical Item Analysis Statistics.....	149
Table 8-E Percentage of Students Attempting Last Operational Item in Test.....	149
Table 8-1 Item Analysis, Grade 3 English Language Arts.....	150
Table 8-2 Item Analysis, Grade 4 English Language Arts.....	151
Table 8-3 Item Analysis, Grade 5 English Language Arts.....	152
Table 8-4 Item Analysis, Grade 6 English Language Arts.....	153
Table 8-5 Item Analysis, Grade 7 English Language Arts.....	154
Table 8-6 Item Analysis, Grade 8 English Language Arts.....	155
Table 8-7 Item Analysis, Grade 3 Mathematics.....	156
Table 8-8 Item Analysis, Grade 4 Mathematics.....	158
Table 8-9 Item Analysis, Grade 5 Mathematics.....	160
Table 8-10 Item Analysis, Grade 6 Mathematics.....	162
Table 8-11 Item Analysis, Grade 7 Mathematics.....	164
Table 8-12 Item Analysis, Grade 8 Mathematics.....	166
Table 8-13 Item Analysis, Grade 4 Science.....	168
Table 8-14 Item Analysis, Grade 8 Science.....	170
Table 8-15 Item Analysis, Grade 4 Social Studies.....	172
Table 8-16 Item Analysis, Grade 8 Social Studies.....	173
Table 8-17 Item Analysis, Grade 10 Social Studies.....	174
Table 8-18 Raw Score Descriptive Statistics.....	176
Table 8-19 Raw Score Descriptive Statistics by Gender.....	177
Table 8-20 Raw Score Descriptive Statistics for English Language Arts by Race/Ethnicity.....	178
Table 8-21 Raw Score Descriptive Statistics for Mathematics by Race/Ethnicity.....	179
Table 8-22 Raw Score Descriptive Statistics for Science by Race/Ethnicity.....	180
Table 8-23 Raw Score Descriptive Statistics for Social Studies by Race/Ethnicity.....	180
Table 8-24 Raw Score Descriptive Statistics by Socioeconomic Status.....	181
Table 8-25 Raw Score Descriptive Statistics by Disability.....	182
Table 8-26 Raw Score Descriptive Statistics by English Language Proficiency.....	183
Table 8-27 Scale Score Descriptive Statistics.....	184
Table 8-28 Scale Score Descriptive Statistics by Gender.....	185
Table 8-29 Scale Score Descriptive Statistics for English Language Arts by Race/Ethnicity.....	186
Table 8-30 Scale Score Descriptive Statistics for Mathematics by Race/Ethnicity.....	187
Table 8-31 Scale Score Descriptive Statistics for Science by Race/Ethnicity.....	188



Table 8-32 Scale Score Descriptive Statistics for Social Studies by Race/Ethnicity .....	188
Table 8-33 Scale Score Descriptive Statistics by Socioeconomic Status .....	189
Table 8-34 Scale Score Descriptive Statistics by Disability .....	190
Table 8-35 Scale Score Descriptive Statistics by English Language Proficiency .....	191
Table 8-36 Performance Level Cut Scores for All Contents .....	192
Table 8-37 Cut Scores and Associated Impact Data, English Language Arts .....	192
Table 8-38 Cut Scores and Associated Impact Data, Mathematics .....	193
Table 8-39 Cut Scores and Associated Impact Data, Science .....	193
Table 8-40 Cut Scores and Associated Impact Data, Social Studies .....	194
Table 8-41 Percentage of Students in Each Performance Level by Subgroup, English Language Arts .....	195
Table 8-42 Percentage of Students in Each Performance Level by Subgroup, Mathematics .....	197
Table 8-43 Percentage of Students in Each Performance Level by Subgroup, Science .....	199
Table 8-44 Percentage of Students in Each Performance Level by Subgroup, Social Studies .....	200
Table 8-45a Summary Statistics for Content Standards Raw and SPI Scores, English Language Arts .....	201
Table 8-45b Summary Statistics for Domain Raw and SPI Scores, English Language Arts .....	204
Table 8-46 Summary Statistics for Content Standards Raw and SPI Scores, Mathematics .....	205
Table 8-47 Summary Statistics for Content Standards Raw and SPI Scores, Science .....	207
Table 8-48 Summary Statistics for Content Standards Raw and SPI Scores, Social Studies .....	208
Table 8-49 SPI Cut Scores, English Language Arts .....	209
Table 8-50 SPI Cut Scores, Mathematics .....	211
Table 8-51 SPI Cut Scores, Science .....	213
Table 8-52 SPI Cut Scores, Social Studies .....	214
Table 8-53 Longitudinal Comparison of State-Level Scale Score Means: ELA .....	215
Table 8-54 Longitudinal Comparison of State-Level Scale Score Means: Mathematics .....	215
Table 8-55 Longitudinal Comparison of State-Level Scale Score Means: Science .....	216
Table 8-56 Longitudinal Comparison of State-Level Scale Score Means: Social Studies .....	216
Table 8-57 Longitudinal Comparison of State-Level Impact Data: ELA .....	216
Table 8-58 Longitudinal Comparison of State-Level Impact Data: Mathematics .....	217
Table 8-59 Longitudinal Comparison of State-Level Impact Data: Science .....	217
Table 8-60 Longitudinal Comparison of State-Level Impact Data: Social Studies .....	217

## PART 9

Table 9-A Example Contingency Table with Three Cut Scores .....	224
Table 9-B Example Classification Table for One Cut Point ( $C_1$ ) .....	225
Table 9-1 Reliability for Total Group and Subgroups Using Cronbach's Alpha .....	230
Table 9-2 Standard Error of Measurement for Total Group and Subgroups .....	231
Table 9-3 Cronbach's Alpha Reliability Coefficients for Content Standard and Domain .....	232
Table 9-4 Standard Error of Measurement per Content Standard and Domain .....	233
Table 9-5 Classification Consistency and Classification Accuracy for English Language Arts Grade 3 .....	234
Table 9-6 Classification Consistency and Classification Accuracy for English Language Arts Grade 4 .....	234
Table 9-7 Classification Consistency and Classification Accuracy for English Language Arts Grade 5 .....	235
Table 9-8 Classification Consistency and Classification Accuracy for English Language Arts Grade 6 .....	235
Table 9-9 Classification Consistency and Classification Accuracy for English Language Arts Grade 7 .....	236
Table 9-10 Classification Consistency and Classification Accuracy for English Language Arts Grade 8 .....	236
Table 9-11 Classification Consistency and Classification Accuracy for Mathematics Grade 3 .....	237
Table 9-12 Classification Consistency and Classification Accuracy for Mathematics Grade 4 .....	237
Table 9-13 Classification Consistency and Classification Accuracy for Mathematics Grade 5 .....	238
Table 9-14 Classification Consistency and Classification Accuracy for Mathematics Grade 6 .....	238
Table 9-15 Classification Consistency and Classification Accuracy for Mathematics Grade 7 .....	239
Table 9-16 Classification Consistency and Classification Accuracy for Mathematics Grade 8 .....	239
Table 9-17 Classification Consistency and Classification Accuracy for Science Grade 4 .....	240
Table 9-18 Classification Consistency and Classification Accuracy for Science Grade 8 .....	240
Table 9-19 Classification Consistency and Classification Accuracy for Social Studies Grade 4 .....	241
Table 9-20 Classification Consistency and Classification Accuracy for Social Studies Grade 8 .....	241
Table 9-21 Classification Consistency and Classification Accuracy for Social Studies Grade 10 .....	242

Table 9-22 Inter-Rater Reliability, English Language Arts .....	243
-----------------------------------------------------------------	-----

**PART 10**

Table 10-1 Items Flagged for DIF by Gender, Focal Group: Female .....	259
Table 10-2 Items Flagged for DIF by Race/Ethnicity, Focal Group: African-American .....	260
Table 10-3 Items Flagged for DIF by Race/Ethnicity, Focal Group: Hispanic .....	261
Table 10-4 Items Flagged for DIF by Race/Ethnicity, Focal Group: Asian .....	262
Table 10-5 Items Flagged for DIF by Race/Ethnicity, Focal Group: American Indian .....	263
Table 10-6 Items Flagged for DIF by English Language Proficiency, Focal Group: Students Not English Language Proficient.....	263
Table 10-7 Items Flagged for DIF by Disability Status, Focal Group: Students with One or More Disabilities .....	264
Table 10-8 Correlations among English Language Arts Test Domains.....	264
Table 10-9 Correlations among English Language Arts Standards .....	265
Table 10-10 Correlations among Mathematics Standards .....	266
Table 10-11 Correlations among Science Standards .....	267
Table 10-12 Correlations among Social Studies Standards .....	267
Table 10-13 Principal Components Analysis .....	268
Table 10-14 Correlations between Content Area Scale Scores .....	268
Table 10-15 Correlations between Content Area Scale Scores by Gender .....	269
Table 10-16 Correlations between Content Area Scale Scores by Ethnicity/Race .....	270
Table 10-17 Correlations between Content Area Scale Scores by English Proficiency Status .....	271
Table 10-18 Correlations between Content Area Scale Scores by SES Status .....	272
Table 10-19 Correlations between Content Area Scale Scores by Disability Status .....	273
Table 10-20 Partial Correlations between Content Area Scale Scores .....	273
Table 10-21 Comparison of Spring 2015 Wisconsin NAEP and Spring 2017 Wisconsin Forward Exam Impact Data .....	274

## TABLE OF FIGURES

### PART 6

Figure 6-A Examples of Likelihood Functions, or the Probability of Each Ability Level Estimate (or Scale Score).....	75
Figure 6-1 Anchor Set TCCs: ELA Grade 3.....	108
Figure 6-2 Anchor Set TCCs: ELA Grade 4.....	108
Figure 6-3 Anchor Set TCCs: ELA Grade 5.....	109
Figure 6-4 Anchor Set TCCs: ELA Grade 6.....	109
Figure 6-5 Anchor Set TCCs: ELA Grade 7.....	110
Figure 6-6 Anchor Set TCCs: ELA Grade 8.....	110
Figure 6-7 Anchor Set TCCs: Mathematics Grade 3.....	111
Figure 6-8 Anchor Set TCCs: Mathematics Grade 4.....	111
Figure 6-9 Anchor Set TCCs: Mathematics Grade 5.....	112
Figure 6-10 Anchor Set TCCs: Mathematics Grade 6.....	112
Figure 6-11 Anchor Set TCCs: Mathematics Grade 7.....	113
Figure 6-12 Anchor Set TCCs: Mathematics Grade 8.....	113
Figure 6-13 Anchor Set TCCs: Science Grade 4.....	114
Figure 6-14 Anchor Set TCCs: Science Grade 8.....	114
Figure 6-15 Anchor Set TCCs: Social Studies Grade 4.....	115
Figure 6-16 Anchor Set TCCs: Social Studies Grade 8.....	115
Figure 6-17 Anchor Set TCCs: Social Studies Grade 10.....	116
Figure 6-18 Item Characteristic Curves for the Flagged Social Studies Grade 4 Anchor.....	116
Figure 6-19 English Language Arts Test Characteristic Curves.....	117
Figure 6-20 English Language Arts Standard Error Curves.....	117
Figure 6-21 English Language Arts Growth at Quartiles.....	118
Figure 6-22 Mathematics Test Characteristic Curves.....	119
Figure 6-23 Mathematics Standard Error Curves.....	119
Figure 6-24 Mathematics Growth at Quartiles.....	120
Figure 6-25 Science Test Characteristic Curves.....	121
Figure 6-26 Science Standard Error Curves.....	121
Figure 6-27 Science Growth at Quartiles.....	122
Figure 6-28 Social Studies Test Characteristic Curves.....	123
Figure 6-29 Social Studies Standard Error Curves.....	123
Figure 6-30 Social Studies Growth at Quartiles.....	124

## Table of Revisions

Date	Revised Section	Revision Description
May 4, 2018	Part 5, Section 5.4.1 Distribution of TDA Item Scores, pp.53, 58-59.	Added explanation of the data in Tables 5.3, 5.4, and 5.5. Revised the data in Tables 5.4 and 5.5.
May 4, 2018	Part 9, Section 9.1 Measures of Internal Consistency and Standard Error of Measurement, p.221.	Addressed the test reliability results for limited English proficiency students.
May 4, 2018	Part 9, Section 9.1.1 Conditional Standard Error of Measurement, pp.222-223.	Added a section on the conditional standard error of measurement
May 4, 2018	Part 9, Section 9.2 Classification Consistency and Accuracy, pp.226-227.	Added a summary of classification consistency and accuracy results by subgroup.
May 4, 2018	Appendix G: Classical Item Analysis Results, pp.327-362.	Added to the Technical Report
May 4, 2018	Appendix I: Conditional Standard Error of Measurement with Cut Scores, pp.371-388.	Added to the Technical Report
May 4, 2018	Appendix J: Classification Consistency and Accuracy Indices by Subgroup, pp.389-418.	Added to the Technical Report
January 7, 2019	Appendix C: Table C-2 English Language Arts, Grade 4 Test Map, p.297	Corrected the test map.

## Part 1: Overview

---

The *Wisconsin Forward Exam Spring 2017 Technical Report* documents the processes and procedures applied in the Spring 2017 test development, administration, and scoring, as well as the assessment results. This report also provides evidence in support of validity and reliability of the testing program in adherence to the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014). This report demonstrates that the Spring 2017 Wisconsin Forward Exam adhered to the appropriate standards and practices of educational assessment. Ultimately, this report provides evidence that valid inferences about Wisconsin student performance can be derived from this assessment.

### 1.1 Historical Background

The *Improving America's Schools Act* of 1994 required that states establish challenging academic standards as well as aligned annual assessments. The *Goals 2000: Educate America Act* and the *Elementary and Secondary Education Act* (ESEA) spelled out additional requirements to ensure that citizens receive coherent information about whether and to what degree students are meeting rigorous academic standards. This Technical Report is an important part of meeting those requirements.

Wisconsin students in grades 4, 8, and 10 began taking the Wisconsin Knowledge and Concepts Examination (WKCE) norm-referenced assessments in the 1997 school year. The assessments used at that time were *TerraNova*<sup>TM</sup> tests developed by CTB/McGraw-Hill (1997, 2000, 2009). The selection of those tests was partly predicated on an awareness of the academic standards being developed. In January 1998, the Wisconsin Model Academic Standards (WMAS) were adopted. These new standards were the work of the Governor's Commission on Wisconsin Model Academic Standards, chaired by then Lieutenant Governor Scott McCallum and the Wisconsin Department of Public Instruction (DPI). The assessments aligned to WMAS would measure student performance in the same subjects as the *TerraNova* tests.

Beginning in the 2005–06 school year, the federal *No Child Left Behind Act* (NCLB) required all states to test all students in Reading and Mathematics in grades 3 through 8 and once in high school (in grade 10 under Wisconsin law § 118.30). Based on the NCLB legislation, student performance, reported in terms of proficiency categories, was used to determine the Adequate Yearly Progress (AYP) of students at the school, district, and state levels.

Beginning with the school year 2007–08, states were also required to administer Science assessments at least once in grades 3–5, once in grades 6–9, and once in grades 10–12. At that time, Wisconsin students in grades 4, 8, and 10 continued to be assessed in Language Arts, Science, and Social Studies as required by state law.

It was within this policy context that the WKCE was constructed, as a criterion-referenced test, for the Fall 2005 administration, replacing the previously existing norm-referenced WKCE in Reading and Mathematics. The criterion-referenced WKCE was designed

specifically for Wisconsin students to measure their performance on the WMAS adopted by the state. These assessments were designed to evaluate students' knowledge and to measure achievement in the basic skills taught in schools at grades 3–8 and 10. The Fall 2013 WKCE was the ninth administration of these assessments and the last administration of Reading, Language Arts, and Mathematics. The assessments in Science and Social Studies under the existing WKCE model continued to be administered until Fall 2014.

A major change in the Wisconsin assessments occurred for the 2014–15 test administration. First, the English Language Arts (ELA) and Mathematics assessments were moved from the Fall testing window to the Spring testing window. Second, the new ELA and Mathematics tests for grades 3 through 8 developed for the Spring 2015 administration consisted of new Smarter Balanced Assessment Consortium (SBAC) items aligned to the Common Core State Standards (CCSS). Thus, the 2014–15 ELA and Mathematics assessments were not comparable content- and construct-wise to the assessments administered in prior years. Third, while the prior year assessments included CTB's *TerraNova* items yielding norm-referenced scores, the 2014–15 assessments did not include such items. Fourth, the regular versions of the 2014–15 assessments were administered as fixed forms in the online mode, in contrast to the previous assessments, which were all administered in the paper-and-pencil mode. Fifth, technology-enhanced item types were introduced in the 2014–15 online test administration. Last, the student test scores for ELA and Mathematics were reported on SBAC scales and the students were classified into performance levels based on SBAC cut scores. Further details on the structure and reporting of the Spring 2015 ELA and Mathematics assessments (called the Wisconsin Badger Exam) can be found at <https://dpi.wi.gov/assessment/historical/smarter>.

The ELA and Mathematics assessments underwent yet another change in the 2015–16 administration year. The Wisconsin DPI partnered with Data Recognition Corporation (DRC) to develop new ELA and Mathematics grades 3 through 8 assessments for the Spring 2016 administration. The items contained in these assessments were drawn from DRC's nationally field-tested College- and Career-Ready (CCR) item bank and aligned with Wisconsin Academic Standards for ELA and Mathematics. The new assessment program is called the Wisconsin Forward Exam, and the new ELA and Mathematics tests were administered online in Spring 2016. Since the new assessments did not contain any items from the 2014–15 Wisconsin Badger Exam tests, they were not statistically linked to the previous scales. The new reporting scales for the ELA and Mathematics tests were developed after the Spring 2016 test administration, and the new performance level cut scores were set for these assessments in the Summer of 2016.

Science (grades 4 and 8) and Social Studies (grades 4, 8, and 10) assessments have been on a different trajectory, and they continued to be aligned with the WMAS. However, the test administration for these assessments was moved from the Fall window to the Spring window for the 2015–16 administration year. The items contained in Science and Social Studies tests were mainly drawn from the pool of previously administered items and also included some new items. Several of the previously administered items were edited to improve item quality and reflect test content changes over time. Despite the fact that many Science and Social Studies items in the Spring 2016 administration came from the previous item pool, the statistical linking of the Spring 2016 forms to the previous forms was not recommended due to the change of the testing

window and the numerous changes to the items themselves. Instead, similar to what was done for the ELA and Mathematics assessments, new scales were developed for the Science and Social Studies tests under the new Wisconsin Forward Exam program. Following the new scale development, the new performance level cut scores were set for Science and Social Studies in the Summer of 2016.

Details regarding development, scaling, reporting, and standard setting for all Spring 2016 assessments are included in the *Wisconsin Forward Exam Spring 2016 Technical Report* available at <https://dpi.wi.gov/assessment/forward/resources>.

Spring 2017 was the second administration year for the Wisconsin Forward Exam in ELA, Mathematics, Science, and Social Studies assessments. The tests were developed based on the input of Wisconsin educators and with adherence to Wisconsin's standards and, with a few exceptions, consisted of items administered to Wisconsin students in Spring 2016 as part of the operational test or a field test. Previously administered operational test items served as linking items between the Spring 2016 and Spring 2017 administrations, allowing the Spring 2017 assessments to be placed on the Spring 2016 scales<sup>2</sup> using statistical equating procedures. Test equating, in turn, allows for direct comparison of student scores within a content area and for evaluation of the year-to-year student performance change. This Technical Report documents all aspects of the 2016–17 testing cycle. The structure of this Technical Report mirrors the testing cycle. A brief content summary of the report is provided later in this part of the report.

## 1.2 Uses of Test Scores

Validity is the overarching component of the Wisconsin Forward Exam program. The following excerpt is from the *Standards for Educational and Psychological Testing* (hereafter the *Standards*); (AERA, APA, & NCME, 2014):

Ultimately, the validity of an intended interpretation of test scores relies on all the available evidence relevant to the technical quality of a testing system. Different components of validity evidence . . . include evidence of careful test construction; adequate score reliability; appropriate test administration and scoring; accurate score scaling, equating, and standard setting; and careful attention to fairness for all test takers, as appropriate to the test interpretation in question. (22)

As stated by the *Standards*, the validity of a testing program hinges on the use of the test scores. Validity evidence that supports the uses of the Wisconsin Forward Exam scores is provided in this Technical Report. In this section, we examine some possible uses of the Wisconsin Forward Exam scores.

The following parts (Parts 2 through 10) of this Technical Report provide additional evidence for these uses as well as technical support for some of the interpretations and uses of test scores. The information in Parts 2 through 10 also provides a firm foundation of evidence that the Wisconsin Forward Exam measures what it is intended to measure. However, this Technical Report cannot anticipate all possible interpretations and uses of the Wisconsin Forward Exam scores. It is recommended that policy and program evaluation studies, in

accordance with the *Standards*, be conducted to support some of the uses of the Wisconsin Forward Exam scores.

The validity of a test score ultimately rests on how that test score is used. To understand whether a test score is being used properly, one must first understand the purpose of the test. The intended uses of the Wisconsin Forward Exam scores include the following:

- Identifying students' strengths and areas in need of improvement,
- Communicating expectations for all students,
- Evaluating school-, district-, and state-level programs,
- Informing stakeholders (i.e., teachers, school administrators, district administrators, DPI staff members, parents, and the public) about the status of the progress toward meeting academic achievement standards of the state, and
- Meeting the requirements of the state's accountability program.

This Technical Report refers to the use of the test-level scores (scale scores and performance levels) and standard-level (objective) scores (Standard Performance Index [SPI] scores and performance levels).

### **1.2.1 Test-Level Scores**

At the test level, an overall scale score that is based on student performance on the entire test is reported. In addition, an associated level of performance is reported. These scores indicate, in varying ways, a student's achievement in ELA, Mathematics, Science, or Social Studies. Test-level scores are reported at four levels: state, school district, school, and student.

Two types of test-level scores are reported to indicate a student's achievement on the Wisconsin Forward Exam: (1) the scale score and (2) its associated level of performance.

#### **Scale Scores**

A scale score indicating a student's performance is determined for each content area. The overall scale score for a content area quantifies the achievement being measured by the ELA, Mathematics, Science, or Social Studies test. In other words, the scale score represents the student's level of performance, where higher scale scores indicate higher levels of performance on the test and lower scale scores indicate lower levels of performance.

#### **Levels of Performance**

A student's performance on the ELA, Mathematics, Science, or Social Studies Wisconsin Forward Exam is reported in one of four levels of performance: *Below Basic*, *Basic*, *Proficient*, or *Advanced*. The cut scores for the levels of performance for all content areas were recommended by Wisconsin educators at the standard setting workshop in June 2016. The cut scores reflect the expectations of Wisconsin educators of what Wisconsin students should know and be able to do in ELA, Mathematics, Science, and Social Studies (see Part 7 of this report for a brief description of the Wisconsin Forward Exam standard setting).



## Use of Test-Level Scores

The Wisconsin Forward Exam scale scores and performance levels provide summary evidence of student achievement in ELA, Mathematics, Science, and Social Studies. Classroom teachers may use these scores as evidence of student achievement in these content areas. At the aggregate level, district and school administrators may use this information for activities such as curriculum planning. The results presented in this Technical Report provide evidence that the scale scores are valid and reliable indicators of student performance in ELA, Mathematics, Science, and Social Studies.

### 1.2.2 Standard-Level Subscores and Performance Levels

The standard-level subscores (i.e., the SPI scores) indicate student performance on a content standard and can be interpreted as an estimate of the number of items a student would be expected to answer correctly if there had been 100 similar items for a given reporting category. The SPI scores are criterion-referenced scores, in that they estimate how much a student knows in a clearly defined skill domain (i.e., the criterion). The SPI scores are computed for content standards measured by at least four items.

Based on their SPI scores, students are classified in one of the four content category performance levels: *Below Basic*, *Basic*, *Proficient*, or *Advanced*. The SPI cut scores separating these performance levels are derived as expected percentages of possible score points for a given standard (content category) for students whose total test score is at the corresponding total test cut score (*Basic*, *Proficient*, or *Advanced*).

### Use of the Standard-Level Subscores

The purpose of reporting SPI scores on the Wisconsin Forward Exam is to show the relationship between the overall achievement being measured (represented by the test score) and the skills within each of the content standards associated with the content area. Teachers may use the SPI scores for individual students as indicators of strengths and needs, but the SPI scores are best corroborated by other evidence, such as homework, class participation, diagnostic test scores, or observation. Part 3 of this Technical Report provides evidence of content validity that supports the use of the standard-level subscores. Part 10 of this Technical Report provides evidence of construct validity that further supports the use of these subscores.

District and school administrators may compare their results by content standard and grade level with the state results to better understand their strengths and needs within a particular content area and grade level. Caution should be exercised when comparing standard-level subscores across years because different items will comprise these subscores and these items may vary in difficulty between test forms or test administrations.

### **1.3 Technical Report Structure**

This Technical Report documents, in the subsequent parts, the major activities of the testing cycle. It provides comprehensive details that confirm that the processes and procedures applied in the Wisconsin Forward Exam adhere to appropriate professional standards and practices of educational assessment. Ultimately, this report provides evidence that valid inferences about Wisconsin student performance can be derived from the Wisconsin Forward Exam. An overview of the subsequent parts within this report is provided below.

#### **Part 2: Test Blueprint and Item Development**

Part 2 of this report describes the test blueprint, the item development process, and some aspects of the content-related validity of the Wisconsin Forward Exam. More specifically, it describes how DRC, DPI, and Wisconsin educators collaborated to ensure that the appropriate content was included in the Wisconsin Forward Exam and to ensure that the test items adequately sampled the domain of content knowledge necessary to make legitimate inferences about student performance. The Wisconsin Academic Standards for ELA and Mathematics were the basis of the test blueprints and item specifications for their respective content areas. For Science and Social Studies, the WMAS formed the basis for test blueprints and item specifications. Wisconsin educators were involved in reviewing the items in all contents to ensure the appropriateness of the test to the standards. The first item review occurred in December 2015 with the convention of approximately 74 educators for grades 3–8 ELA and Mathematics, grades 4 and 8 Science, and grades 4, 8, and 10 Social Studies. This item review served to establish the accessibility of the items and reading passages. Simultaneously, DRC created the test specifications documents that were later approved by DPI and will continue to serve as a foundation for item and test development. The second item review, supported by the item data acquired after the Spring 2016 test administration, occurred in August 2016 and was conducted by DPI content experts. The purpose of this review was to refine the pool of items from which the Spring 2017 operational test forms were selected.

#### **Part 3: Test Form Development**

Part 3 presents the Wisconsin Forward Exam test design and the discusses key development tasks related to creating the Spring 2017 Wisconsin Forward Exam forms. The Spring 2017 Wisconsin Forward Exam was an online assessment with a single print-on-demand form at each grade level. Student responses to the print-on-demand form were transcribed by a proctor into the online assessment system. Other variations of the forms included stacked Spanish translation forms, video sign language, and closed captioning. These were provided in an online format at each grade level.

Item selection was based on the approved test blueprints. DRC's CCR item bank contained a sufficient number of items to fulfill the test design needs for the ELA and Mathematics exams. Science and Social Studies forms were supplemented through the use of *TerraNova* items (CTB/McGraw-Hill, 2009). Part 3 also discusses the process of selecting operational test items and the process of obtaining DPI approvals. As detailed in Part 3, in addition to the operational test items, there were numerous unique field test items on each form.

Selection of the Spring 2017 test forms was done using the approved test blueprints, test designs, and psychometric specifications as guides.

#### **Part 4: Test Administration**

Part 4 briefly describes test administration and accommodations. The Wisconsin Forward Exam is a component of the Wisconsin Student Assessment System (WSAS), considered to be a comprehensive statewide program of assessments. In the 2015–16 school year, this assessment replaced the Wisconsin Badger Exam (SBAC) in the areas of ELA and Mathematics in grades 3–8 and the WKCE in the areas of Science (grades 4 and 8) and Social Studies (grades 4, 8, and 10). In the 2016–17 school year, the Wisconsin Forward Exam was administered to Wisconsin students for the second time.

Test administration was conducted over an eight-week window: March 20–May 5, 2017. All testing was conducted online, administered via DRC’s INSIGHT platform.

Part 4 of the Technical Report serves to describe the processes and activities implemented and information disseminated to help ensure standardized test administration procedures and, thus, uniform test administration conditions for students.

#### **Part 5: Scoring**

Part 5 documents the scoring process for different item types: scanning of multiple-choice (MC) and multi-select (MS) items, auto-scoring of technology-enhanced (TE), short answer (SA), evidence-based selected response (ESR) items, and artificial intelligence (AI) scoring and handscoring of text-dependent analysis (TDA) items. The description of the handscoring process includes the development and review of the scoring rubrics, anchor (sample) paper selection, training of scoring personnel, ongoing quality assurance, and a systematic review of the resulting score distributions supporting reliable and valid reported test scores. The scoring rubric used in handscoring of the TDA writing items is presented in detail.

#### **Part 6: Calibration, Equating, and Deriving Scale Scores**

The Spring 2017 administration year is the second administration year for the Wisconsin Forward Exam in all grades and content areas. Part 6 discusses characteristics of the sample of student data used for data analysis and describes the calibration, equating, and scoring methods implemented for the Wisconsin Forward Exam after the Spring 2017 test administration. The data were calibrated using two different item response theory (IRT) models, one for constructed-response items and one for MC items, which are the item types used for most large-scale standardized testing programs in education. Evaluation of the sufficiency of the IRT model results include model-to-data fit and the standard error of measurement (SEM). The equating of Spring 2017 test forms to the scales established after the Spring 2016 administration was performed using the Stocking and Lord procedure. Item-pattern scoring was applied to the Spring 2017 Wisconsin Forward Exam. As discussed in Part 6, item-pattern scoring is generally recommended over number-correct scoring because it produces more accurate scores for

individual students. Part 6 also explains how a student’s scale score is derived from the raw score using item-pattern scoring.

### **Part 7: Standard Setting**

Part 7 provides a brief overview of the standard setting process during which the performance level cut scores were set for the Wisconsin Forward Exam. The standard setting methodology and results, including performance level descriptors and cut scores, are presented.

### **Part 8: Test Results**

Part 8 summarizes results of item analyses as well as test reliability reported using Cronbach’s alpha and SEM. Summary descriptive statistics for all scores (i.e., raw scores, scale scores, SPI scores, and performance levels) are reported for the total population and for subgroups identified by gender, race/ethnicity, socioeconomic status, disability status, and English language proficiency. In addition, the longitudinal test results are presented in Part 8.

### **Part 9: Reliability**

Part 9 elaborates on the reliability of the test based on results presented in previous parts of the report. SEM was assessed for raw scores and scale scores. Inter-rater reliability was computed for TDA items on ELA tests that were scored using the AI scoring engine with human scorer verification. Internal consistency was evaluated for all tests for the total student population and for subgroups identified by gender, race/ethnicity, socioeconomic status, disability status, and English language proficiency. Classification consistency and accuracy were estimated for performance classification.

### **Part 10: Validity**

Part 10 reviews the validity evidence presented in all previous parts of the report and provides additional validity evidence supporting the Wisconsin Forward Exam. Factor analysis, correlations among content standards, and relationship between the Wisconsin Forward Exam scores and external variables are presented in the context of construct validity. An analysis of differential item functioning is presented. Forensic analysis procedures, implemented to detect possible aberrant testing behavior, are also discussed.

### **Part 11: Summary Recommendations**

Key findings of the Spring 2017 Wisconsin Forward Exam administration are presented in the body of the report. However, some items of a more technical nature, which stand out as key recommendations and summary statements that should be considered in subsequent administrations, are presented in Part 11. Recommendations based on the Spring 2017 Wisconsin Forward Exam administration cover three different phases of the testing cycle: item development; scoring; and psychometric, or measurement-based, research and evaluation.

## Part 2: Test Blueprint and Item Development

---

The purpose of this section is to describe how DRC, DPI, and Wisconsin educators collaborated through a series of test development processes to ensure that appropriate content was included in the Wisconsin Forward Exam and to ensure that test items adequately sampled the domain of content knowledge necessary to make accurate inferences about student performance. Part 2 documents the test blueprint and item development process for the Spring 2017 administration.

This part of the Technical Report is particularly relevant to AERA, APA, & NCME (2014) Standards 3.1, 3.2, and 4.0. Each of these Standards and the way each Standard is addressed will be presented in this section of the report. AERA, APA, & NCME (2014) Standard 4.0 states the following:

Tests and testing programs should be designed and developed in a way that supports the validity of interpretations of the test scores for their intended uses. Test developers and publishers should document steps taken during the design and development process to provide evidence of fairness, reliability, and validity for intended uses for individuals in the intended examinee population. (85)

DRC's College- and Career-Ready (CCR) item bank contains nationally field-tested CCR items that support the next generation of standards and assessments. It is aligned to the CCR standards in Mathematics and English Language Arts (ELA) in grades 3–8, as well as Science items aligned to Wisconsin's Model Academic Standards for Science (WMASS) and enhanced by the Next Generation Science Standards (NGSS) based on the National Research Council's Framework for K–12 Science Education. The item bank is designed to support states like Wisconsin that have adopted, or are preparing to adopt, more rigorous content standards, curricula, and assessments that better prepare students for college and careers.

Alignment to standards, grade-level appropriateness, depth of knowledge (DOK), item/task level of complexity, estimated difficulty level, relevancy of context, rationale for distractors, style, accuracy, and correct terminology were major considerations in the item development process. DRC's item development processes for the CCR item bank followed the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). DRC's item development work was and continues to be designed to produce reliable and instructionally valid tests that reflect the complete range of performance articulated in the AERA, APA, and NCME *Standards*.

Furthermore, DRC's item development work adheres to the Principles of Universal Design (Thompson, Johnstone, & Thurlow, 2002) and reflects how items and tests must lend themselves to accessibility by diverse groups of students. Members of DRC's item development team have received direct training from the National Center on Educational Outcomes (NCEO). Therefore, DRC employs the Principles of Universal Design throughout all stages of both the item development process and the test development process.

All items in the DRC CCR items bank for ELA and Mathematics were reviewed for content and for fairness not only by DRC’s content experts but also by a panel of external experts and more recently by Wisconsin educators. The external reviewers have a broad range of experience in the educational field. All the reviewers have bachelor’s-level, master’s-level, or doctoral-level degrees and teaching experience in their specific area of expertise. Table 2-1 provides a high-level sequence of the activities that occurred in the development of the DRC CCR item bank for ELA and Mathematics items.

Various item types were developed in order to best assess students’ understandings of the standards. Descriptions of each item type used in the CCR item bank are included in Table 2-2.

The efforts by DRC in developing items are in alignment with multiple best practices of the test industry and, in particular, support the following AERA, APA, & NCME (2014) Standards:

**Standard 3.1** Those responsible for test development, revision, and administration should design all steps of the testing process to promote valid score interpretations for intended score uses for the widest possible range of individuals and relevant subgroups in the intended population. (63)

**Standard 3.2** Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests’ being affected by construct-irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics. (64)

It was determined that the State of Wisconsin would license ELA, Mathematics, and Science items from DRC’s CCR item bank for the Spring 2017 test administration. Due to the state-specific nature of the Social Studies standards, DPI owns the items for that content area. Details regarding the development of the items in the CCR bank created prior to their field testing on the Forward Exam are provided in the *Wisconsin Forward Exam Spring 2016 Technical Report* as well as in the *2010 Wisconsin Knowledge and Concept (WKCE) Technical Report*. Both reports are available on the DPI website at <https://dpi.wi.gov>.

## 2.1 Test Blueprints

The test blueprints specify the number of items for each reporting category and subskill as well as the allowable DOK levels for the respective reporting categories. The process used for developing the blueprints for Wisconsin Forward Exam was a collaborative effort between DRC and DPI. The DPI-approved blueprints can be found in Tables 2-3 through 2-6.

## 2.2 Reading Passage and Item Selection for Spring 2016 Field Test

The purpose of the Spring 2016 field test was to expand the pool of items eligible for inclusion in the Spring 2017 operational test forms. ELA, Mathematics, Science, and Social

Studies passages and items field tested during the Spring 2016 test administration were selected, reviewed, and approved for placement on the Wisconsin Forward Exam in December 2015 by both DPI and Wisconsin educators. For these reviews, educators from across the state convened in Madison, Wisconsin, to review items in an online format so that items could be evaluated in the same testing engine and style in which items are presented to the student during the actual administration. The training PowerPoint presentations used at the reviews can be found in Appendix A of the *Wisconsin Forward Exam Spring 2016 Technical Report*. Details regarding the number of items for each content area taken to the item reviews may be found in Appendix B of the *Wisconsin Forward Exam Spring 2016 Technical Report*.

Using the approved test blueprints as a guide, DRC content specialists determined the focus of the items that would be taken to item review. Using an electronic tally sheet, Wisconsin educators made the determinations of standard alignment, DOK levels, and key(s). They noted any bias and sensitivity concerns and had the opportunity to determine whether items were accepted as is or accepted with revisions or to register a “dissenting view” in which the committee preference was that the item not be selected to appear on the Wisconsin Forward Exam in a field test position.

### **2.3 Field Testing**

New items were field tested in Spring 2016 during the operational test administration. Field test items were fully embedded in the operational forms, and students were not able to distinguish between the operational and field test items. The field test items were embedded in several test forms administered in each grade and content area. Each test form contained the same operational test items and unique field test items. The test forms were spiraled at the student level within a grade and a content area. A total of 214 new items (30 to 45 per grade) were field tested for ELA. A total of 96 items (16 per grade) were field tested for Mathematics. A total of 32 items (16 per grade) were field tested for Science, and a total of 52 items (16 to 20 per grade) were field tested for Social Studies in Spring 2016 test administration.

### **2.4 Statistical Analysis of Spring 2016 Field Test Data**

Following the field test data acquisition, the field test data analyses were conducted. The analyses included classical item analysis, differential functioning item (DIF) analysis, and item response theory (IRT). The classical item analysis included computation and evaluation of the following statistics: item  $p$ -values (difficulty), item-total test correlation, percentage of students selecting incorrect responses, point-biserial correlation for incorrect responses for the multiple-choice (MC) items, score point distribution for items worth more than 1 point, and omit rates for all items. More details on classical item analysis methodology is provided in Part 8 of this report.

DIF was conducted for all field test items to examine potential item bias and to determine whether item performance differences between identifiable subgroups were due to factors other than student ability, making the items unfairly difficult for a particular subgroup in the student population. DIF analyses were conducted based on gender, race/ethnicity, socioeconomic status,

disability status, and English language proficiency (ELP) groups. More details on the DIF methodology is provided in Part 10 of this report.

As the last step of the field test data analysis, the field test items were calibrated and equated to operational test scales using the IRT methodology (explained in detail in Part 6 of this report).

Item statistics are used as a means of detecting items that deserve closer scrutiny, rather than being a mechanism for automatic retention or rejection. Toward this end, a set of criteria was used as a screening tool to identify items that needed a closer review. For an item to be flagged for an additional review, the criteria included any of the following:

- $p$ -value  $<0.20$  or  $>0.90$ ,
- item-total test correlation (point biserial for MC items)  $<0.15$ ,
- positive point biserial on a distractor for an MC item,
- omit rate  $>5\%$ , and
- large DIF.

Items flagged for any of the above reasons were reviewed by the content-area specialists prior to their review by DPI.

## **2.5 Review of Items with Data**

In the preceding section, it was stated that test development content-area specialists used certain statistics from item and DIF analyses of the 2016 field test to identify items for further review. Specific flagging criteria for this purpose were specified in the previous section. Items without statistical flags were regarded as statistically acceptable and were not included in the data review. Likewise, items of extremely poor statistical quality were regarded as unacceptable and needed no further review. Such items were excluded from the Wisconsin item pool prior to the data review with DPI. The remaining flagged items were regarded by DRC content-area test development specialists and DRC psychometric specialists as needing further review. The intent was to capture all items that needed an additional review based on their statistical properties; thus, the criteria employed for item flagging tended to overidentify rather than under-identify potential item issues.

The review of the items with data was conducted by DPI staff and DRC content specialists broken out into content-area and/or grade-level groups. The data review took place in Madison, Wisconsin, September 26–27, 2016. In these sessions, reviewers were first trained by a representative from DRC’s staff with regard to the statistical indices used in item evaluation. This was followed by a discussion with examples concerning reasons that an item might be retained regardless of the statistics. The review process involved a brief exploration of possible reasons for the statistical profile of an item (e.g., possible bias, grade appropriateness, instructional issues) and a decision regarding acceptance. DRC content-area test development specialists facilitated the review of the items. Each group reviewed the pool of field-tested items and made recommendations on each item and/or scenario/passage. The training presentation used



at the data review meeting may be found in Appendix A. A summary of the data review results, including the number of items field tested, the number and percentage of items with statistical flags, and the number and percentage of items rejected by DPI during the data review, is presented in Appendix B. Items accepted for subsequent use in the Wisconsin Forward Exam were included in the pool of items for Spring 2017 operational test form selection.

## **2.6 Summary**

In summary, the items included in the Spring 2017 Wisconsin Forward Exam were reviewed by DRC, DPI, and Wisconsin educators for accessibility, bias, sensitivity, and content. During the reviews, experts identified (1) issues that could negatively affect a student's ability to access stimuli and items, (2) content in stimuli and items that could unfairly affect a student's response because of his or her background, (3) developmental appropriateness, and (4) alignment of stimuli and items to the content specifications. Item content was checked for the accuracy of the content, answer keys, and scoring rules. Items flagged for accessibility, bias and sensitivity, and/or other content concerns were removed from the Wisconsin item pool prior to the form construction for the first administration year of the Wisconsin Forward Exam. In addition, item statistics from the Spring 2016 operational and field test administration were used to refine the item pool used in selection of Spring 2017 Wisconsin Forward Exam forms.

Table 2-1 College- and Career-Ready Item Bank Development Activities

<b>DRC College- and Career-Readiness Item Bank Development Activities</b>
Establish item/passage development specifications and style guides, and prepare item writing training manuals.
Determine item development plans.
Train item writers and/or passage developers in the project requirements and specifications.
Develop passages and write items.
Review, edit, code, and track items and produce graphics.
Produce review forms for content and bias/fairness/sensitivity reviews by external reviewers.
Modify items based on external reviewers' recommendations.
Review and approve field test ready items and passages.
Develop field-test forms and administer field test.
Internally review field-test item data.
Approve items to be included in the item bank.

Table 2-2 CCR Item Bank Item Type Descriptions

Item Type	Name	Description
ESR	Evidence-Based Selected Response	Each evidence-based selected-response item has two parts, and each two-part item is designed to elicit an evidence-based response from a student who has read a literature text passage, an informational text passage, or a writing concept. In part one, which is similar to a multiple-choice item, the student analyzes a passage or writing concept and chooses the best answer from four response options. In part two, the student uses evidence from the passage or writing concept to select one or more answers based on the response to part one. Each of these items is worth one point.
MC	Multiple Choice	Each multiple-choice item has four response options, only one of which is correct. Multiple-choice items are used to assess a variety of skill levels, from short-term recall of information to inference and problem solving. Each of these items is worth one point.
MS	Multiple Select	Each multiple-select item requires a student to evaluate information presented and respond by choosing two or more correct responses. Multiple-select items can be used to assess multiple skills and concepts in both Mathematics and ELA. Each of these items is worth one point.
SA	Short Answer	Each short-answer item requires a student to enter a short numeric or algebraic response. These items are designed to assess a student’s ability to formulate a solution to a pure or applied math problem without the assistance of response options. The short-answer items are scored on a 0–1-point scale using item-specific autoscoring rules.
SCR	Short Constructed Response	Each short-constructed response item is designed to address writing through a short response as opposed to an essay. It assesses writing skills in ways a multiple-choice item cannot. The short-constructed response items are scored on a 0–2 point scale using item-specific scoring rubric.
TE	Technology Enhanced	Each technology-enhanced item is designed to elicit evidence of a broad range of student understanding. A student interacts with the enhanced features of these computer-delivered, auto-scorable test items to show understanding of skills and concepts. Item types such as drag-and-drop, hot-spot, number line and coordinate graphing, data displays, matching interaction, and drop-down menus are just some of the technology-enhanced items presented to a student. The technology-enhanced items are scored on a 0–2 point scale using item-specific scoring rules.
TDA	Text-Dependent Analysis	Each text-dependent analysis item is a text-based analysis based on a passage or a multiple-passage set that each student has read during the assessment. Both literature and informational texts are addressed through this item type. Students must draw on basic writing skills while inferring and synthesizing information from the passage in order to develop a comprehensive, holistic essay response. The demand required of a student’s reading and writing skills in response to a TDA item coincides with the similar demands required for a student to be college and career ready. The TDA prompts are scored using a holistic scoring guideline on a 1–4-point scale. This item type is supported by all Wisconsin ELA standards across all grades for both Reading Literature and Reading Informational Texts and by the Writing standards 1, 2, 3, 4, and 9 across all grades. The TDA items were scored using artificial intelligence (AI) scoring, with an appropriate level of human scoring to validate the AI algorithms for all TDA items used in the Wisconsin ELA grades 3–8 assessments.

Table 2-3 English Language Arts Test Blueprints for Grades 3–8

Domain (Reporting Category)	Total Points by Grade					
	3	4	5	6	7	8
<b>Reading</b>	20	20	20	20	20	20
Key Ideas and Details	6–10	6–10	6–10	6–10	6–10	6–10
Craft and Structure/Integration of Knowledge and Ideas	4–10	4–10	4–10	4–10	4–10	4–10
Vocabulary Use Includes Language Standards 4 and 5	2–6	2–6	2–6	2–6	2–6	2–6
Literature	about 60%	about 60%	about 60%	about 50%	about 50%	about 50%
Informational Text	about 40%	about 40%	about 40%	about 50%	about 50%	about 50%
<b>Writing/Language</b>	14	16	16	16	16	16
Text Types and Purposes	3–8	3–8	3–8	3–8	3–8	3–8
Research	3–8	3–8	3–8	3–8	3–8	3–8
Language Conventions	3–8	3–8	3–8	3–8	3–8	3–8
<b>Text-Dependent Writing</b>	12	12	12	12	12	12
Text-Dependent Analysis	12	12	12	12	12	12
<b>Listening</b>	7	8	8	8	8	8
<b>ELA Points Total</b>	<b>53</b>	<b>56</b>	<b>56</b>	<b>56</b>	<b>56</b>	<b>56</b>

Table 2-4 Mathematics Test Blueprints for Grades 3–8

Reporting Category	Total Points by Grade					
	3	4	5	6	7	8
Operations and Algebraic Thinking	8–10	9–11	8–10			
Number and Operations in Base Ten	7–9	8–10	8–10			
Number and Operations–Fractions	7–9	9–11	8–10			
Measurement and Data	9–11	9–11	9–11			
Geometry	6–8	6–8	8–10	6–8	9–11	9–11
Ratios and Proportional Relationships				6–8	7–9	
The Number System				10–12	6–8	7–9
Expressions and Equations				10–12	9–11	9–11
Statistics and Probability				9–11	10–12	7–9
Functions						9–11
<b>Mathematics Points Total</b>	<b>42</b>	<b>46</b>	<b>46</b>	<b>46</b>	<b>46</b>	<b>46</b>

Table 2-5 Science Test Blueprints for Grades 4 and 8

Reporting Category	Total Points by Grade	
	4	8
Science Connections & Nature of Science	7–10	6–9
Science Inquiry	6–9	7–10
Physical Science	5–7	5–7
Earth and Space Science	5–7	5–7
Life and Environmental Science	5–7	5–7
Science Applications and Science in Social and Personal Perspectives	6–9	6–9
<b>Science Total Points</b>	<b>40</b>	<b>40</b>

Table 2-6 Social Studies Test Blueprints for Grades 4, 8, and 10

Reporting Categories	Total Points by Grade		
	4	8	10
Geography: People, Places, and Environments	7–11	8–12	9–11
History: Time, Continuity, and Change	6–10	10–15	11–14
Political Science and Citizenship: Power, Authority, Governance, and Responsibility	5–9	5–7	11–14
Economics: Production, Distribution, Exchange, and Consumption	5–9	5–7	7–10
The Behavioral Sciences: Individuals, Institutions, and Cultures	5–9	4–6	7–10
<b>Social Studies Total Points</b>	<b>38</b>	<b>40</b>	<b>50</b>

## **Part 3: Test Form Development**

---

Part 3 of this report focuses on key development tasks related to creating the Spring 2017 Wisconsin Forward Exam operational forms. The test blueprint and item development activities described in Part 2 explain how specific development processes provided evidence to support test validity, primarily content validity, through the use of expert professional judgment from Wisconsin educators and from DRC test development specialists. The foundational documents—test blueprints and test designs—developed and approved during the initial phases of the project served as critical guides throughout development of the test forms. These documents contributed to ensuring that each test form accurately measured the content in consistent and stable ways, thus providing evidence supporting the test’s use as an indicator of student achievement of state standards. Information is provided in Part 3 relating to the following topics:

- Presentation of the detailed test design
- A general discussion of DRC’s test creation and form review process
- The process of selecting operational and field test items
- The process of obtaining DPI approvals

### **3.1 Design of the Wisconsin Forward Exam**

The following sections provide detailed test design of the content areas assessed on the Spring 2017 Wisconsin Forward Exam assessments.

#### **3.1.1 English Language Arts**

Table 3-1 highlights the details of the ELA forms, including the number of passages and items at each grade level that were used in the core and embedded field test positions. There was one common set of core items in each of the eight field test forms at each grade level. Table 3-1 also identifies the various item types that appeared on the ELA forms, including the points for item scoring. A detailed description of the item types is provided in Part 2 of this report.

The ELA section of the online Wisconsin Forward Exam was divided into four sessions: text-dependent writing prompt, writing/language, listening, and reading. Students were able to take the sessions in any order. Recommended testing times for all sessions were included in the test design document as well as in the test administration manual.

#### **3.1.2 Mathematics**

Table 3-2 shows the operational Mathematics test design. The Mathematics exams for grades 3–8 were administered in two testing sessions, with students able to take the sessions in either order. Table 3-2 also illustrates the embedded field test item count. Grade 3 had three forms with one common set of core items while grades 4–8 had four field test forms with a common set of core items within each grade level.

In grades 6–8, the first session included both a non-calculator part and a calculator part in which the use of an embedded online calculator was allowed. Once students had completed the non-calculator part of the session, they were not allowed to return to those specific items and continued on with the remainder of that session. Recommended testing times for both sessions were included in the test design document as well as in the test administration manual.

### **3.1.3 Science**

Table 3-3 presents the operational Science test design. The Science test at grades 4 and 8 included one common set of core items at each assessed grade level and twelve sets of embedded field test forms that included the use of scenarios or tasks for students to respond to. Reporting for the operational items for 2017 remained aligned to the WMASS standards. The Science exam included two sessions which could be administered in either order.

The Science test design details the number of points and recommended testing times for each grade level. These recommended testing times were also made available in the test administration manual.

### **3.1.4 Social Studies**

Table 3-4 represents the the Social Studies test design. Each grade-level exam was administered in two testing sessions, with students able to complete the sessions in either order. The Social Studies exam at grades 4, 8, and 10 included custom items developed specifically for the Wisconsin Forward Exam.

The Social Studies test design detailed in Table 3-4 portrays the number of points and recommended testing times for each grade level. These recommended testing times were also made available in the test administration manual.

## **3.2 Test Development Process**

The creation of test forms involved the expertise of multiple DRC departments and DPI. The activities that contributed to the creation of the test forms are described below. The Wisconsin Forward Exam test development complied with the following AERA, APA, & NCME (2014) standards:

**Standard 4.1** Test specifications should describe the purpose(s) of the test, the definition of the construct or domain measured, the intended examinee population, and interpretations for intended uses. The specifications should include a rationale supporting the interpretations and uses of test results for the intended purpose(s). (85)

**Standard 4.7** The procedures used to develop, review, and try out items and to select items from the item pool should be documented. (87)



**Standard 4.12** Test developers should document the extent to which the content domain of a test represents the domain defined in the test specifications. (89)

### **3.2.1 Wisconsin Forward Test Form Creation**

The DRC team worked cooperatively with DPI content and assessment specialists to select passages and prompts with associated content-specific items for the online assessments. The DRC team constructed forms that complied with the approved test blueprints and form construction guidelines. DRC used an integrated team approach to test development, including content area specialists, psychometricians, and scoring specialists working as a unit in collaboration with DPI content experts.

### **3.2.2 Item Selection**

As a first step in building the online assessments, the DRC team prepared all items that could be considered in the process in DRC's item banking system called IDEAS. The form, format, extent, and organization of items in their respective test sessions were determined in consultation with DPI.

Following preparation of all necessary materials and resources, forms construction began. Construction of the test forms themselves was a collaborative effort between DRC's integrated development team of assessment specialists, psychometric services specialists, and scoring specialists.

Before test forms were created, passages, item/performance tasks, and artwork were carefully selected. Below, we have described the process used for item selection:

- Using the pool of vendor-owned items, DRC test development specialists first selected items to match the approved test blueprints.
- DRC test development specialists checked to see that each item clearly aligned with the standards where applicable and that each item, with available item statistics, met psychometric guidelines for inclusion in the test.
- DRC test development specialists verified that each item met technical quality for well-crafted items, including that each item
  - had one clearly correct answer (or answers if multi-select);
  - used clear and concise wording;
  - was grammatically correct;
  - had an appropriate range of difficulty;
  - was free of any offensive, inappropriate, or biased content; and
  - met the Principles of Universal Design and maximum accessibility.

In addition to content requirements, the following statistical criteria were used in item selection:

- Test length and item types match the DPI-approved test design.
- Content coverage matches the DPI-approved test blueprint.
- The following items are avoided, whenever possible:

- $p$ -value  $\leq 0.20$  or  $\geq 0.90$
- Item-total test correlation  $< 0.15$
- Omit rates  $\geq 5\%$
- Poor item fit statistics (misfit flag)
- Significant DIF statistics—If an item with DIF had to be included in the test to maintain blueprint coverage, the item was examined to determine whether any content reason exists for the DIF flag (sometimes items demonstrate statistical bias but no content reason can be determined for the bias)

The statistical properties of the Spring 2016 test forms were used as targets for selection of the Spring 2017 test forms. The form selection was conducted in two phases.

In the first phase, the anchor (linking) items were selected. The anchor items are used for statistical linking of the new forms to the previous test forms on already established test scales. The anchor items on the Spring 2017 test forms were selected from the Spring 2016 operational item pool. The anchor set was selected as a “mini” version of the full operational test for each grade level and content area in regard to its length, content coverage, and psychometric properties.

The length of the anchor sets was at least one-third of the length of the total test. The items included in the anchor sets meet the same blueprint specification as the full test in regard to the percentage of score points measuring each content standard. In addition, the psychometric properties of the anchor sets matched the corresponding properties of the target forms as closely as possible. Anchor selections were reviewed and approved by a DRC psychometrician.

In the second phase of the item selection, non-anchor operational items were selected. With the exception of ELA TDA items, the non-anchor operational items came from the Wisconsin Forward Exam Spring 2016 operational and field test item pool. The TDA items selected for the Spring 2017 test administration were new items that were not previously field tested in Wisconsin. The non-anchor operational items were selected using the item selection guidelines presented earlier in this section. Full form selections were reviewed and approved by a DRC psychometrician.

After selection of all operational items, the new field test items were added to each form. In constructing the forms, the DRC content area test development specialists followed the guidelines provided below:

- Forms included adequate standards coverage, as required by test blueprints.
- No item in a form “clued” another item on that same form.
- “Clang” was avoided (i.e., distractors were unique from one another).
- Forms were ethnically diverse as needed, in terms of artwork and graphics.
- Forms included a wide range of topics and a variety of questions.
- Correct answer distributions were psychometrically sound.
- Forms did not contain any items that had been released to the public.
- DPI reviewed and gave final approval of all online test forms.

The test maps in Appendices C, D, E, and F provide details on the operational items placed on the Spring 2017 Wisconsin Forward Exam per grade and content area. The test maps include the session number, item sequence, item usage, item maximum score, standard code, and domain name. The ELA test map is included in Appendix C, the Mathematics test map is contained in Appendix D, the Science test map is provided in Appendix E, and the Social Studies test map is given in Appendix F.

### 3.2.3 Quality Reviews

Content area test development specialists and content editorial specialists reviewed items and passages for technical quality; alignment with the standards; bias, fairness, and sensitivity; depth of knowledge; estimated difficulty; and adherence to the Principles of Universal Design in all steps of the forms creation and forms review process. The aim for this team approach was to conduct a multi-tiered internal review of all passages and items prior to submission for review by DPI and then, with approval by DPI prior to submission, for external committees to ensure that all items align with Wisconsin's standards and adhere to DPI's standards for high-quality items.

DRC content and editorial teams reviewed all passages and items to ensure that they possessed

- content alignment or congruence with the knowledge and skills specified in the standards;
- a range of estimated difficulty levels;
- appropriate grade-level vocabulary, subject matter, and assumed student knowledge;
- freedom from issues or concerns regarding bias, sensitivity, or fairness;
- accessibility, following the Principles of Universal Design; and
- correct grammar, usage, and structure/format.

As a part of DRC's internal review of the items, the test development team members and graphic specialists ensured that item art could be reproduced clearly and accurately when electronically displayed and when used in the print-on-demand form.

Test specifications were reviewed to identify any potential display requirements that may present challenges in an electronic display environment. Display tolerances are impacted by line thickness, percentage of screening for shading, specialized fonts and symbols, photographs, and color. These are defined in the early stages of the item and test development process to help guide the delineation of style requirements and specifications.

Item art was produced using transparent vector graphics that allow for adjustments without the breakdown of image clarity, which is common with lower-quality formats, and provide for the online accommodation of alternate background colors. The DRC multi-tiered quality assurance process made certain converted item art was carefully compared to the original format throughout the test development and production process.

In reviewing forms in the online environment, multiple reviewers checked passages and items on multiple electronic platforms on which students took the test to ensure a smooth testing experience.

### **3.3 DPI Approvals**

The phases during which DPI had the opportunity to review passages and items to be placed on the Spring 2017 Wisconsin Forward Exam included

- prior to item content review,
- at item content review, and
- during forms construction.

Prior to the opening of the testing window, all online forms were made accessible to DPI for review in DRC's secure INSIGHT testing engine.

### **3.4 Summary**

In summary, the efforts and procedures used in the development of the Spring 2017 Wisconsin Forward Exam balanced the content and psychometric requirements for the form development. The content of the Spring 2017 test forms adhered to the test blueprint requirements. The psychometric properties of the new test forms were comparable to the psychometric properties of the Spring 2016 forms. Overall, the process implemented in the Spring 2017 operational form development was in alignment with multiple best practices of the test industry.

Table 3-1 English Language Arts Test Design

Test Design		Grade					
		3	4	5	6	7	8
Number of Passage Sets	Literature	2-3	2-3	2-3	2-3	2-3	2-3
	Informational	1-2	1-2	1-2	2-3	2-3	2-3
	Listening	2-3	2-3	2-3	2-3	2-3	2-3
Number of Core (OP) Items	Item Type: SR/TE (1 pt.)	23	28	28	24	26	26
	Item Type: SR/TE/EBSR (2 pts)	9	8	8	10	9	9
	Item Type TDA (12 pts)	1	1	1	1	1	1
	Total Core Items	33	37	37	35	36	36
Total Core Points		53	56	56	56	56	56
Embedded Field Test (FT)	Number of Forms	8	8	8	8	8	8
	Passages	1	1	1	1	1	1
	FT Items per Form	8	8	8	8	8	8
	Total Items Field Tested	59	58	59	59	59	58
Total Items (Core + FT) per Form		41	45	45	43	44	44
Total Estimated Testing Time (minutes)		125	125	125	125	125	125

Table 3-2 Mathematics Test Design

Test Design		Grade					
		3	4	5	6	7	8
Number of Core (OP) Items	Item Type: MC/ESR/SA (1 pt.)	40	43	41	42	40	41
	Item Type: TE (1 pt.)	2	3	5	4	6	5
	<b>Total Core Items</b>	42	46	46	46	46	46
<b>Total Core Points</b>		42	46	46	46	46	46
Embedded Field Test (FT)	Number of Forms	3	4	4	4	4	4
	FT Items per Form	8	8	8	8	8	8
	<b>Total Items Field Tested</b>	24	32	32	32	32	32
<b>Total Items (Core + FT) per Form</b>		50	54	54	54	54	54
<b>Total Estimated Testing Time (minutes)</b>		90	90	90	105	105	105

Table 3-3 Science Test Design

Test Design		Grade	
		4	8
Number of Core (OP) Items	Item Type: SR (1 pt.)	40	40
Total Core Points		40	40
Embedded Field Test (FT)	Number of Forms	12	12
	Scenarios/Tasks	12	12
	FT Items per Form	8	8
	Total Items Field Tested	89	90
Total Items (Core + FT) per Form		48	48
Total Estimated Testing Time (minutes)		100	100

Table 3-4 Social Studies Test Design

Test Design		Grade		
		4	8	10
Number of Core (OP) Items	Item Type: SR (1 pt.)	38	40	50
Total Core Points		38	40	50
Embedded Field Test (FT)	Number of Forms	4	4	4
	FT Items per Form	8	8	10
	Total Items Field Tested	32	32	40
Total Items (Core + FT) per Form		46	48	60
Total Estimated Testing Time (minutes)		90	90	90

## Part 4: Test Administration

---

In the Spring of 2017, Wisconsin administered assessments in ELA and Mathematics for grades 3–8. Science was administered in grades 4 and 8 and Social Studies in grades 4, 8, and 10. The test administration window was March 20–May 5, 2017. Part 4 of the Technical Report describes a set of standardized procedures and policies applied to administer the Wisconsin Forward Exam. The issue of test security in test administration which has important implications for the integrity of the results and thus the validity of Wisconsin Forward Exam scores is also discussed. Documentation citing the written procedures provided to test administrators and school personnel in order to standardize the administration of the test are provided in this part as well. The following American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) (2014) standards are addressed in Part 4: 4.15, 4.16, 6.1, 6.2, 6.3, 6.4, 6.6, and 6.7. Each standard will be explicated within the relevant section of this part of the report.

DPI is committed to the proposition that all schools and all students within schools will be held accountable to a common set of high academic content standards, the Wisconsin Academic Standards. As an alternate assessment for students being instructed using alternate academic achievement standards, the Wisconsin Essential Elements. The Dynamic Learning Maps (DLM) assessment measures the academic progress of students with the most significant cognitive disabilities in the subject areas of ELA and Mathematics at grades 3–11, and Science at grades 4 and 8–11. A teacher rater form is used to assess these students in Social Studies at grades 4, 8, and 10.

All other students are accountable to the grade-level knowledge and skills outlined in the Wisconsin Academic Standards. Those students who have an Individualized Education Program (IEP)—a 504 plan (under Section 504 of the Rehabilitation Act of 1973)—or are identified as limited English proficient (LEP) or formerly limited English proficient (FLEP) may be eligible to receive testing accommodations. Accommodations are changes in the routine conditions under which a student takes an assessment in order to provide the student an equal opportunity to demonstrate his or her knowledge. Accommodations provided to a student must be documented in his or her current IEP and used as a component of his or her regular instructional setting. DPI guidance makes it clear that the accommodations or supports provided to a student must be consistent for classroom instruction, classroom assessments, and district and state assessments. It is important to note that while some accommodations or supports may be appropriate for instructional use, they may not be appropriate for use on a standardized assessment. AERA, APA, & NCME (2014) Standard 6.2 states the following:

When formal procedures have been established for requesting and receiving accommodations, test takers should be informed of these procedures in advance of testing. (115)

An overview of the types of accommodations and guidelines for test administration conditions are described below. Additionally, IEP teams were directed to the Wisconsin Forward Exam Accommodations and Supports page at <http://dpi.wi.gov/assessment/forward/accommodations> for guidance regarding all available



accommodations and supports intended to provide equitable access to grade-level content and assessments.

Test administrators indicated which accommodations were to be available for use by each student within the student learning profile in DRC's eDIRECT system. All student accommodations are managed and can be monitored through DRC's eDIRECT system. This system is the interface to the administrative functions of the DRC INSIGHT Online Learning System, where students interface with their online assessments. As a function of this roles-based system, the primary users of eDIRECT were District Assessment Coordinators and School Assessment Coordinators who were approved by DPI and assigned permissions accordingly for security purposes. The major functions are those of managing users and managing students. As such, eDIRECT was used to manage and update student information, including demographic and accommodations/accessibilities information. All eDIRECT user roles and permission levels were approved by DPI.

#### **4.1 Accessibility Resources**

Accommodations were allowed for eligible individual students participating in the Wisconsin Forward Exam. Accommodations provided to a student must be documented in a current IEP and used during routine instruction. IEP teams were directed to refer to the Wisconsin Forward Exam accommodations policy and guidance at <http://dpi.wi.gov/assessment/forward/accommodations>.

It is important to note that students were provided access to a range of supports that included universal tools (available to all students), designated supports, and accommodations, including the Braille version of the Wisconsin Forward Exam, based on their needs. Those are defined as follows.

##### **4.1.1 Universal Tools**

Universal tools are accessibility features that are available to all students based on student preference and selection. These access features of the assessment are either provided as digitally-delivered components of the test administration system (embedded) or separate from it (non-embedded).

##### **4.1.2 Designated Supports**

Designated supports are those features that are available for use by any student for whom the need has been indicated by an educator or team of educators (with parent/guardian and student input as appropriate) and are part of the students classroom instruction. They are either provided as part of the online test administration system or separate from it (embedded or non-embedded). All designated supports (embedded and non-embedded) must be entered into eDIRECT prior to test administration. Embedded and non-embedded supports will appear on student test tickets.

### **4.1.3 Accommodations**

Accommodations are features that increase equitable access but do not compromise the grade-level standard or intended outcome of the assessment, and are available for students for whom there is documentation of the need in the Individualized Education Program (IEP) or 504 accommodation plan, and who use a similar accommodation as part of their classroom instruction. Accommodations are either provided as part of the test administration system or separate from it (embedded or non-embedded). All accommodations must be entered into eDIRECT prior to test administration. Embedded and non-embedded accommodations will appear on student test tickets.

#### **Embedded Universal Tools (online)**

- Pause
- Breaks
- Sticky Notes
- Highlighter
- Keyboard Navigation
- Flag/Mark for Review
- Review Page
- Measuring Tools (Math)
- Cross-off Tool (Strikethrough)
- Magnifier Tool (Zoom)
- Help/What's This?
- Click to Enlarge
- Go to Question
- Tool Tips
- Test Directions

#### **Embedded Designated Supports (online)**

- Color Choices
- Contrasting Color
- Reverse Contrast
- Masking
- Text-to-Speech
- Spanish Translations (Stacked)

#### **Embedded Accommodations (online)**

- Closed Captioning
- Visual Sign Language (online VSL delivery)
- Braille
- Text-to-Speech (reading passages)
- Print-on-Demand

## **Non-Embedded Universal Tools, Designated Supports, and Accommodations**

- Pause (Breaks)
- Scratch Paper
- Word-to-Word Bilingual Dictionary
- Color Overlay
- Magnification
- Noise Buffers
- Read Aloud
- Scribe
- Separate Setting
- Abacus
- Alternate Response Options
- Multiplication Table
- Used translation
- Used Braille
- Used assistive device (e.g., text-talker, adaptive keyboard, picture symbols)
- Used a print-on-demand, paper-based version of the Wisconsin Forward Exam
- Used another DPI-approved accommodation
- Used a non-allowed accommodation resulting in the invalidation of test results

### **4.1.4 Translation**

For the Spring 2017 Wisconsin Forward Exam administration, the State of Wisconsin used Spanish translation scripts. The aim of these scripts is to better help students demonstrate their knowledge on the Wisconsin Forward Exam when English language is part of the test construct. Students whose native language is Spanish were given the choice to use all or parts of the translation accommodation, which included a bilingual word list of commonly used content area vocabulary, translation of the test directions, and a written translation script of Mathematics, Science, and Social Studies test items. DPI recommended that educators also consult the list of allowable accommodations (referenced above) to create the most appropriate testing situation for their students.

DPI recognizes that approximately 5 percent of the Wisconsin limited English proficient population speaks a language other than Spanish, and specific guidelines are provided for these students. Districts that serve students who speak languages other than Spanish may have used qualified translators to provide oral translation support to students. However, the use of translation support was restricted to Mathematics, Science, and Social Studies tests, given that the test constructs are not specific to the English language.

#### 4.1.5 Additional Accessibility Resources

Additional accessibility resources guidance available at the testing sites included the following:

- **Multiplication Table:** This resource is a non-embedded accommodation available for students who have it in their IEP or 504 plan for grades 4–8 Mathematics.
- **Read Aloud Guidelines:** This document outlines the qualifications, guidelines, and procedures required for a test reader. The test reader must sign the Read Aloud Agreement to Maintain Security and Confidentiality prior to test administration. Completed agreement forms should be retained by the Site Assessment Coordinator.
- **Scribing Guidelines:** This document outlines the qualifications, guidelines, and procedures required when using a scribe.
- **Interpreter Guidelines:** This document outlines the qualifications, guidelines, and procedures required when using an interpreter.

Tables 4-1 through 4-7 provide the list of accommodations or designated supports made available for the Spring 2017 Wisconsin Forward Exam along with the number and percentage of students provided these accommodations or supports. The counts are based on the accommodations and designated supports selected via the eDIRECT portal.

#### 4.2 Reporting Results of Assessments Taken with Accommodations

Scores of assessments taken with accommodations were included with the results for students who took these tests under standard conditions and presented at the school, district, and state levels.

#### 4.3 Test Security

Maintaining the security of all test materials is crucial to preventing the possibility of random or systematic errors, such as unauthorized exposure of test items that would affect the valid interpretation of test scores. Several test security measures have been implemented for the Wisconsin Forward Exam with compliance to the following AERA, APA, & NCME (2014) standards:

**Standard 6.6** Reasonable efforts should be made to ensure the integrity of test scores by eliminating opportunities for test takers to attain scores by fraudulent or deceptive means. (116)

**Standard 6.7** Test users have the responsibility of protecting the security of test materials at all times. (117)

The primary goal of test security is to protect the integrity of the assessments and ensure that scores retain their interpretability. To ensure that trends in achievement results can be

calculated across years and to provide longitudinal data, a certain number of test questions must be repeated from year to year. If any of these questions are made public, the validity of the test may be compromised. Because the Wisconsin Forward Exam is administered virtually 100 percent online, printed test materials are limited to the very few cases where a student requires a printed version of the test as provided in the IEP (Braille and Print-on-Demand), so the assessment exposure is limited to those educators who required access for those purposes. DPI and DRC ensured that all who had access to any materials associated with the Wisconsin Forward Exam understood the critical need for test security. They presented security requirements during the Pre-Test Workshops and outlined the acceptable and unacceptable test preparation and administration practices. The Wisconsin Forward Exam was administered under secure testing conditions established by DPI.

Other security measures for Wisconsin Forward Exam test administrations are described below.

- The use of any unauthorized electronic device is prohibited during testing.
- Password-protected, role-based administrator access to all test setup, management, and reporting functions is required.
- Student Test Login Tickets provide secure student access to the test using a unique username and password.
- Test content is securely transferred using leading encryption technologies; content is decrypted when the student login is validated.
- Decrypted test content is purged from the system's memory upon completion of the test session.
- Device lockdown during testing prevents students from copying, pasting, printing, and accessing other applications.
- If the test is paused, content is removed from the screen to ensure security of test content. The system will time out and close the test after a defined period of inactivity.
- Extensive SQA tests ensure that all data are scanned, captured, and accurately scored in the secure database and all associated reports contain accurate data.

The online systems provided by DRC that are associated with the administration of the Wisconsin Forward Exam have all been designed to provide the level of security required by DPI and described in the DPI Test Security Manual for its assessment programs. Student testing environments are designed to ensure the protection of responses as well as student data (as required under the federal Family Educational Rights and Privacy Act). DRC's information security policies and procedures are based on the National Institute of Standards and Technology (NIST) criteria (NIST Standard 800-53). This is a nationally recognized standard for information security practices.

#### **4.3.1 Secure Student Access**

Students are required to provide a valid username and password to access the online testing system. The test administrator provides each student with a Student Test Login Ticket,

which contains the student's username and a unique, pre-generated password. A separate, unique password is generated for each assessment, ensuring that students can only access the content designated for that particular test. Passwords are generated randomly for each student to use. Test tickets are generated from within the eDIRECT secure administrative system, which is pre-populated with student records. As an additional security measure, upon logging in, a Student Verification Page prompts the student to verify his or her profile information, including any assigned accommodations, prior to initiating the test. The student's name is also displayed on the screen during the test, providing an additional verification check for the student and the test administrator.

Test tickets and rosters are considered secure materials. As such, it is recommended that test tickets be printed as close to the date of testing as possible, and sites are instructed that test tickets and rosters should be kept in a secure location until the session is scheduled to begin. Test tickets are distributed just prior to student login and are collected after all students have logged in and begun testing; directions also include a request to count the number of tickets that are distributed and collected after sign in to make sure the numbers of tickets are the same. After a testing session is complete, all test tickets are returned to the Site Assessment Coordinator for secure destruction or secure storage.

#### **4.3.2 Test Security during Breaks**

Test security must be maintained during all breaks within a testing session. To lessen the risk of a security breach occurring during these breaks, students requiring the use of restroom facilities must be escorted by either a proctor or a test examiner. In addition, students must not be allowed to use any form of wireless communication during these breaks.

#### **4.4 Test Administration Training**

Training workshops for district and school assessment personnel for the Spring 2017 administration of the Wisconsin Forward Exam were conducted by DPI and DRC staff. The purpose of the training workshops and the ancillary materials was to keep districts and schools informed about policies and procedures related to the Wisconsin Forward Exam administration. The information covered during the workshops included standardizing the administration of the Wisconsin Forward Exam, maintaining the security of the assessment, allowing access to the assessments for special populations by providing appropriate designated supports or accommodations, and providing guidance on appropriate interpretations of the test results. These communication and training efforts by DPI and the ancillary information developed by DRC are in alignment with multiple best practices of the testing industry and, in particular, support the following *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014):

**Standard 4.15** The directions for test administration should be presented with sufficient clarity so that it is possible for others to replicate the administration conditions under which the data on reliability, validity, and (where appropriate) norms were obtained. Allowable variations in administration procedures should be clearly described. The

process for reviewing requests for additional testing variations should also be documented. (90)

**Standard 4.16** The instructions presented to test takers should contain sufficient detail so that test takers can respond to a task in the manner that the test developer intended. When appropriate, sample materials, practice or sample questions, criteria for scoring, and a representative item identified with each item format or major area in the test's classification or domain should be provided to the test takers prior to the administration of the test, or should be included in the testing material as part of the standard administration instructions. (90)

**Standard 6.1** Test administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer and any instructions from the test user. (114)

**Standard 6.2** When formal procedures have been established for requesting and receiving accommodations, test takers should be informed of these procedures in advance of testing. (115)

**Standard 6.3** Changes or disruptions to standardized test administration procedures or scoring should be documented and reported to the test user. (115)

**Standard 6.4** The testing environment should furnish reasonable comfort with minimal distractions to avoid construct-irrelevant variance. (116)

In order to ensure standardized testing administration for all students, a Guide for District Assessment Coordinators and School Assessment Coordinators was made available to all assessment coordinators. The guide included the following topics:

- Responsibilities of District Assessment Coordinators (DACs),
- Responsibilities of School Assessment Coordinators (SACs),
- Responsibilities of District Technology Coordinators,
- Responsibilities of Test Administrators (TA)/Proctors,
- Test Times and Schedules,
- Test Security,
- Testing Procedures,
- Accessibility Information,
- Before Online Testing,
- Technology Resources,
- Additional Materials,
- After Online Testing,
- Packaging the Test Materials,
- Procedures for Returning Materials,
- Test Results, and
- Checklists for Responsible Parties (DACs, SACs, TAs).

In addition, Test Administration Manuals were made available to all test administrators. The manuals included the following:

- Test Administrator (TA)/Proctor Responsibilities,
- Test Times/Schedules,
- Test Security,
- Accessibility Information,
- Before Testing,
- Test Tickets,
- Testing Materials,
- Setting Up Testing Environment,
- During Online Testing, and
- After Testing.

These topics were also addressed in the face-to-face training workshops held across the state, and subsequently posted for online access and review.

### **Student Preparation for Online Testing**

Prior to testing, sites were encouraged to provide students with time to complete both a tutorial video series and an online tools training.

### **Student Tutorial Video**

The Student Tutorial video was available for students and test administrators to become familiar with the online testing environment. The video is broken into multiple chapters. A table was provided to help educators determine which chapters students should view and the time required for each. Tutorials could be viewed as a class or at an individual student machine by launching INSIGHT and clicking on DRC INSIGHT Online Assessment Tutorials.

### **Online Tools Training**

The Online Tools Training (OTTs) are provided for students to allow them a hands-on opportunity to practice the types of items and tools available in the online testing system. OTTs are available publicly for practice using a Chrome browser. Users (at home or school) could visit <https://dpi.wi.gov/assessment/forward/sample-items> to access the public OTTs. OTTs could also be accessed on student testing devices once INSIGHT was installed. General OTTs were made available for each content area and grade level. Separate OTTs were available for students to practice using Video Sign Language (VSL), Text-to-Speech (TTS), Spanish translation, Masking, Color Choice, and Closed Captioning tools. VSL and Spanish OTTs were available by grade band (3–5, 6–8, and 10). The OTT was not scored and was not intended for content practice.



## **Administration Supports Before and Following Testing**

With a few exceptions (accommodated student versions), the Wisconsin Forward Exam was administered fully online. Because DRC produced a variety of Wisconsin-specific manuals with process reviews by DRC program management staff, DRC editorial staff, and DPI staff, substantial consideration was given to the information required for successful online testing to occur. DPI provided a final signoff for each document prior to delivery and public posting.

Table 4-8 displays a list of electronic materials that DRC developed in conjunction with DPI. A final PDF of each deliverable was provided to DPI to post to the DPI informational website to allow districts to review and/or print.

For additional or specific information related to test administration, refer to the Test Coordinator's Guide and/or the Test Administration Manuals that are available online at <https://dpi.wi.gov/assessment>.

### **4.5 Summary**

This part of the report summarizes the processes and activities implemented and information disseminated to help ensure standardized test administration procedures and, thus, uniform test administration conditions for students. It describes how the test administration procedures implemented for the Wisconsin Forward Exam were in alignment with best practices of the testing industry.

Table 4-1 Number and Percentage of Students Using Accommodations or Designated Supports:  
Grade 3

Grade 3 Accommodation or Support	English Language Arts		Mathematics	
	N Count	Percent	N Count	Percent
Used Braille [BRL]	7	0.01	8	0.01
Used Print-on-Demand [POD]	1	0.00	1	0.00
Used Bilingual Dictionary			546	0.85
Used Magnification	172	0.27	172	0.27
Used Noise Buffers	911	1.42	910	1.42
Used Read Aloud	1992	3.12	2339	3.65
Used Scribe	731	1.14	684	1.07
Used Separate Setting	7230	11.31	7251	11.32
Used Alternate Response Options	27	0.04	29	0.05
Used Read Aloud (Reading Passages)	487	0.76	488	0.76
Provided Color Choices [CC]	368	0.58	373	0.58
Used Contrasting Color [CTC]	357	0.56	362	0.57
Used Reverse Contrast [RC]	289	0.45	290	0.45
Used Masking [MSK]	937	1.47	927	1.45
Used Text-to-Speech [TTS]	10162	15.89	11438	17.85
Used Spanish Translation [ST]	664	1.04	988	1.54
Used Video Sign Language [VSL (ASL)]	13	0.02	19	0.03
Used Color Overlay	29	0.05	28	0.04
Used Closed Captioning [C CAP] ELA	51	0.08		
Used Listening Scripts [LS] ELA	7	0.01		
Used Text-to-Speech for Reading Passages [TTS (PSGS)] ELA	739	1.16		
Used Abacus Math			39	0.06
Used Non-Embedded Calculator Math			235	0.37
Used Multiplication Table Math			1145	1.79

Table 4-2 Number and Percentage of Students Using Accommodations or Designated Supports:  
Grade 4

Grade 4 Accommodation or Support	English Language Arts		Mathematics		Science		Social Studies	
	N Count	Percent	N Count	Percent	N Count	Percent	N Count	Percent
Used Braille [BRL]	5	0.01	5	0.01	5	0.01	4	0.01
Used Print-on-Demand [POD]	3	0.00	4	0.01	4	0.01	3	0.00
Used Bilingual Dictionary			440	0.68	437	0.68	438	0.68
Used Magnification	158	0.25	152	0.24	151	0.23	150	0.23
Used Noise Buffers	865	1.34	851	1.32	840	1.30	840	1.30
Used Read Aloud	1917	2.98	2212	3.43	2134	3.31	2128	3.30
Used Scribe	804	1.25	768	1.19	760	1.18	757	1.17
Used Separate Setting	7751	12.03	7794	12.08	7626	11.82	7613	11.80
Used Alternate Response Options	13	0.02	13	0.02	13	0.02	13	0.02
Used Read Aloud (Reading Passages)	492	0.76	492	0.76	492	0.76	492	0.76
Provided Color Choices [CC]	388	0.60	388	0.60	386	0.60	385	0.60
Used Contrasting Color [CTC]	372	0.58	373	0.58	372	0.58	372	0.58
Used Reverse Contrast [RC]	299	0.46	299	0.46	298	0.46	298	0.46
Used Masking [MSK]	968	1.50	976	1.51	967	1.50	964	1.49
Used Text-to-Speech [TTS]	9837	15.27	11481	17.79	11251	17.44	11238	17.42
Used Spanish Translation [ST]	579	0.90	745	1.15	667	1.03	667	1.03
Used Video Sign Language [VSL (ASL)]	16	0.02	22	0.03	22	0.03	22	0.03
Used Color Overlay	35	0.05	34	0.05	34	0.05	35	0.05
Used Closed Captioning [C CAP] ELA	59	0.09						
Used Listening Scripts [LS] ELA	7	0.01						
Used Text-to-Speech for Reading Passages [TTS (PSGS)] ELA	889	1.38						
Used Abacus Math			50	0.08				
Used Non-Embedded Calculator Math			343	0.53				
Used Multiplication Table Math			2156	3.34				

Table 4-3 Number and Percentage of Students Using Accommodations or Designated Supports:  
Grade 5

Grade 5 Accommodation or Support	English Language Arts		Mathematics	
	N Count	Percent	N Count	Percent
Used Braille [BRL]	4	0.01	4	0.01
Used Print-on-Demand [POD]	4	0.01	2	0.00
Used Bilingual Dictionary			376	0.60
Used Magnification	137	0.22	131	0.21
Used Noise Buffers	805	1.28	799	1.27
Used Read Aloud	1678	2.66	1953	3.09
Used Scribe	705	1.12	653	1.03
Used Separate Setting	7167	11.38	7197	11.40
Used Alternate Response Options	17	0.03	16	0.03
Used Read Aloud (Reading Passages)	460	0.73	459	0.73
Provided Color Choices [CC]	328	0.52	331	0.52
Used Contrasting Color [CTC]	321	0.51	322	0.51
Used Reverse Contrast [RC]	280	0.44	282	0.45
Used Masking [MSK]	952	1.51	942	1.49
Used Text-to-Speech [TTS]	9018	14.32	10485	16.60
Used Spanish Translation [ST]	471	0.75	650	1.03
Used Video Sign Language [VSL (ASL)]	15	0.02	21	0.03
Used Color Overlay	49	0.08	49	0.08
Used Closed Captioning [C CAP] ELA	58	0.09		
Used Listening Scripts [LS] ELA	4	0.01		
Used Text-to-Speech for Reading Passages [TTS (PSGS)] ELA	895	1.42		
Used Abacus Math			31	0.05
Used Non-Embedded Calculator Math			384	0.61
Used Multiplication Table Math			2401	3.80

Table 4-4 Number and Percentage of Students Using Accommodations or Designated Supports:  
Grade 6

Grade 6 Accommodation or Support	English Language Arts		Mathematics	
	N Count	Percent	N Count	Percent
Used Braille [BRL]	2	0.00	2	0.00
Used Print-on-Demand [POD]	3	0.00	3	0.00
Used Bilingual Dictionary			254	0.40
Used Magnification	105	0.17	102	0.16
Used Noise Buffers	610	0.97	605	0.96
Used Read Aloud	1344	2.14	1490	2.37
Used Scribe	508	0.81	468	0.74
Used Separate Setting	6503	10.36	6522	10.38
Used Alternate Response Options	14	0.02	14	0.02
Used Read Aloud (Reading Passages)	465	0.74	464	0.74
Provided Color Choices [CC]	494	0.79	492	0.78
Used Contrasting Color [CTC]	434	0.69	434	0.69
Used Reverse Contrast [RC]	385	0.61	386	0.61
Used Masking [MSK]	2223	3.54	2122	3.38
Used Text-to-Speech [TTS]	7736	12.33	9062	14.42
Used Spanish Translation [ST]	199	0.32	262	0.42
Used Video Sign Language [VSL (ASL)]	20	0.03	25	0.04
Used Color Overlay	33	0.05	33	0.05
Used Closed Captioning [C CAP] ELA	67	0.11		
Used Listening Scripts [LS] ELA	2	0.00		
Used Text-to-Speech for Reading Passages [TTS (PSGS)] ELA	922	1.47		
Used Abacus Math			22	0.04
Used Non-Embedded Calculator Math			748	1.19
Used Multiplication Table Math			2692	4.28

Table 4-5 Number and Percentage of Students Using Accommodations or Designated Supports:  
Grade 7

Grade 7 Accommodation or Support	English Language Arts		Mathematics	
	N Count	Percent	N Count	Percent
Used Braille [BRL]	5	0.01	4	0.01
Used Print-on-Demand [POD]	4	0.01	4	0.01
Used Bilingual Dictionary			218	0.34
Used Magnification	91	0.14	92	0.15
Used Noise Buffers	517	0.82	510	0.81
Used Read Aloud	1001	1.59	1154	1.83
Used Scribe	310	0.49	281	0.44
Used Separate Setting	6435	10.20	6487	10.26
Used Alternate Response Options	10	0.02	10	0.02
Used Read Aloud (Reading Passages)	357	0.57	356	0.56
Provided Color Choices [CC]	484	0.77	485	0.77
Used Contrasting Color [CTC]	451	0.71	451	0.71
Used Reverse Contrast [RC]	410	0.65	410	0.65
Used Masking [MSK]	2076	3.29	2073	3.28
Used Text-to-Speech [TTS]	7245	11.48	8440	13.35
Used Spanish Translation [ST]	184	0.29	272	0.43
Used Video Sign Language [VSL (ASL)]	8	0.01	19	0.03
Used Color Overlay	44	0.07	43	0.07
Used Closed Captioning [C CAP] ELA	51	0.08		
Used Listening Scripts [LS] ELA	16	0.03		
Used Text-to-Speech for Reading Passages [TTS (PSGS)] ELA	849	1.35		
Used Abacus Math			9	0.01
Used Non-Embedded Calculator Math			854	1.35
Used Multiplication Table Math			2510	3.97

Table 4-6 Number and Percentage of Students Using Accommodations or Designated Supports: Grade 8

Grade 8 Accommodation or Support	English Language Arts		Mathematics		Science		Social Studies	
	N Count	Percent	N Count	Percent	N Count	Percent	N Count	Percent
Used Braille [BRL]	3	0.00	3	0.00	3	0.00	3	0.00
Used Print-on-Demand [POD]	5	0.01	5	0.01	5	0.01	5	0.01
Used Bilingual Dictionary			269	0.43	264	0.43	263	0.42
Used Magnification	94	0.15	87	0.14	87	0.14	87	0.14
Used Noise Buffers	416	0.67	410	0.66	404	0.65	408	0.66
Used Read Aloud	804	1.29	922	1.48	905	1.46	904	1.46
Used Scribe	242	0.39	228	0.37	229	0.37	229	0.37
Used Separate Setting	6210	10.00	6228	10.02	6110	9.84	6071	9.78
Used Alternate Response Options	13	0.02	13	0.02	12	0.02	12	0.02
Used Read Aloud (Reading Passages)	303	0.49	302	0.49	302	0.49	303	0.49
Provided Color Choices [CC]	476	0.77	457	0.74	457	0.74	456	0.73
Used Contrasting Color [CTC]	382	0.62	383	0.62	383	0.62	383	0.62
Used Reverse Contrast [RC]	335	0.54	333	0.54	333	0.54	333	0.54
Used Masking [MSK]	1793	2.89	1793	2.88	1791	2.88	1789	2.88
Used Text-to-Speech [TTS]	6750	10.87	7910	12.72	7710	12.41	7747	12.48
Used Spanish Translation [ST]	197	0.32	276	0.44	268	0.43	269	0.43
Used Video Sign Language [VSL (ASL)]	16	0.03	25	0.04	25	0.04	25	0.04
Used Color Overlay	34	0.05	33	0.05	32	0.05	32	0.05
Used Closed Captioning [C CAP] ELA	65	0.10						
Used Listening Scripts [LS] ELA	3	0.00						
Used Text-to-Speech for Reading Passages [TTS (PSGS)] ELA	844	1.36						
Used Abacus Math			8	0.01				
Used Non-Embedded Calculator Math			865	1.39				
Used Multiplication Table Math			2095	3.37				

Table 4-7 Number and Percentage of Students Using Accommodations or Designated Supports:  
Grade 10

<b>Grade 10</b>	<b>Social Studies</b>	
	<b>N Count</b>	<b>Percent</b>
Used Braille [BRL]	2	0.00
Used Print-on-Demand [POD]	1	0.00
Used Bilingual Dictionary	109	0.17
Used Magnification	56	0.09
Used Noise Buffers	49	0.08
Used Read Aloud	505	0.79
Used Scribe	78	0.12
Used Separate Setting	3792	5.95
Used Alternate Response Options	9	0.01
Provided Color Choices [CC]	273	0.43
Used Contrasting Color [CTC]	266	0.42
Used Reverse Contrast [RC]	246	0.39
Used Masking [MSK]	437	0.69
Used Text-to-Speech [TTS]	3283	5.15
Used Spanish Translation [ST]	101	0.16
Used Video Sign Language [VSL (ASL)]	15	0.02
Used Color Overlay	14	0.02



Table 4-8 Summary Table of Manual Materials

Material	Configuration
<p><b>DAC/SAC Guide (District Assessment Coordinator/School Assessment Coordinator Guide)</b></p>	<p>The DAC/SAC Guide is a 36-page handbook that includes the following information:</p> <ul style="list-style-type: none"> <li>• Key dates</li> <li>• Roles and responsibilities</li> <li>• Test security</li> <li>• Accessibility information</li> <li>• Procedures before testing begins</li> <li>• Technology resources</li> <li>• Testing times and schedules</li> <li>• Braille ordering</li> <li>• Overview of testing and test management software</li> <li>• Procedures for once testing is finished</li> <li>• Transferring students</li> <li>• Coordinator checklists</li> <li>• Guidelines and procedures for documenting a test security incident</li> <li>• Multiplication chart (for use with some tests)</li> <li>• Sample test schedules</li> </ul>
<p><b>eDIRECT User Guide: User Management</b></p>	<p>The Manage Users Guide is a 32-page guide that includes the following information:</p> <ul style="list-style-type: none"> <li>• Managing user’s own eDIRECT account</li> <li>• Adding and editing other eDIRECT users</li> <li>• Adding and removing eDIRECT user permissions</li> </ul>
<p><b>eDIRECT User Guide: Students and Testing</b></p>	<p>The Students and Testing Guide is a 72-page guide that includes the following information:</p> <ul style="list-style-type: none"> <li>• Adding and editing students and student demographics, accommodations, and testing codes</li> <li>• Viewing, adding, and editing student test session information</li> <li>• Printing and managing student test tickets</li> <li>• Transferring students between schools and districts</li> </ul>
<p><b>Accessibility Guide</b></p>	<p>The Accessibility Guide is a 22-page document that outlines the various accessibility options available to students taking the Wisconsin Forward Exam. Guidelines for using the various accessibility features were also included.</p>
<p><b>Student Tutorial</b></p>	<p>The Student Tutorial includes 12 video “chapters” intended for students. It is designed to show students the interface of the online testing system and familiarize them with the tools and features available. It is intended to accompany the Online Tools Training (OTT).</p> <p>The 2017 tutorial also includes ten chapters for test coordinators and proctors to familiarize them with the administrative features and functionality of eDIRECT as well as the accessibility features of the Wisconsin Forward Exam.</p>

Table 4-8 Summary Table of Manual Materials (cont.)

Material	Configuration
<p><b>TAM (Test Administration Manual) and Test Directions</b></p>	<p>The TAMs was a 47-page document intended for test proctors. It includes the following information:</p> <ul style="list-style-type: none"> <li>• Key dates</li> <li>• Test times and schedules</li> <li>• Test security</li> <li>• Accessibility information</li> <li>• Procedures for before testing</li> <li>• Test ticket management</li> <li>• Test material management</li> <li>• Setting up the testing environment</li> <li>• Procedures for during testing</li> <li>• Procedures for after testing</li> <li>• Proctor checklist and guidelines</li> <li>• Read-aloud protocol</li> <li>• Scribe guidelines</li> </ul> <p>Test Directions are presented in seven documents, one per grade. Each set of test directions includes a script for test proctors as they guide students through logging in to the INSIGHT test software and through the online test directions screens.</p>
<p><b>Technology User Guide (TUG)</b></p>	<p>The TUG is an approximately 248-page document intended for Technology Coordinators. It includes detailed instructions on the installation and configuration of INSIGHT and the TSM for all supported platforms.</p>
<p><b>Interpretive Guide</b></p>	<p>The Interpretive Guide is a 30-page document that includes the following information:</p> <ul style="list-style-type: none"> <li>• Interpreting Wisconsin Forward Exam scores</li> <li>• Accessing Individual Student Reports (ISRs) and summary reports via the eDIRECT Portal.</li> </ul>
<p><b>Technology Readiness Package</b></p>	<p>The Technology Readiness Package is a suite of documents and tools for Technology Coordinators to prepare for the Wisconsin Forward Exams that includes the following:</p> <ul style="list-style-type: none"> <li>• Capacity Estimator</li> <li>• System requirements</li> <li>• Technology overview presentation</li> <li>• Technology Coordinator Checklist</li> <li>• Tech FAQ</li> </ul>
<p><b>Online Tools Training (OTT)</b></p>	<p>The OTT is a hands-on opportunity for students to become familiar with logging in, navigating, using tools, using accessibility features, reviewing, and submitting the test prior to signing in to an actual test. It is designed to be a second step after viewing the student tutorials.</p>

Table 4-8 Summary Table of Manual Materials (cont.)

Material	Configuration
<p><b>Technical Report</b></p>	<p>The Technical Report is a manual that covers all grades and all psychometric details associated with administering the Wisconsin Forward Exam. The Technical Report provided by DRC presents thorough documentation to demonstrate the assessment validity. The document contains the following information:</p> <ul style="list-style-type: none"> <li>• Description of the item pool used in the Wisconsin form-development process</li> <li>• Description of the test administration process and test security</li> <li>• Scoring of various types of items</li> <li>• Summary information of student performance (including means and standard deviations of scaled scores, percentage of examinees within each performance level for each content area and grade level, and scale score distribution tables)</li> <li>• Item- and test-level analysis information for each content area and grade level, test scaling procedure, and student scoring process</li> <li>• Measures of scoring reliability for text-dependent analysis items</li> <li>• Evidence of test validity</li> </ul>
<p><b>Data Forensic Report</b></p>	<p>A separate Data Forensic Report will include analyses of the following:</p> <ul style="list-style-type: none"> <li>• Evaluation of response changes</li> <li>• Evaluation of student response time to items</li> </ul>

## Part 5: Scoring

---

The purpose of Part 5 is to demonstrate adherence to the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) (2014) Standards 4.18, 4.20, 6.8, and 6.9. Standard 4.18 provides some general guidance for Part 5:

Procedures for scoring and, if relevant, scoring criteria, should be presented by the test developer with sufficient detail and clarity to maximize the accuracy of scoring. Instructions for using rating scales or for deriving scores obtained by coding, scaling, or classifying constructed responses should be clear. This is especially critical for extended-response items such as performance tasks, portfolios, and essays. (91)

Part 5 describes

- the scoring process of multiple-choice (MC) and multi-select (MS) items;
- the auto-scoring process of technology-enhanced (TE), short-answer (SA), and evidence-based selected response (ESR) items; and
- the scoring of text-dependent analysis (TDA) items, including
  - scoring rubrics,
  - Artificial Intelligence (AI) scoring process,
  - handscoring process,
  - electronic handscoring system,
  - scoring personnel selection,
  - anchor papers selection, and
  - TDA item scores distribution.

### 5.1 Multiple-Choice and Multi-Select Item Scoring Process

Responses to MC and multi-select items were captured during the online test administration. In the case of the Braille or paper-and-pencil form administrations, student responses to these items were transcribed into the online system by a test administrator. All MC and multi-select items had one and only one correct item response for each item.

### 5.2 Technology-Enhanced, Short-Answer, and Evidence-Based Selected Response Item Scoring Process

All TE, SA, and ESR items were processed through DRC's autoscoring engine and scored according to the assigned scoring rules. DRC ensured that all rubrics and scoring rules were verified for accuracy before scoring any of these items. DRC established an adjudication process for these items and any gridded responses to verify that correct answers were identified. The quality process for DRC's TE, SA, or ESR item scoring included the following:

- A scoring rubric was created for each TE, SA, or ESR item. It was similar to describing the one-and-only correct answer for dichotomously scored items (scored as either right or wrong). For ELA ESR items worth 2 points, the rubric described in detail the type of response that could receive partial credit for 1 score point.
- The information from the scoring rubric was entered into the scoring system within the item banking system so that all information about the item resided in one place, along with the item image and other metadata. This scoring information designated specific information that varied by item type. For example, for a drag-and-drop item, the information included which objects are to be placed into which drop region to receive credit.
- The information was then verified by another autoscoring expert.
- After testing started, reports were generated that showed every response, how many students gave that response, and the score the scoring system provided.
- The scoring was then checked against the scoring rubric using two levels of verification.
- If any discrepancies were found, the scoring information was modified and verified again. Scoring was then rerun. This checking and modification process continued until no other issues were found.
- As a final check, a final report was run that showed all student responses, along with frequencies and received scores.

In the case of the Braille or paper-and-pencil form administrations, student responses to paper-and-pencil TE, SA, or ESR or TE-equivalent items were transcribed (entered) into the online system by a test administrator.

### **5.3 Scoring of Text-Dependent Analysis Items**

Sections 5.3 and 5.4 document the scoring processes used for TDA items. This documentation forms part of the validity evidence supporting the scoring process used for these items. Sections 5.3 and 5.4 describe the scoring rubrics, the scoring process, the selection of sample (anchor) papers used to train scoring personnel, the process of selecting personnel, and the distributions of scores for TDA items.

#### **5.3.1 Description of Scoring Rubrics and Non-Score Codes**

In the 2017 administration, the ELA forms in grades 3–8 contained one TDA item at each grade level. The TDA item responses were scored using a 4-point holistic rubric. The responses were scored using an AI engine, and then validation scoring was performed by human scorers on approximately 10 percent of the AI scored responses. Table 5-1 presents the scoring rubric. In cases where student responses could not be scored, a non-score code was used. The non-score codes are presented in Table 5-2. All non-score codes were converted to a score of “0” in derivation of student total test scores.

### **5.3.2 Artificial Intelligence Scoring**

DRC partnered with Measurement Incorporated (MI) to score the TDA tasks. MI is a recognized leader in the field of automated essay scoring. MI employed its essay scoring engine (PEG) to score all student responses. The AI model for scoring the Wisconsin student responses was built by first having DRC expert scorers score a representative sample of Wisconsin responses twice, independently. Once the sample was scored, responses and corresponding scores were delivered to the AI team at MI for model development. MI's linguistics, software developers, psychometricians, and human-computer interaction specialists created task-specific algorithms that were then used to accurately predict how humans would score these responses.

MI's AI scoring software flagged a small percentage of student responses that could not be AI scored. The software has various triggers for identifying alert responses and responses in which it has low confidence. These responses lack proper development, lack enough content to be scored, are written in an unsupported language, or contain inappropriate language or represent a bad-faith effort to complete the test (e.g., repeated text, off-topic text). The limited number of responses that could not be scored by AI were routed to DRC for human scoring with a condition code indicating why the response could not be AI scored.

### **5.3.3 Handscoring Process**

Human scoring of TDA items is referred to as "handscoring." The scoring personnel who score TDA items are referred to as scorers. The scorers were trained using customized training materials, such as the anchor papers described in Section 5.3.5. Once qualified, scorers were required to maintain accuracy standards throughout the project. These requirements were assessed primarily through each scorer's daily agreement rates with the AI scores (described below) and targeted read-behinds with team leaders (described below). Reports were generated daily and monitored by the scoring director, team leaders, and project manager. Any scorers falling below the established quality standards for any item were retrained with the supervisors, who monitored scoring trends (such as difficulty with any particular score point). These scorers also received additional reviews and read-behinds. Failure to recalibrate resulted in dismissal from the scoring assignment. This process was in place throughout the entire handscoring window.

### **5.3.4 Handscoring System**

Scoreboard, DRC's handscoring system, was used to score TDA items as a validation method and to resolve cases where the AI engine returned a non-scoreable condition code. Scoreboard presented images of rendered online responses to trained scorers who assigned scores for the TDA items. The rendered student responses were viewed on high-quality workstation monitors. Images of each student's responses were automatically routed to designated groups of scorers trained to score these items.

### 5.3.5 Anchor Papers and Training Papers

All training materials, including scoring guides and rubrics, anchor papers, training papers, and qualification papers, were selected from live student work. Prior to actual scoring, a group of papers written by Wisconsin students were selected as models to train scorers. These papers, referred to as anchor papers, played an important role in deciding which level of writing should receive which score. The range-finding committee, made up of six scoring directors (one from each grade), then chose those papers that had a high level of agreement to create a set of anchor papers and a set of training papers for each grade. These anchor and training papers were then used to train a select group of scorers who scored approximately 2,000 student responses used to train the AI engine (model building). For this model-building activity, each student response was independently scored by two separate scorers. If there was any disagreement between the two readers, the scores were adjudicated to 100 percent agreement. Once trained, the AI engine scored the remaining Wisconsin student responses. Upon completion of the AI scoring, a random sample consisting of approximately 10 percent of the student responses scored by the AI engine was sent to DRC for a human read. DRC then scored the 10 percent read-behind sample using the original AI engine scoring group to ensure consistency. The 10 percent read-behind with human scorers served as a validation check of the AI engine scoring data.

### 5.3.6 Scoring Personnel and Qualifications

AERA, APA, & NCME (2014) Standard 4.20 specifies the following:

The process for selecting, training, qualifying, and monitoring scorers should be specified by the test developer. The training materials, such as the scoring rubrics and examples of test takers' responses that illustrate the levels on the rubric score scale, and the procedures for training scorers should result in a degree of accuracy and agreement among scorers that allows the scores to be interpreted as originally intended by the test developer. Specifications should also describe processes for assessing scorer consistency and potential drift over time in raters' scoring. (92)

DRC recruited, trained, and managed personnel to complete all of the handscoring operations within the timelines of the contract. The recruitment process and requirements of the scorers, team leaders, and scoring supervisors are described in the following sections.

**Scorers**—The DRC scorer pool included many retired and current educators, as well as engineers, editors, published authors, and individuals with advanced degrees. The minimum qualification for all scorers was a bachelor's degree. Scorers were required to participate in training and successfully pass a qualification round. Once qualified, scorers could start scoring, but throughout the scoring process, scorer performance was assessed by a scoring director, a team leader, and the project manager through read-behinds and reviews of inter-rater reliability statistics, as described in Sections 5.3.8, 5.4, and Part 9.

**Team Leaders**—Team leaders were selected on the basis of their ability to maintain a high degree of scoring accuracy and consistency, often across multiple content areas and grades. Team leaders were also required to possess good interpersonal and leadership skills in order to

be effective when training and counseling scorers. Team leaders were each responsible for a small team of scorers. In addition to performing read-behinds on scorers, team leaders also coached scorers when needs were identified through data review or otherwise by supervisory staff.

**Scoring Directors**—Scoring directors comprised the core group at DRC who directed and organized the scoring process, and trained team leaders and scorers. Scoring directors had extensive experience as team leaders prior to their qualification and selection, and most had previous scoring director experience. Scoring directors were content area experts. They oversaw all team leaders and scorers.

### **5.3.7 Scorer Training**

AERA, APA, & NCME (2014) Standard 6.9 specifies the following:

Those responsible for test scoring should establish and document quality control processes and criteria. Adequate training should be provided. The quality of scoring should be monitored and documented. Any systematic source of scoring errors should be documented and corrected. (118)

Qualification was a critical task in the training process and the final determinant of scorer readiness. All scorers, including team leaders, were required to achieve a certain level of scoring accuracy in the qualifying round that followed training. The standard to which they were held was industry standard for TDA items: at least 70% exact agreement. Only those who were successfully validated were qualified as scorers to score tests.

### **5.3.8 Monitoring the Scoring Process**

AERA, APA, & NCME (2014) Standard 6.8 states the following:

Those responsible for test scoring should establish scoring protocols. Test scoring that involves human judgment should include rubrics, procedures, and criteria for scoring. When scoring of complex responses is done by computer, the accuracy of the algorithm and processes should be documented. (118)

The read-behind was used as a valuable monitoring technique. Each team leader was able to read a random selection of a scorer's scored responses. This reading could be targeted at the item and score-point level. The scores (the scorer score and the team leader score) were compared, and if they agreed, the team leader was able to offer feedback, which enhanced the scorer's confidence and ability to score quickly and accurately. However, if a scorer strayed from the standards established in the training samples, the aberrant scoring was detected, and the team leader was able to offer guidance necessary to refocus the scorer's effort. Read-behinds by team leaders were more frequent for the scorers who had inconsistent scores, thus correcting any scoring variations.



### 5.3.9 Final Scores

All TDA responses were sent to the AI engine for scoring. The AI scores were the final scores (i.e., scores of record). In all cases where the AI engine returned a non-scorable condition code, the student responses were reviewed and scored by humans and a resolution was reached. If a human scorer was able to assign a score for a response that the AI engine was not able to score, then a score from a human scorer became the score of record.

## 5.4 Inter-Rater Reliability

A random 10 percent of the AI-scored responses were sent to human scorers for the second reads and used to validate (assess the accuracy of) the AI score. The statistics for the inter-rater reliability were calculated for all TDA items. To determine the reliability of scoring, the score distribution and percentage of agreement of the two readers were examined. In this section, the distribution of TDA item scores is presented. Additional inter-rater reliability measures including intra-class correlation and weighted kappa statistics are presented in Part 9 of the Technical Report.

### 5.4.1 Distribution of TDA Item Scores

Table 5-3 shows the score and non-scorable code distributions for TDA items for all Wisconsin students with valid ELA scores. The presented scores are from the AI engine supplemented by non-scorable responses resolved by human readers. It should be noted that a large number of records displayed condition code “N” (insufficient to score). Such an outcome may be influenced by the fact that the TDA item type is relatively new to Wisconsin students and many students might not have been familiar with that item type. It is expected that the number of students in this code category will decrease over time.

Table 5-4 shows the score and non-scorable code distribution for TDA items for responses selected for the second read (handscoring). Table 5-5 shows the associated percentage of scores and non-scorable code for TDA items for responses selected for the second read. In both tables, Scorer 1 is the AI engine and Scorer 2 is a human scorer. It should be noted that all non-scorable responses, returned by the AI engine, were reviewed by the scoring directors and assigned either a specific condition code or a score. The data in Tables 5-4 and 5-5 (Non-Scorable Code columns) show the number and percentage of the non-scorable responses from AI engine and detailed condition codes for these responses assigned by the human scorers (scoring directors).

As shown in Tables 5-4 and 5-5, there was a generally high degree of agreement between the AI engine and the human scorer for all grades with the following exceptions: 52.10% of students received a score of 1 from the AI engine compared to 46.18% of students who received a score of 1 from the human scorer (about 6% difference) and 11.72% of students received a score of 2 from the AI engine compared to 16.20% of students who received a score of 2 from the human scorer (over 4% difference) in grade 5; 34.82% of students received a score of 1 from the AI engine compared to 38.36% of students who received a score of 1 from the human scorer

(over 3% difference) in grade 6; 31.89% of students received a score of 1 from the AI engine compared to 36.80% of students who received a score of 1 from the human scorer (about 5% difference), and 9.48% of students received a score of 3 from the AI engine compared to 5.86% of students who received a score of 3 from the human scorer (over 3% difference) in grade 7. All other differences between the AI engine and the human scorer were 2 percent or less.

## **5.5 Summary**

Taken together, the information presented in this part of the Technical Report summarizes the scoring procedures for different types of items and the steps taken by DRC to ensure accuracy in the TE item scoring, AI scoring, and handscoring processes. The score distribution statistics from the AI engine and the human scorer presented in Section 5.4 demonstrate that the items are scored reliably during the scoring process. These efforts by DRC follow multiple best practices of the testing industry and support AERA, APA, & NCME (2014) Standards 4.18, 4.20, 6.8, and 6.9 as presented in Part 5.

Table 5-1 TDA Item Scoring Guidelines, Grades 3–8

Score Value	Score Description	Scoring Rubrics
4	Demonstrates effective analysis of text and skillful writing	<ul style="list-style-type: none"> <li>• Effective addressing of all parts of the task to demonstrate an in-depth understanding of the text(s)</li> <li>• Thorough analysis based on explicit and implicit meanings from the text(s) to support claims, opinions, and ideas</li> <li>• Strong organizational structure and focus on the task with logically grouped and related ideas, including an effective introduction, development, and conclusion</li> <li>• Substantial, accurate, and direct reference to the text(s) using an effective combination of details, examples, quotes, and/or facts</li> <li>• Substantial reference to the main ideas and relevant key details of the text(s)</li> <li>• Skillful use of transitions to link ideas within categories of textual and supporting information</li> <li>• Effective use of precise language and domain-specific vocabulary drawn from the text(s)</li> <li>• Few errors, if any, in sentence formation, grammar, usage, spelling, capitalization, and punctuation that do not interfere with meaning</li> </ul>
3	Demonstrates adequate analysis of text and appropriate writing	<ul style="list-style-type: none"> <li>• Adequate addressing of all parts of the task to demonstrate a sufficient understanding of the text(s)</li> <li>• Clear analysis based on explicit and implicit meanings from the text(s) to support claims, opinions, and ideas</li> <li>• Appropriate organizational structure and focus on the task with logically grouped and related ideas, including a clear introduction, development, and conclusion</li> <li>• Sufficient, accurate, and direct reference to the text(s) using an appropriate combination of details, examples, quotes, and/or facts</li> <li>• Sufficient reference to the main ideas and relevant key details of the text(s)</li> <li>• Appropriate use of transitions to link ideas within categories of textual and supporting information</li> <li>• Appropriate use of precise language and domain-specific vocabulary drawn from the text(s)</li> <li>• Some errors in sentence formation, grammar, usage, spelling, capitalization, and punctuation that seldom interfere with meaning</li> </ul>

Table 5-1 TDA Item Scoring Guidelines, Grades 3–8 (cont.)

Score Value	Score Description	Scoring Rubrics
2	Demonstrates limited analysis of text and inconsistent writing	<ul style="list-style-type: none"> <li>• Inconsistent addressing of some parts of the task to demonstrate a partial understanding of the text(s)</li> <li>• Inconsistent analysis based on explicit and/or implicit meanings from the text(s) that ineffectively supports claims, opinions, and ideas</li> <li>• Weak organizational structure and focus on the task with ineffectively grouped ideas, including a weak introduction, development, and/or conclusion</li> <li>• Limited and/or vague reference to the text(s) using some details, examples, quotes, and/or facts</li> <li>• Limited reference to the main ideas and relevant details of the text(s)</li> <li>• Limited use of transitions to link ideas within categories of textual and supporting information</li> <li>• Inconsistent use of precise language and domain-specific vocabulary drawn from the text(s)</li> <li>• Errors in sentence formation, grammar, usage, spelling, capitalization, and punctuation that may interfere with meaning</li> </ul>
1	Demonstrates minimal analysis of text and inadequate writing	<ul style="list-style-type: none"> <li>• Minimal addressing of part(s) of the task to demonstrate an inadequate understanding of the text(s)</li> <li>• Minimal analysis based on the text(s) that may or may not support claims, opinions, and ideas</li> <li>• Minimal evidence of an organizational structure and focus on the task with arbitrarily grouped ideas that may or may not include an introduction, development, and/or conclusion</li> <li>• Insufficient reference to the text(s) using few details, examples, quotes, and/or facts</li> <li>• Minimal reference to the main ideas and relevant details of the text(s)</li> <li>• Few, if any, transitions to link ideas</li> <li>• Little or no use of precise language or domain-specific vocabulary drawn from the text(s)</li> <li>• Many errors may in sentence formation, grammar, usage, spelling, capitalization, and punctuation that often interfere with meaning</li> </ul>

Table 5-2 TDA Item Non-scorable Codes, Grades 3–8

Non-scorable Code	Definition/Example/Notes
B – Blank	<p>A response that is completely blank. This includes responses that</p> <ul style="list-style-type: none"> <li>• are completely erased (so that words are unreadable).</li> <li>• are completely crossed out (so that words are unreadable).</li> <li>• are online and consist solely of “white space” (e.g., spaces, tabs, returns).</li> </ul>
R – Refusal	<p>A response indicates a refusal to attempt the task. This includes the following examples:</p> <ul style="list-style-type: none"> <li>• <i>“I don’t care”; “I’m not taking this test”; “This is stupid”; “I won’t do it”; “you can’t make me answer this question”</i></li> <li>• <i>“I don’t know”; “IDK”; “we never learned this”; “X”; “NA”</i></li> <li>• <i>Unrelated song lyrics/rap lyrics/poetry (e.g., the lyrics to “Hotel California” in answer to a writing prompt asking whether backpacks should be allowed in class)</i></li> <li>• <i>Intentionally off-task response (e.g., a detailed description of what the student ate for breakfast that morning in answer to a question about Mozart’s childhood)</i></li> </ul> <p>This also includes responses that consist solely of scribbles, random keystrokes (“yyyyyyy”, “av:aeoiahvb”; “e, hrrttuuvv”), indecipherable writing/keystrokes (“swensts mengetstets arawnstets”) emoticons, stray marks, doodles, drawings, circles, underlines, a couple of random letters (not a word), or other evidence that no attempt was made to address the task.</p>
N – Non-scorable	<p>This category includes</p> <ul style="list-style-type: none"> <li>• responses written entirely in a language other than English.</li> <li>• responses that are completely illegible due to poor handwriting.*</li> <li>• online or typed responses that are incoherent due to consisting of incomprehensible strings of words that are not clearly a Refusal or Off Topic (e.g., <i>“best day school teacher inspired so I car”</i>)</li> <li>• responses too insufficient to be assessed by the criteria on the rubric.</li> <li>• (for TDAs only) responses that address some part of the question but do not contain any logical/accurate/relevant reference to the passage(s) or any ideas contained in the passage(s).</li> <li>• (for TDAs only) responses that consist solely, or almost solely, of text copied directly from the passage(s).</li> </ul> <p>* If a response is difficult to read, every effort is made to read the response. Multiple people, including a team leader and/or a scoring director, will attempt to decipher the response, and the original answer document will be reviewed if necessary. If, ultimately, only a portion of the response is legible, that verbiage will be scored on its own merits.</p>
T – Off Topic	<p>A response makes no reference to the item or (if applicable) the passage provided but does not seem to constitute an intentional refusal.</p> <p>If any part of the response relates to the item in any way, score the response.</p>
C – Copied Item/Directions	<p>A response consists of text copied from the item and/or test directions.</p>

Note: Crossed out but legible/partially legible responses are scored according to the rubric based on whatever verbiage is legible.

Table 5-3 TDA Item Score Distribution

Grade	Item Number	Total Count	Item Score				Non-Scorable Code				
			1	2	3	4	B	C	N	R	T
3	4	63862	34766	12222	1253	1	187	77	14964	363	29
4	6	64361	33990	15765	4758	38	210	51	8002	485	1062
5	4	62952	40533	15831	1020	3	90	7	5242	197	29
6	4	62698	30459	22076	4920	388	169	11	4281	357	37
7	4	63004	22119	28333	8209	1223	199	7	2502	393	19
8	5	62012	26538	20818	6841	1177	458	13	5514	614	39

Table 5-4 TDA Item Score Distribution: AI Engine vs. Human Scorer

Grade	Scorer	Total Count	Score Count				Non-Scorable Code Count				
			1	2	3	4	B	C	N	R	T
3	Scorer 1 (AI Engine)	25478	8086	1711	202				15479		
3	Scorer 2 (Human)	25478	8489	1348	151	11	1	76	15005	368	29
4	Scorer 1 (AI Engine)	18760	6389	2169	568	2			9632		
4	Scorer 2 (Human)	18760	6739	1984	359	46	1	50	8020	496	1065
5	Scorer 1 (AI Engine)	15506	8079	1818	112	2			5495		
5	Scorer 2 (Human)	15506	7161	2512	292	46		7	5255	204	29
6	Scorer 1 (AI Engine)	11588	4035	2316	483	41			4713		
6	Scorer 2 (Human)	11588	4445	2127	280	23		10	4302	364	37
7	Scorer 1 (AI Engine)	10422	3324	3056	988	116			2938		
7	Scorer 2 (Human)	10422	3835	2994	611	44		6	2513	400	19
8	Scorer 1 (AI Engine)	15070	5574	2386	757	134			6219		
8	Scorer 2 (Human)	15070	5881	2370	540	60		10	5547	622	40

Table 5-5 TDA Item Percentage Score Distribution: AI Engine vs. Human Scorer

Grade	Scorer	Total Count	Score Percentage				Non-Scorable Code Percentage				
			1	2	3	4	B	C	N	R	T
3	Scorer 1 (AI Engine)	25478	31.74	6.72	0.79				60.75		
	Scorer 2 (Human)	25478	33.32	5.29	0.59	0.04	0.00	0.30	58.89	1.44	0.11
4	Scorer 1 (AI Engine)	18760	34.06	11.56	3.03	0.01			51.34		
	Scorer 2 (Human)	18760	35.92	10.58	1.91	0.25	0.01	0.27	42.75	2.64	5.68
5	Scorer 1 (AI Engine)	15506	52.10	11.72	0.72	0.01			35.44		
	Scorer 2 (Human)	15506	46.18	16.20	1.88	0.30		0.05	33.89	1.32	0.19
6	Scorer 1 (AI Engine)	11588	34.82	19.99	4.17	0.35			40.67		
	Scorer 2 (Human)	11588	38.36	18.36	2.42	0.20		0.09	37.12	3.14	0.32
7	Scorer 1 (AI Engine)	10422	31.89	29.32	9.48	1.11			28.19		
	Scorer 2 (Human)	10422	36.80	28.73	5.86	0.42		0.06	24.11	3.84	0.18
8	Scorer 1 (AI Engine)	15070	36.99	15.83	5.02	0.89			41.27		
	Scorer 2 (Human)	15070	39.02	15.73	3.58	0.40		0.07	36.81	4.13	0.27

## **Part 6: Calibration, Equating, and Deriving Scale Scores**

---

This part of the Technical Report describes the analyses involving test calibrating, equating, and student scoring that occurred for the Wisconsin Forward Exam after the 2017 test administration. Part 6 demonstrates adherence in the Wisconsin Forward Exam program data analysis to AERA, APA, & NCME (2014) Standards 1.8, 2.13, 5.2, and 7.2. Each standard will be explicated within the appropriate section of this chapter. Standard 7.2 provides general guidance that is relevant to this chapter:

The population for whom a test is intended and specifications for the test should be documented. If normative data are provided, the procedures used to gather the data should be explained; the norming population should be described in terms of relevant demographic variables; and the year(s) in which the data were collected should be reported. (126)

Student responses on the Wisconsin Forward Exam are inputted into complex mathematical algorithms designed to model the relationship between a student’s ability in a content area and a test item. The group of algorithms is collectively known as item response theory (IRT). Wisconsin Forward Exam scores are established through the processes of calibration, scaling, and item-pattern scoring.

Calibration is the mathematical process of estimating characteristics of individual items. These characteristics are termed “item parameters.” Section 6.1 serves to explain this process, beginning with a description of the calibration methods that were applied to the Spring 2017 Wisconsin Forward Exam, followed by a presentation of a calibration sample and a discussion of the calibration models and the software used. The results of the calibration process, using model-to-data fit statistics, and the outcomes of test scaling are also discussed in Section 6.1. Section 6.2 describes test equating procedures and results. Section 6.3 addresses the process for derivation of scale scores from raw scores.

Readers should note that calibration, equating, and scoring using IRT are mathematically complex and computationally intensive processes. A full understanding of these topics requires a background in psychometrics. However, in order to make these processes more accessible and transparent to a wider range of audiences, a brief, nontechnical explanation of how scale scores are derived from raw scores is provided in Section 6.3. Additional references are also provided.

### **6.1 Item Calibration**

This section of the report outlines the calibration procedures and results for the Spring 2017 Wisconsin Forward Exam.

#### **6.1.1 Calibration Models**

The three-parameter logistic (3PL) model and the two-parameter partial credit (2PPC) IRT model (Bock & Aitkin, 1981; Thissen, 1982) were used to estimate parameters for MC and



CR items, respectively. All non-MC items, including TE, ESR, MS, SA, and TDA items, were treated as CR items in calibrations. Item parameters for items contained in all Wisconsin assessments were estimated using a marginal maximum-likelihood procedure.

Under the 3PL model, the probability that a student with a trait or scale score  $\theta$  will respond correctly to MC item  $j$  is

$$P_j(\theta) = c_j + (1 - c_j) / [1 + \exp(-1.7a_j(\theta - b_j))].$$

In the equation,  $a_j$  is the item discrimination,  $b_j$  is the item difficulty, and  $c_j$  is the probability of a correct response by a very low-ability student. Under the 2PPC model, the probability that a student with a trait or scale score  $\theta$  will respond in category  $k$  to partial-credit item  $j$  is

$$P_{jk}(\theta) = \exp(z_{jk}) / \sum_{i=1}^{m_j} \exp(z_{ji}),$$

$$\text{where } z_{jk} = (k - 1)f_j - \sum_{i=0}^{k-1} g_{ji}, \text{ and } g_{j0} = \mathbf{0} \text{ for all } j.$$

The summary output of the 3PL and 2PPC models is in two different metrics. The discrimination and location parameters for the MC items are in the traditional 3PL metric and are labeled  $a$  and  $b$ , respectively. In the 2PPC model,  $f$  (alpha) and  $g$  (gamma) are analogous to  $a$  and  $b$ , 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 parameters  $a$  and  $b$  are not directly comparable to the 2PPC parameters  $g$  and  $f$ ; however, they can be converted to a common metric. The two metrics are related by  $a = f / 1.7$  and  $b = g/f$  (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.

Using the 3PL/2PPC model for estimation of ELA, Mathematics, and Science grade 4 item parameters and the 3PL model for estimation of Science grade 8 and Social Studies item parameters was consistent with the past methodology (except for administration year 2014–15 for ELA and Mathematics) implemented for these content areas in the Wisconsin testing program. Item parameters estimated after the 2016–17 test administration were used to score Wisconsin students who took these tests.

### 6.1.2 Calibration Sample

The calibration of the Wisconsin Forward Exam occurred after the Spring 2017 test administration and was based on student data from an early return sample of the state test data. This arrangement was chosen in order to expedite the data analysis in preparation for reporting.

This section provides information on the comparability of the calibration sample to the census data in terms of demographic characteristics in adherence to Standard 1.8 of the AERA, APA, & NCME (2014) *Standards*:

The composition of any sample of test takers from which validity evidence is obtained should be described in as much detail as is practical and permissible, including major relevant socio-demographic and developmental characteristics. (25)

The calibration sample consisted of the student data acquired before the testing window ended and included students from public, choice, and private schools. The characteristics of the calibration sample compared to the total population of students are presented in Tables 6-1 through 6-4 for ELA, Mathematics, Science, and Social Studies, respectively. The 2017 calibration sample was comparable to the Wisconsin student population.

### 6.1.3 Calibration Procedure

The calibrations were conducted separately for each grade level and content area using the marginal maximum-likelihood procedures implemented with the expected maximum algorithm (Bock & Aitkin, 1981; Thissen, 1982). In a process of item calibration, the number of estimation cycles was set to 99 with the convergence criterion of 0.001 for all content areas. The maximum value of  $a$ -parameter was set to 5.0, and the range for  $b$ -parameter was set between -7.5 and 7.5. For all items, the estimated  $a$ - and  $b$ -parameters were within the prescribed parameter ranges. The  $c$ -parameters for anchor items were fixed to their Spring 2016 values. It should be noted that there was a small number of items with the default value for the  $c$ -parameter on all tests. When the PARDUX (Burket, 2002) program, used to calibrate the items, encounters difficulty estimating the  $c$ -parameter, it assigns a default  $c$ -parameter value of 0.20.

### 6.1.4 Calibration Software

Calibrating of the Wisconsin Forward Exam data was performed using PARDUX software (Burket, 2002). PARDUX is designed to produce a single scale by jointly analyzing data resulting from students' responses to both MC items and CR items for assessments that include both item types. In PARDUX, items are calibrated based on IRT, using the 3PL model (Lord & Novick, 1968) for MC items and the 2PPC model (Yen, 1993) for CR items.

PARSCALE, MULTILOG, and BIGSTEPS are among the most widely known and used IRT programs. Extensive simulation studies and comparisons between PARDUX and MULTILOG (Thissen, 1990)—a program widely used for research purposes—have shown that PARDUX provides precise parameter and ability estimates and performs more efficiently than MULTILOG (Fitzpatrick, 1991). Simulation studies have also compared PARDUX with PARSCALE (Muraki & Bock, 1991) and with BIGSTEPS (Wright & Linacre, 1992). Fitzpatrick and Julian (1996) found that PARDUX provided precise parameter and ability estimates and performed more efficiently than the other programs. Extensive research with simulation data has also shown that the IRT procedures used here produce accurate vertical scaling (Yen & Burket, 1997).

### 6.1.5 Calibration Results

This section describes the calibration results in terms of the estimation of item parameters and model-to-data fit for all content areas and grades.

#### IRT Item Parameters

When calibrating items, items may not converge, meaning the characteristics of the item are not able to be determined. When this occurs, items may be suppressed from student scoring and future assessments. In Spring 2017, no convergence issues occurred for any item on the operational tests.

#### IRT Item Fit

The calibration process produces ability and item parameter estimates that can be used to predict student response patterns to each item. For example, based on the item parameter estimates for item difficulty and item discrimination, we may expect that low-ability students are less likely to answer a difficult and highly discriminating item correctly than higher-ability students. After parameters are produced, we can compare the predicted scoring patterns to the observed scoring patterns in what are referred to as item-to-model fit comparisons. Where there is little difference between the predicted scoring patterns and the observed scoring patterns, the model can be said to “fit” the data.

A procedure developed by Yen (1981) was used to assess model-to-data fit for all test items. In this procedure, students are rank ordered on the basis of their  $\hat{\theta}$  values and sorted into ten cells, with 10 percent of the sample in each cell. Each item  $j$  in each decile  $i$  has a response from  $N_{ij}$  examinees. The fitted IRT models are used to calculate an expected proportion  $E_{ijk}$  of examinees who respond to item  $j$  in category  $k$ . The observed proportion  $O_{ijk}$  is also tabulated for each decile. The fit index for item  $i$  is

$$Q_{1j} = \sum_{i=1}^{10} \sum_{k=1}^{m_j} \frac{N_{ij} (O_{ijk} - E_{ijk})^2}{E_{ijk}}.$$

$Q_{1j}$  should be approximately chi-square distributed with degrees of freedom ( $DF$ ) equal to the number of “independent” cells,  $10(m_j - 1)$ , minus the number of estimated parameters. For the 3PL model,  $m_j = 2$ , so  $DF = 10(2 - 1) - 3 = 7$ . For the 2PPC model,

$$DF = 10(m_j - 1) - m_j = 9m_j - 10.$$

DRC evaluated item-to-model fit in a two-step process. First, item-to-model fit information was obtained for each item using a  $Z$ -statistic. The  $Z$ -statistic is an index of the degree to which obtained proportions of students with each item score match the proportions predicted by the estimated student ability and item parameters. When the difference between the obtained proportions of students with each item score and the proportions predicted by the

estimated student ability and item parameters reached a certain threshold, the item was flagged for “misfit.”

The Z-statistic is a transformation of the chi-square ( $Q_1$ ) statistic that takes into account differing numbers of score levels as well as sample size using the equation

$$Z_j = \frac{(Q_{1j} - DF_j)}{\sqrt{2DF_j}},$$

where  $Q_{1j}$  is the item chi-square statistic,  $j$  is an item, and  $DF$  is the degrees of freedom for a given item  $j$ .

Because the value of  $Z$  increases as the sample size increases, with other things being equal, the critical values for  $Z$  were established using the following equation (Yen & Candell, 1991)

$$Z_{crit,j} = \frac{4N_j}{1500},$$

where  $Z_{crit,j}$  is the critical value of  $Z$  for item  $j$  and  $N_j$  is the number of students who responded to item  $j$ . These values, along with the associated chi-squares ( $Q_1$ ), are computed for ten intervals corresponding to deciles of the ability distribution (Yen, 1984).

Table 6-5 presents items that were flagged for less than optimal fit when the obtained Z-statistic exceeded the critical Z-statistic value. This table specifies the content area, grade level, item number in the calibration, item type (MC or CR),  $N$  size (the number of students who took this item),  $Z$ , and critical  $Z$ , as described previously. Eight items were flagged for poor fit for ELA, three items were flagged for Mathematics, and one item was flagged for Social Studies. Most of the flagged items were CR items (TE and ESR). For example, ELA grade 4 item #36 in calibration was flagged because the observed  $Z$  of 242.75 is larger than the critical  $Z$  value of 171.33 based on a sample size of 64,248. While for many of the flagged items the observed  $Z$  and the critical  $Z$  are not very far apart, indicating small misfit, it was observed that for some items the misfit was moderate (for example, item #9 in ELA grade 7 or item #2 in ELA grade 8). No items were flagged for poor fit for Science tests.

In order to evaluate item-to-model fit further, DRC inspected the observed-to-predicted item characteristic curve (ICC) for each flagged item. These ICCs simultaneously plot the characteristics of an item (e.g., item difficulty, item discrimination, level of guessing) using IRT model predications and the observed student responses. The ICCs show exactly where along the ability continuum the misfit occurs and the extent of the misfit.

All three cases of MC items flagged for misfit had empirical (observed) information that differed from the model in the lower-ability range, where there are fewer students to provide information at the tail of the distribution. Similarly, for CR items, there were, in general, fewer

students at the lower score levels, which provides less information at the tails of the student distribution. Items that only show misfit at the tails of the distribution provide stable information about the majority of the students—those in the middle range of the distribution. However, if the misfit happens around the middle of the ability range, where there are many students, this may be a concern and may lead to the item being dropped from the item pool.

In a large-scale assessment, such as the Wisconsin Forward Exam, with 17 combinations of grades and content areas, it is expected that some items will be flagged for misfit. As noted, the difference between the obtained Z-statistic and the critical Z-statistic was often small or moderate. Items flagged for misfit were reported to the DRC Test Development team for additional review. Such items are flagged in the Wisconsin Forward Exam item bank and are avoided during the form selection process unless there is a compelling reason that they should be included, such as meeting the test blueprint.

## 6.2 Test Equating

Test equating is a statistical process of placing scores from two or more parallel assessments onto a common scale resulting in direct comparability of scores from two different test forms. A common-item design was used to link the 2017 year's assessments to the established Wisconsin Forward Exam ELA, Mathematics, Science, and Social Studies scales. Sets of items administered to Wisconsin students in Spring 2016, and included in the Spring 2017 assessments, served as the anchor sets in each grade and content area. The anchor sets constituted at least 25% of the Spring 2017 assessments and were representative of the Spring 2017 test content. After the item calibration, item parameters were linked to the Wisconsin Forward Exam scales using the Stocking & Lord (1983) equating procedure.

The Stocking & Lord procedure minimizes the mean squared difference between the two test characteristics curves (TCCs), one based on estimates from the previous calibration and the other on transformed estimates from the current calibration. Let  $\hat{\Psi}_j$  be the TCC based on estimates from a previous calibration and  $\hat{\Psi}_j^*$  be the TCC based on transformed estimates from the current calibration.

$$\hat{\Psi}_j = \hat{\Psi}(\theta_j) = \sum_{i=1}^n P_i(\theta_j; a_i, b_i, c_i)$$

$$\hat{\Psi}_j^* = \hat{\Psi}(\theta_j) = \sum_{i=1}^n P_i\left(\theta_j; \frac{a_i}{A}, Ab_i + B, c_i\right)$$

The TCC method determines the equating constants ( $A$  and  $B$ ) by minimizing the following quadratic loss function ( $F$ ):

$$F = \frac{1}{N} \sum_{a=1}^N (\hat{\Psi}_j - \hat{\Psi}_j^*)^2 .$$

The Stocking & Lord equating procedure is commonly used in large-scale assessments. The standard error of the equating (SEE) is difficult and cumbersome to estimate for IRT equating

procedures like the Stocking & Lord procedure (Kolen & Brennan, 1995; Michaelides & Haertel, 2004). The estimation of the SEE is beyond the scope of this report.

### **6.2.1 Evaluation of Anchor Items**

AERA, APA, & NCME (2014) Standard 5.15 requires information about the anchors, stating the following:

In equating studies that employ an anchor test design, the characteristics of the anchor test and its similarity to the forms being equated should be presented, including both content specifications and empirically determined relationships among test scores. If anchor items are used in the equating study, the representativeness and psychometric characteristics of the anchor items should be presented. (105)

Two statistical methods were used to evaluate anchor items: (1) iterative linking (Candell & Drasgow, 1988) using Stocking & Lord's (1983) test characteristic curve method and (2) differences between the item-ability regression curves.

#### **Test Characteristic Curve Method**

The Stocking & Lord (1983) procedure, also called the test characteristic curve method for which the mathematical equation was provided in a previous section of this document, minimizes the mean squared difference between the two TCCs, one based on estimates from the previous calibration and the other on transformed estimates from the current calibration.

Differential item functioning was evaluated by examining previous (input) and transformed (estimated) item parameters. Items with an absolute difference of parameters greater than two times the root mean square deviation were flagged for review. These differences are monitored by plotting input and estimated item parameters.

#### **Item Response Theory Item-Ability Regression Curves**

Differences between the item-ability regression curves of the anchor items in the Spring 2017 Wisconsin Forward Exam administration were also compared to previous calibrations (from Spring 2016). The differences between the item curves were evaluated using the following statistics:

- UnWtd Mean = Average signed difference in estimated probability
- UnWtd Mean Abs = Average absolute (unsigned) difference in estimated probability
- UnWtd RMSD = Root mean squared difference
- Wtd Mean = Weighted average signed difference in estimated probability
- Wtd Mean Abs = Weighted average absolute (unsigned) difference in estimated probability
- Wtd RMSD = Weighted root mean squared difference

Both unweighted and weighted versions of these statistics were calculated. Unweighted differences give equal weight to differences across the ability spectrum. Weighted differences assign weights according to the number of test takers that are impacted, that is, the frequency distribution of estimated student abilities during the calibration.

For the six statistics listed above, differences greater than  $+0.10$  are considered large and differences between  $+0.07$  and  $+0.10$  are considered moderate.

Additionally, the Maximum Absolute difference (Max Abs) will be identified. For Max Abs, large differences are those greater than  $+0.15$  and moderate differences are all differences between  $+0.125$  and  $+0.15$ .

### **6.2.2 Removal of Anchor Items**

One of the key requirements of anchor items in deriving valid and reliable linking results is that the anchor items form a miniature of the test, in terms of content coverage or test blueprint. While dropping an anchor item that is flagged based solely on statistical criteria has its simplicity, this option may change the content coverage and invalidate results. Before an anchor item is dropped from an anchor set, the item characteristics, adequacy of the content coverage, and impact to the size of the anchor set must be evaluated.

An item may be removed from the anchor set only if it adversely affects the quality of scaling, not the desirability of the results. As such, DRC does not consider how the removal of an item affects the overall mean scale score or the impact data (percentage of students in each achievement level) when recommending items for removal.

Items removed from the anchor set are still scored as part of the whole test. DRC recommends that the anchor items be considered for exclusion from the Wisconsin Forward Exam equating sets under the following conditions:

1. An item may be a candidate for removal when it is flagged for large differences on four of the seven statistics (listed in Section 6.2.1) considered when examining the differences between the IRT item-ability regression curves.
2. Removal of the item will only be considered after alternative explanations have been considered that may explain shifts in performance. For example, performance on the anchor item may improve because of a statewide initiative emphasizing instruction on a particular set of skills. In this case, improved performance on the item represents true growth in that area. Removing the anchor item may artificially lower test scores.
3. Removal of the item may not significantly alter the content distribution of the anchor set. The distribution of the anchor items across the content standards must remain within 10% of the Wisconsin Forward Exam test blueprint.
4. The number of remaining items will remain at an acceptable level of anchor set reliability. Operationally, this means the anchor set will still be representative of the total test blueprint and that the anchor set may not be less than 20% of the total test length.

Flagged items are reviewed by DRC test development experts to verify that no changes to item content or format occurred between the administration in which the anchor item was used and the current administration. In addition, for the flagged CR or TE anchor items, verification that no changes to scoring rubrics occurred between the two administrations is performed.

### 6.2.3 Evaluation of Equating Results

Table 6-6 provides equating results for the TCC method for ELA, Mathematics, Science, and Social Studies. This table summarizes the following information for each grade content area: grade level, number of anchors, number of iterations, quadratic loss function (F), correlation between the *a*-parameter input and estimates, correlation between the *b*-parameter input and estimates, number of *a*- and *b*-parameter outliers as indicated by the root mean square deviation method, and equating constants (A and B). Note that two sets of equating results are included for Social Studies grade 4 due to exclusion of one flagged anchor item from equating.

The overall alignment of the anchor TCCs was very good for all grades and content areas. Figures 6-1 through 6-17 show the TCC alignment of the anchor set before and after equating for all grades and content areas. In these figures, the input anchor set TCC (before equating) is indicated in a dashed red line and the new anchor estimate TCC (after equating) is indicated in the solid blue line. The correlations between the *a*-parameter input and estimates and between the *b*-parameter input and estimates were 0.96 or higher for all grades and content areas. One anchor item was flagged as an *a*-parameter outlier in each of the following: ELA grades 5 and 8, Mathematics grades 3, 6, and 8, and Social Studies grade 10. Three anchor items were flagged as *a*-parameter outliers in Mathematics grade 4. No anchor items were flagged as *a*-parameter outliers in any other grades or content areas. One anchor item was flagged as a *b*-parameter outlier in each of the following: ELA grade 8, Mathematics grades 4 through 8, Science grade 4, and Social Studies grades 4 and 10. Two anchor items were flagged as *b*-parameter outliers in ELA grade 7 and Mathematics grade 3. No anchor items were flagged as *b*-parameter outliers in the remaining grades or content areas. Overall the number of anchor items flagged using the TCC method was small.

Table 6-7 presents the item-ability regression statistics for the Social Studies grade 4 anchor item (anchor position 24; Question 7 in Session 2 of the test) flagged using the item-ability regression curve criteria described in an earlier section of this report. This item was flagged using four or more of the statistics used to examine ICC differences using the IRT item-ability regression curve method. Figure 6-18 shows the ICCs before and after equating for the flagged item in Social Studies grade 4. In this figure the dashed red line is the ICC before equating (based on input parameters) and the solid blue line is the ICC after equating (based on new parameter estimates). Examination of statistical properties of the flagged anchor items revealed that students performed less well on this item in Spring 2017 compared to the Spring 2016 test administration. No other anchor items in any other grades or content areas were flagged using the IRT item-ability regression curve method.

The flagged anchor item was reviewed by DRC test development experts who verified that no changes to item content or format occurred between the Spring 2016 and Spring 2017 administrations. Because no plausible explanation was found for differential item performance



between the two administration years, the flagged anchor item was excluded from the equating of the Social Studies grade 4 test. Exclusion of the flagged anchor item from the anchor set did not significantly affect the anchor set content coverage or the equating results for this grade.

#### 6.2.4 Test Scales

The purpose of scaling a test is to enhance its validity by increasing the comparability of test takers' scores. This section explicates the way in which the Wisconsin Forward Exam scales are produced to comply with Standard 5.2 of the AERA, APA, & NCME (2014) *Standards*, which states the following:

The procedures for constructing scales used for reporting scores and the rationale for these procedures should be described clearly. (102)

The Wisconsin Forward Exam scales were established after the Spring 2016 test administration. In this section the results of the test scaling in the second year the Wisconsin Forward Exam was administered are described and evaluated.

Following the test equating, the equated item parameter estimates in the theta metric were transformed into the scale score metric for the purpose of the evaluation of the scale properties. The scale evaluation included

- evaluation of the TCCs,
- evaluation of the standard error (SE) curves, and
- examination of the growth at quartiles.

The scaling constants,  $M1$  and  $M2$  used to transform equated item parameters in the theta metric into the scale score metric are the same as used in the Spring 2016 scale development and are presented in Table 6-8. The transformation formulae used are presented below:

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

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 level (gamma) for category  $m_j$  in scale score metric for CR items,

$a_{\theta}$  is a discrimination parameter in the original theta metric for MC items,

$b_{\theta}$  is a difficulty parameter in the original theta metric for MC items,

$f_{\theta}$  is a discrimination parameter in the original theta metric for CR items,

$g_{\theta}$  is a difficulty level (gamma) for category  $m_j$  in the original theta metric for CR items, and

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

## ELA Scale

*Test Characteristic Curves*—Figure 6-19 shows the TCCs for ELA tests. As shown in Figure 6-19, the ELA TCCs for grades 3, 4, and 5 are ordinal, indicating increasing difficulty of these assessments as the grade level increases. The grade 6 TCC overlaps with the grade 5 TCC at most ability levels, indicating comparable difficulty of ELA grade 5 and grade 6 assessments. The grade 7 TCC crosses the grade 5 and grade 6 TCCs at the upper end of the ability scale, meaning that while the grade 7 test is more difficult for lower middle-ability students than the grade 5 and grade 6 tests, the grade 7 test tends to be easier for high-ability students compared to the grade 5 and grade 6 assessments. The grade 8 TCC is ordinal in relation to the grades 5, 6, and 7 TCCs indicating that the grade 8 ELA assessment is more difficult for grade 8 students at all ability levels compared to lower grades.

It should be noted that while TCC ordinality is a desirable property of a vertical scale, the lack of it does not necessarily affect student scores or grade-to-grade growth interpretation. As demonstrated by the grade 3–8 pattern of scale scores at quartiles (see *Growth at Quartiles* paragraph below), student ability on ELA assessments increases as grade level increases at all grade levels, indicating grade-to-grade growth.

*Standard Error Curves*—The SE curves for ELA presented in Figure 6-20 are generally U-shaped, indicating smaller errors around ability estimates roughly in the middle of the scale score distribution. The SE is expected to be higher at the top and bottom ends of the ability scale, where fewer items measuring very high- and very low-achieving students are found. Overall, the SEs around the scale score were found to be reasonable for ELA assessments (for more details see Section 6.3.1 of this report).

*Growth at Quartiles*—The estimated scale scores for the ELA calibration sample at the 25th, 50th, and 75th percentiles for all grade levels are presented in Figure 6-21. It can be observed that the scale scores increase as the percentile increases within each grade level. Consistent with the properties of a vertical scale, the scale scores also increase at the same percentile across grade levels, indicating growth on the ELA ability scale as students move from one grade to the next.

## Mathematics Scale

*Test Characteristic Curves*—Figure 6-22 shows the TCCs for Mathematics assessments, which are on a vertical scale. As observed in Figure 6-22, the TCCs for Mathematics, with the exception of the grade 5 and grade 6 TCCs, are ordinal, indicating increasing difficulty of the assessment as the grade level increases. The crossing of the grade 5 and grade 6 TCCs at the lower end of the ability scale indicates that the grade 5 assessment may be more difficult for lower-ability students compared to the grade 6 assessment. This TCC pattern is similar to the one observed for Mathematics in the Spring 2016 test administration.

*Standard Error Curves*—The SE curves for Mathematics presented in Figure 6-23 are U-shaped (as expected), indicating smaller errors around ability estimates roughly in the middle of the scale score distribution. The SE is expected to be higher at the top and bottom ends of the

ability scale, where fewer items measuring these students are found. Overall, the SEs around the scale score were found to be reasonable for Mathematics assessments (for more details see Section 6.3.1 of this report).

*Growth at Quartiles*—The estimated scale scores for the calibration sample at the 25th, 50th, and 75th percentiles for all grade levels are presented in Figure 6-24. It can be observed that the scale scores increase as the percentile increases within each grade level. Consistent with the properties of a vertical scale, the scale scores also increase at the same percentile across grade levels, indicating growth on the Mathematics ability scale as students move from one grade to the next.

## Science Scale

*Test Characteristic Curves*—Although the Science assessments are not vertically scaled, the TCCs for grades 4 and 8 are presented together in Figure 6-25 for comparison purposes. The TCCs are S-shaped, indicating increasing probability of a higher test score as a student’s ability increases. The grade 4 and grade 8 TCCs are parallel to each other, indicating similar overall test discrimination of the two assessments.

*Standard Error Curves*—Figure 6-26 shows Science test SE curves for grades 4 and 8. The SE curves are U-shaped, indicating smaller errors around ability estimates approximately in the middle of the scale score distribution. The SE is expected to be higher at the top and bottom ends of the ability scale, where fewer items measuring these students are found. Overall, the SEs around the scale score were found to be reasonable for Science assessments (for more details see Section 6.3.1 of this report).

*Growth at Quartiles*—The estimated scale scores for the Science calibration sample at the 25th, 50th, and 75th percentiles for both grade levels are presented in Figure 6-27. The data pattern presented in this figure indicates that the scale scores increase as the percentile increases within each grade level. Because the Science assessments are not on a vertical scale, it is not appropriate to compare scale scores between grades.

## Social Studies Scale

*Test Characteristic Curves*—Similar to Science, although the Social Studies assessments are not vertically scaled, the TCCs for grades 4, 8, and 10 are presented together in Figure 6-28 for comparison purposes. The TCCs are S-shaped, indicating increasing probability of a higher test score as a student’s ability increases. The grade 4 and grade 8 TCCs are parallel to each other, indicating similar overall test discrimination of the two assessments.

*Standard Error Curves*—Figure 6-29 shows Social Studies SE curves for grades 4, 8, and 10. The SE curves are U-shaped, indicating smaller errors around ability estimates approximately in the middle of the scale score distribution. The SE is expected to be higher at the top and bottom ends of the ability scale, where fewer items measuring these students are found. Overall, the SEs around the scale score were found to be reasonable for Science assessments (for more details see Section 6.3.1 of this report).

*Growth at Quartiles*—The estimated scale scores for the Social Studies calibration sample at the 25th, 50th, and 75th percentiles for all grade levels are presented in Figure 6-30. The data pattern presented in this figure indicates that the scale scores increase as the percentile increases within each grade level. Because the Social Studies assessments are not on a vertical scale, it is not appropriate to compare scale scores between grades.

### 6.3 Deriving Scale Scores in the Wisconsin Forward Exam

A scale score can be interpreted as a highly probable estimate of a student’s ability in a given content area. Scale scores are based on the student’s responses to all items on a given test and account for the characteristics of the items that are in the test (such as item difficulty).

Scale scores in the Wisconsin Forward Exam are based on the theoretical models of the item response process described above and elaborated upon below. The essential idea behind these models is that the probability of a correct response to a given item is a function of examinee ability and the characteristics of the item, such as the difficulty of the item. IRT models expect that as examinee ability increases, the probability of a correct response to a given item also increases, given certain conditions and assumptions. This description applies specifically to MC items; non-MC items are treated as CR items and are handled slightly differently but follow logic that is essentially the same.

Whether looking at an individual item or at a group of items that make up a complete test, IRT uses probability models to describe the relationship between a student’s ability and his or her observed scores. As described above, the 3PL model is used to estimate the probability of a correct response for each of the MC items. The model is provided here because its components are reviewed in the following paragraphs.

$$P(u_i = 1 | \theta) = c_i + \frac{1 - c_i}{1 + e^{-1.7a_i(\theta - b_i)}} \quad (1)$$

In this model,  $\theta$  denotes a measured ability (e.g., ELA ability) and  $u_i$  represents an observed score on a particular item. For MC items, the observed score  $u_i$  is either 0 or 1, indicating either an incorrect or correct response, respectively. For an MC item, the probability model can be denoted as  $P(u_i = 1 | \theta)$ . That is,  $P$  is an estimation of the probability that a student with an ability value  $\theta$  would answer item  $i$  correctly.

The terms on the right side of the equation above ( $a_i, b_i, c_i$ ) represent the parameters in the model: discrimination, difficulty (or location), and a pseudo-guessing factor. Discrimination refers to how well an item sorts students by ability level; difficulty represents the difficulty of the item or its location on an ability continuum; and the pseudo-guessing factor represents the probability of a low-ability student guessing the correct response.

Given any particular response pattern ( $u_1u_2 \cdots u_n$ ) on a test with some number of items ( $n$  items), the “likelihood function,” or the probability that a student with a given ability value ( $\theta$ ) would produce this particular response pattern, is given by

$$P(u_1u_2 \cdots u_n | \theta) = \prod_{i=1}^n P(u_i | \theta). \quad (2)$$

The formula indicates that the “estimated maximum likelihood” IRT item-pattern scoring method searches for the ability estimate ( $\theta_0$ ) that maximizes the probability function in (2) and it assigns an ability estimate ( $\theta_0$ ) as the test score for the student with the response pattern ( $u_1u_2 \cdots u_n$ ). In other words, the scale score is the most likely, or most probable, estimate of student ability produced in a context where item parameters are known and based on all of the items in a given test.

As indicated, the item-pattern scoring method takes into account not only a student’s total raw score but also the psychometric characteristics of all items the student responded to, including the items the student responded to incorrectly. It should be noted that a weight of 3 was applied to ELA TDA item scores in estimation of the student total test scale scores.

Consider the following example. Suppose six examinees in grade 4 take an ELA test with 30 MC items. Suppose further that the properties, or parameters, of the items on that test are as follows:

Table 6-A Example of Item Parameters for a Test

Item	Discrimination (a)	Location (b)	Guessing (c)	Item	Discrimination (a)	Location (b)	Guessing (c)
1	0.0341	318.75	0.16	16	0.0398	286.13	0.13
2	0.0342	244.62	0.20	17	0.0523	290.65	0.26
3	0.0234	257.56	0.20	18	0.0387	280.23	0.14
4	0.0306	235.00	0.20	19	0.0329	315.71	0.21
5	0.0125	342.39	0.17	20	0.0370	287.88	0.25
6	0.0305	261.51	0.16	21	0.0387	280.25	0.18
7	0.0316	296.93	0.19	22	0.0321	285.86	0.17
8	0.0228	252.70	0.20	23	0.0219	302.52	0.13
9	0.0383	266.28	0.20	24	0.0551	301.11	0.26
10	0.0229	308.84	0.11	25	0.0165	324.24	0.19
11	0.0536	259.00	0.21	26	0.0279	297.19	0.11
12	0.0478	245.19	0.20	27	0.0423	296.06	0.28
13	0.0418	276.25	0.28	28	0.0658	324.76	0.21
14	0.0377	287.60	0.23	29	0.0488	281.56	0.32
15	0.0177	316.08	0.24	30	0.0237	345.32	0.37

Now suppose that the student response patterns for these six examinees are as follows, where 0 represents an incorrect response and 1 represents a correct response:

Table 6-B Example of Item Response Pattern

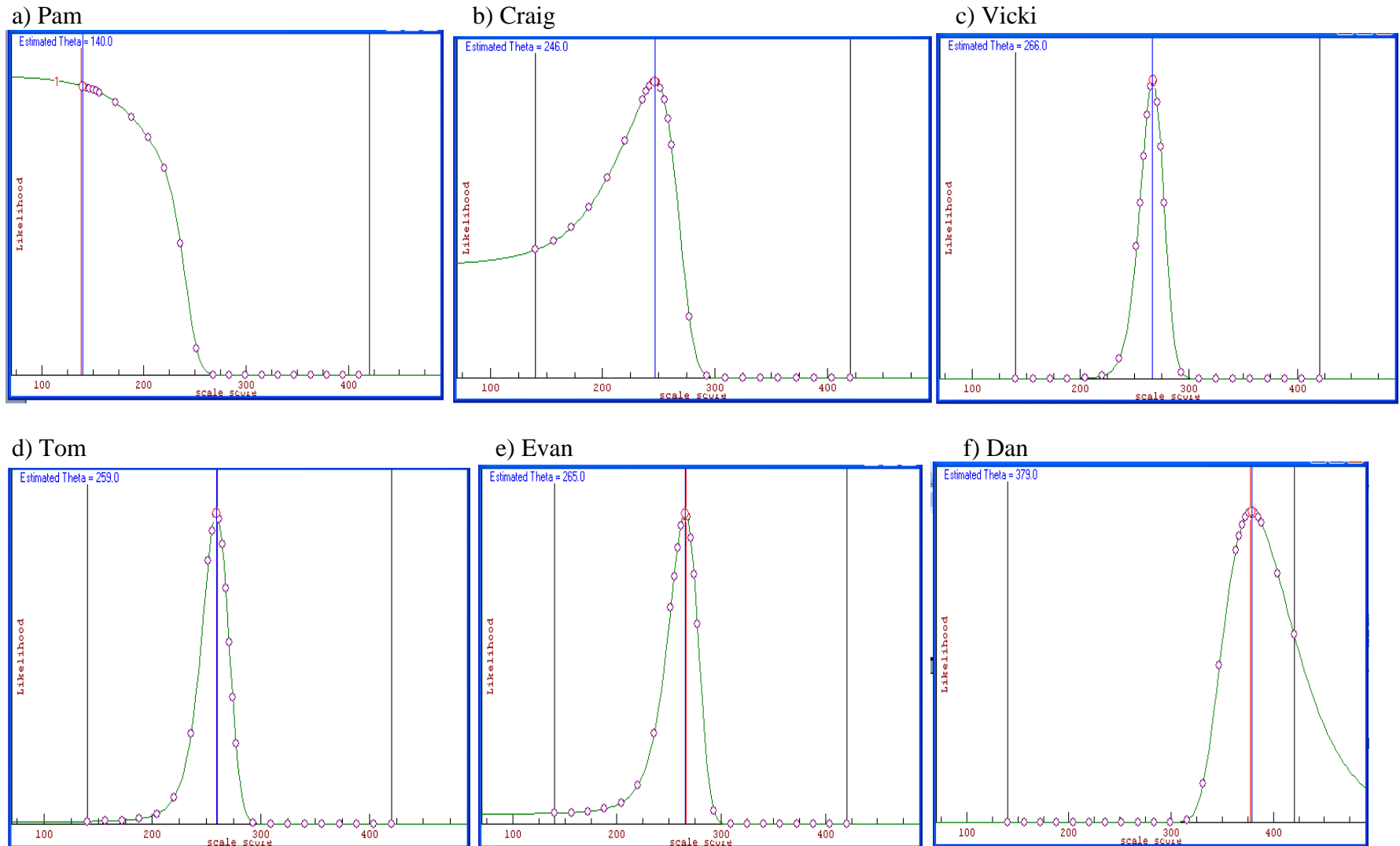
Student	Response Pattern ( $u_1u_2 \cdots u_n$ )	Raw Score	Item-Pattern Score
Pam	1000011001010000000000000101	7	140
Craig	101010101010101010101010101010	15	246
Vicki	010101010101010101010101010101	15	266
Tom	001100110011001100110011001101	15	259
Evan	110011001100110011001100110010	15	265
Dan	1111111111111111111111111011111	29	379

The first student, Pam, answered 7 of the items correctly and obtained a scale score of 140, which is equal to the lowest point on the scale score range, called the “lowest obtainable scale score,” or LOSS. The next four students each answered 15 out of 30 items correctly, but the response pattern of each of these students is different. The raw score of each of these students is 15. However, the maximum likelihood item-pattern scoring method produced a different scale score for each examinee. Scale scores were 246 for Craig, 266 for Vicki, 259 for Tom, and 265 for Evan. These scores can be accounted for by considering the pattern of the student responses on the test together with the properties (or parameters) of the items, as shown in Table 6-A. By referring to Table 6-A, the reader can observe that Vicki and Evan answered some difficult and highly discriminating items correctly, whereas Craig and Tom did not. The remaining student, Dan, scored 29 out of the 30 items correctly and obtained a scale score of 379, which is near the upper limit of the scale score range, called the “highest obtainable scale score,” or HOSS.

Figure 6-A below shows the probability of each ability estimate (or scale score) for the six examinees. The total scale score range for the test is plotted on the horizontal axis. As indicated by the two vertical lines in the plot, the lower and upper limits of the scale score range are 140 and 420, respectively. The likelihood, or probability, of all possible ability estimates for each examinee is plotted on the vertical axis and ranges from 0 to 1.0. The higher the likelihood, the more probable it is that the ability estimate actually reflects the examinee’s ability level.

As indicated above, scale scores are the most likely, or the maximum likelihood, estimates of examinee ability. As can be observed for Vicki, Tom, and Evan, scores that are plus or minus only a few scale score points are markedly less likely estimates of their ability. The same is true for Craig and Dan, though to a slightly lesser extent. In the case of Pam, a few scores were almost as likely as the maximum likelihood estimate reported. Those scores that appear to be more likely than the reported score are outside of the scale score range of the test (below the LOSS).

Figure 6-A Examples of Likelihood Functions, or the Probability of Each Ability Level Estimate (or Scale Score)



Note: The circular dots in the likelihood functions indicate that the software program used is searching for a maximum likelihood estimate (scale score) for the student.

There are two IRT-based scoring methods generally used for large-scale assessments: number-correct scoring and item-pattern scoring. Item-pattern scoring may be recommended over number-correct scoring for several reasons. Two reasons, accuracy and reliability, are pertinent for present purposes.

Item-pattern scoring generally produces more accurate scores for individual students. Specifically, it produces a smaller conditional standard error of measurement (CSEM) across the scale score range for a given test compared to number-correct scoring. The smaller the CSEM, the more confident one can be in the accuracy of the test results. The increase in accuracy provided by item-pattern scoring is equivalent, on average, to approximately a 15% to 20% increase in test length (Yen, 1984; Yen & Candell, 1991).

Second, reliability tends to be higher using item-pattern scoring, which means (a) fewer items are needed to achieve a given level of reliability and (b) a given test with a given number of items will have higher reliability than when using number-correct scoring. Yen (1984) has demonstrated that an equivalent level of reliability for a 20-item test scored by the number-correct scoring method could be obtained with a 16- or 17-item test scored by the item-pattern scoring method.

The procedures applied here are consistent with student scoring in prior Wisconsin Knowledge and Concepts Examinations. Several supplements to this simplified outline of IRT are available. Introductory discussions of IRT can be found in *Educational Measurement* (Linn, 1989) or Chapter 11 in *Introduction to Measurement Theory* (Allen & Yen, 1979). More advanced discussions of partial-credit models may be found in Muraki (1990, 1992), Yen (1993), and van der Linden and Hambleton (1997). For additional information on the technical details of item-pattern scoring, readers can also refer to Yen & Candell (1991).

### **6.3.1 Conditional Standard Error of Measurement**

One way of characterizing the reliability of a reported test score is by examining the standard error associated with the score. An observed score should not be regarded as an absolute value but as a point within a range that with a certain degree of probability includes a student's true score. The CSEM is defined as the reciprocal of the square root of the test information function and can be estimated across all points of the ability continuum (Hambleton & Swaminathan, 1985). The CSEM can be used to obtain the range within which a student's true score is likely to fall, that is, with a certain degree of probability. It is expected that 68% of the time a student's score obtained from a single testing will fall within one CSEM of that student's true score and that 95% of the time the obtained score will fall within two CSEMs of the true score.

Standard 2.13 of the AERA, APA, & NCME (2014) *Standards* states the following:

The standard error of measurement, both overall and conditional (if reported), should be provided in units of each reported score. (45)



The CSEM of the scale scores in the Spring 2017 Wisconsin Forward Exam is displayed graphically for each grade and content area in Figures 6-20 (for ELA), 6-23 (for Mathematics), 6-26 (for Science), and 6-29 (for Social Studies). The CSEM provided is based on item-pattern scoring. Each CSEM curve is plotted as a function of the scale scores. These figures show the scale score range within which measurement is most accurate. The figures also show that extreme scale scores have more measurement error than scores in the middle of the distribution. Scale scores in the high or low extremes of the student distribution are less precise than those in the middle of the distribution because there tend to be fewer test items in these score areas and fewer students. The lower and upper limits of the scale, referred to as the lowest obtainable scale score (LOSS) and highest obtainable scale score (HOSS), are the starting scale score and the last scale score in these figures. LOSS and HOSS are further discussed in the next section.

Because of the nature of item-pattern scoring, a scoring table showing a simple, direct conversion of raw score to scale score cannot be generated for the Spring 2017 Wisconsin Forward Exam. However, scoring tables showing an approximate raw score-to-scale score relationship, and the associated CSEM can be produced, and they are provided in Tables 6-9 through 6-25. These tables are provided to illustrate the approximate raw score-to-scale score relationship for each unique raw score and do not include all combination of raw score-to-scale score associations.

### **6.3.2 LOSS and HOSS**

As has been established, a scale score is a maximum likelihood ability estimate. The maximum-likelihood procedure cannot produce scale score estimates for students with perfect scores or scores below the scoring level expected by guessing. Although maximum likelihood estimates are available for students with extreme scores other than zero or a perfect score, these estimates generally have large SEMs. Therefore, scores are established for these extreme highs and lows based on a rational, but necessarily non maximum, likelihood procedure. These values are set separately by grade and called the LOSS and the HOSS. The LOSS and HOSS values for the Wisconsin Forward Exam were established after the Spring 2016 test administration and remained unchanged in the Spring 2017 test administration.

Table 6-26 shows the number and percentage of students at the LOSS and the HOSS. In general, there should not be many students clustered at the LOSS or HOSS. An accumulation of a high proportion of students in the LOSS or HOSS may indicate a floor or ceiling effect.

It should be noted that for ELA and Mathematics the LOSS and HOSS values were set in such a way during the Spring 2016 scale development, that they increase as the grade level increases. Setting increasing LOSS as the grade level increases is an important property of a vertical scale and constrains student ability in each grade in such a way that the lowest-ability students in a given grade will always have a higher scale score than the lowest-ability students in a grade below and a lower scale score than the lowest-ability students in a grade above. Conversely, setting increasing HOSS values as the grade level increases constrains student ability in each grade in such a way that the highest-ability students in a given grade will always have a higher scale score than the highest-ability students in a grade below and a lower scale score than the highest-ability students in a grade above.

In most grades and content areas, the percentage of students at the LOSS and HOSS was small: less than 1%. However, in some grades and content areas the LOSS percentages were larger. In Mathematics, all grades, except grade 3 had more than 1% of students at the LOSS: grade 4—3.14%, grade 5—1.87%, grade 6—2.16%, grade 7—3.09%, and grade 8—4.67%. These percentages at the LOSS indicate that the Mathematics assessments were difficult for some students and that they can be considered as a point of reference when developing future forms. The percentage at the LOSS in these grades may be reduced in future years by including some additional items that are less difficult. The percentage of students scoring at the HOSS ranged from 0 in ELA grades 3 through 6 to 1% in Social Studies grade 4 and over 1% in Social Studies grade 8 (1.40%). The percentage scoring at the HOSS may be reduced by including some additional difficult items or by including more items on the test.

## **6.4 Summary**

In summary, the overall purpose of the test scaling and equating is to ensure that the test items, as well as the overall test, are functioning appropriately. It also helps maintain the test scale across years so that test results may be appropriately compared across years. The data analyses undertaken by DRC are in alignment with multiple best practices of the testing industry and, in particular, support the following AERA, APA, & NCME (2014) *Standards*: 1.8, 2.13, 5.2, and 7.2.

Table 6-1 English Language Arts Calibration Sample Demographics Compared to Population

Grade 3	Calibration Sample		Population		Difference
	N	%	N	%	%
All Students	63795		63946		
<b>Gender</b>					
Male	32479	50.91	32569	50.93	-0.02
Female	31316	49.09	31377	49.07	0.02
<b>Race/Ethnicity</b>					
White	42291	66.29	42335	66.20	0.09
Black	7093	11.12	7151	11.18	-0.06
Hispanic	8524	13.36	8557	13.38	-0.02
Asian/Pacific Islander	2555	4.01	2565	4.01	-0.01
American Indian	766	1.20	767	1.20	0.00
Other	2566	4.02	2571	4.02	0.00
<b>LEP</b>					
No	58081	91.04	58206	91.02	0.02
Yes	5714	8.96	5740	8.98	-0.02
<b>Disability</b>					
No	56695	88.87	56823	88.86	0.01
Yes	7100	11.13	7123	11.14	-0.01
<b>SES Disadvantaged</b>					
No	37414	58.65	37459	58.58	0.07
Yes	26381	41.35	26487	41.42	-0.07
<b>Grade 4</b>	<b>N</b>	<b>%</b>	<b>N</b>	<b>%</b>	<b>%</b>
All Students	64249		64423		
<b>Gender</b>					
Male	32884	51.18	32975	51.19	0.00
Female	31365	48.82	31448	48.81	0.00
<b>Race/Ethnicity</b>					
White	42947	66.85	43005	66.75	0.09
Black	6971	10.85	7038	10.92	-0.07
Hispanic	8514	13.25	8543	13.26	-0.01
Asian/Pacific Islander	2499	3.89	2509	3.89	0.00
American Indian	834	1.30	838	1.30	0.00
Other	2483	3.86	2490	3.87	0.00
<b>LEP</b>					
No	59403	92.46	59561	92.45	0.01
Yes	4845	7.54	4862	7.55	-0.01
<b>Disability</b>					
No	56927	88.61	57066	88.58	0.02
Yes	7321	11.39	7357	11.42	-0.02
<b>SES Disadvantaged</b>					
No	38275	59.57	38334	59.50	0.07
Yes	25973	40.43	26089	40.50	-0.07

Table 6-1 English Language Arts Calibration Sample Demographics Compared to Population (cont.)

Grade 5	Calibration Sample		Population		Difference
	N	%	N	%	%
All Students	62898		62995		
<b>Gender</b>					
Male	32251	51.28	32305	51.28	-0.01
Female	30647	48.72	30690	48.72	0.01
<b>Race/Ethnicity</b>					
White	42200	67.09	42228	67.03	0.06
Black	6751	10.73	6799	10.79	-0.06
Hispanic	8372	13.31	8389	13.32	-0.01
Asian/Pacific Islander	2521	4.01	2524	4.01	0.00
American Indian	803	1.28	803	1.27	0.00
Other	2249	3.58	2252	3.57	0.00
<b>LEP</b>					
No	59247	94.20	59339	94.20	0.00
Yes	3649	5.80	3656	5.80	0.00
<b>Disability</b>					
No	55576	88.36	55649	88.34	0.02
Yes	7320	11.64	7346	11.66	-0.02
<b>SES Disadvantaged</b>					
No	37757	60.03	37784	59.98	0.05
Yes	25139	39.97	25211	40.02	-0.05
<b>Grade 6</b>	<b>N</b>	<b>%</b>	<b>N</b>	<b>%</b>	<b>%</b>
All Students	62598		62754		
<b>Gender</b>					
Male	31945	51.03	32028	51.04	-0.01
Female	30653	48.97	30726	48.96	0.01
<b>Race/Ethnicity</b>					
White	42973	68.66	43028	68.57	0.10
Black	6434	10.28	6502	10.36	-0.08
Hispanic	7854	12.55	7886	12.57	-0.02
Asian/Pacific Islander	2447	3.91	2453	3.91	0.00
American Indian	791	1.26	793	1.26	0.00
Other	2086	3.33	2092	3.33	0.00
<b>LEP</b>					
No	59713	95.41	59866	95.40	0.01
Yes	2872	4.59	2888	4.60	-0.01
<b>Disability</b>					
No	55693	88.99	55812	88.94	0.05
Yes	6892	11.01	6942	11.06	-0.05
<b>SES Disadvantaged</b>					
No	39083	62.45	39136	62.36	0.08
Yes	23502	37.55	23618	37.64	-0.08

Table 6-1 English Language Arts Calibration Sample Demographics Compared to Population (cont.)

Grade 7	Calibration Sample		Population		Difference
	N	%	N	%	%
All Students	62911		63091		
<b>Gender</b>					
Male	32261	51.28	32352	51.28	0.00
Female	30650	48.72	30739	48.72	0.00
<b>Race/Ethnicity</b>					
White	43792	69.62	43857	69.51	0.10
Black	6261	9.95	6317	10.01	-0.06
Hispanic	7677	12.20	7717	12.23	-0.03
Asian/Pacific Islander	2414	3.84	2422	3.84	0.00
American Indian	789	1.25	793	1.26	0.00
Other	1972	3.13	1985	3.15	-0.01
<b>LEP</b>					
No	60275	95.82	60448	95.81	0.01
Yes	2630	4.18	2643	4.19	-0.01
<b>Disability</b>					
No	55913	88.88	56055	88.85	0.04
Yes	6992	11.12	7036	11.15	-0.04
<b>SES Disadvantaged</b>					
No	40121	63.78	40187	63.70	0.08
Yes	22784	36.22	22904	36.30	-0.08
<b>Grade 8</b>	<b>N</b>	<b>%</b>	<b>N</b>	<b>%</b>	<b>%</b>
All Students	61863		62109		
<b>Gender</b>					
Male	31791	51.39	31914	51.38	0.01
Female	30072	48.61	30195	48.62	-0.01
<b>Race/Ethnicity</b>					
White	43400	70.16	43481	70.01	0.15
Black	6071	9.81	6186	9.96	-0.15
Hispanic	7551	12.21	7583	12.21	0.00
Asian/Pacific Islander	2327	3.76	2332	3.75	0.01
American Indian	766	1.24	771	1.24	0.00
Other	1747	2.82	1756	2.83	0.00
<b>LEP</b>					
No	59278	95.82	59509	95.81	0.01
Yes	2584	4.18	2600	4.19	-0.01
<b>Disability</b>					
No	55089	89.05	55266	88.98	0.07
Yes	6773	10.95	6843	11.02	-0.07
<b>SES Disadvantaged</b>					
No	40136	64.88	40221	64.76	0.12
Yes	21726	35.12	21888	35.24	-0.12

Table 6-2 Mathematics Calibration Sample Demographics Compared to Population

Grade 3	Calibration Sample		Population		Difference
	N	%	N	%	%
All Students	60266		64066		
<b>Gender</b>					
Male	30674	50.90	32629	50.93	-0.03
Female	29592	49.10	31437	49.07	0.03
<b>Race/Ethnicity</b>					
White	40095	66.53	42346	66.10	0.43
Black	6534	10.84	7162	11.18	-0.34
Hispanic	8010	13.29	8618	13.45	-0.16
Asian/Pacific Islander	2471	4.10	2601	4.06	0.04
American Indian	737	1.22	768	1.20	0.02
Other	2419	4.01	2571	4.01	0.00
<b>LEP</b>					
No	54767	90.88	58197	90.84	0.04
Yes	5499	9.12	5869	9.16	-0.04
<b>Disability</b>					
No	53613	88.96	56938	88.87	0.09
Yes	6653	11.04	7128	11.13	-0.09
<b>SES Disadvantaged</b>					
No	35584	59.04	37492	58.52	0.52
Yes	24682	40.96	26574	41.48	-0.52
<b>Grade 4</b>	<b>N</b>	<b>%</b>	<b>N</b>	<b>%</b>	<b>%</b>
All Students	61537		64533		
<b>Gender</b>					
Male	31506	51.20	33028	51.18	0.02
Female	30031	48.80	31505	48.82	-0.02
<b>Race/Ethnicity</b>					
White	41234	67.01	43021	66.67	0.34
Black	6446	10.47	7042	10.91	-0.44
Hispanic	8202	13.33	8599	13.32	0.00
Asian/Pacific Islander	2464	4.00	2544	3.94	0.06
American Indian	814	1.32	837	1.30	0.03
Other	2377	3.86	2490	3.86	0.00
<b>LEP</b>					
No	56742	92.21	59552	92.28	-0.07
Yes	4795	7.79	4981	7.72	0.07
<b>Disability</b>					
No	54573	88.68	57159	88.57	0.11
Yes	6964	11.32	7374	11.43	-0.11
<b>SES Disadvantaged</b>					
No	36830	59.85	38392	59.49	0.36
Yes	24707	40.15	26141	40.51	-0.36

Table 6-2 Mathematics Calibration Sample Demographics Compared to Population (cont.)

Grade 5	Calibration Sample		Population		Difference
	N	%	N	%	%
All Students	59148		63152		
<b>Gender</b>					
Male	30358	51.33	32373	51.26	0.06
Female	28790	48.67	30779	48.74	-0.06
<b>Race/Ethnicity</b>					
White	39740	67.19	42238	66.88	0.30
Black	6221	10.52	6842	10.83	-0.32
Hispanic	7898	13.35	8456	13.39	-0.04
Asian/Pacific Islander	2442	4.13	2562	4.06	0.07
American Indian	768	1.30	802	1.27	0.03
Other	2079	3.51	2252	3.57	-0.05
<b>LEP</b>					
No	55572	93.95	59362	94.00	-0.04
Yes	3576	6.05	3790	6.00	0.04
<b>Disability</b>					
No	52328	88.47	55801	88.36	0.11
Yes	6820	11.53	7351	11.64	-0.11
<b>SES Disadvantaged</b>					
No	35585	60.16	37851	59.94	0.23
Yes	23563	39.84	25301	40.06	-0.23
<b>Grade 6</b>	<b>N</b>	<b>%</b>	<b>N</b>	<b>%</b>	<b>%</b>
All Students	60667		62847		
<b>Gender</b>					
Male	30973	51.05	32061	51.01	0.04
Female	29694	48.95	30786	48.99	-0.04
<b>Race/Ethnicity</b>					
White	41673	68.69	43044	68.49	0.20
Black	6121	10.09	6516	10.37	-0.28
Hispanic	7661	12.63	7923	12.61	0.02
Asian/Pacific Islander	2424	4.00	2476	3.94	0.06
American Indian	772	1.27	795	1.26	0.01
Other	2016	3.32	2093	3.33	-0.01
<b>LEP</b>					
No	57776	95.23	59875	95.27	-0.04
Yes	2891	4.77	2972	4.73	0.04
<b>Disability</b>					
No	54090	89.16	55891	88.93	0.23
Yes	6577	10.84	6956	11.07	-0.23
<b>SES Disadvantaged</b>					
No	37901	62.47	39172	62.33	0.14
Yes	22766	37.53	23675	37.67	-0.14

Table 6-2 Mathematics Calibration Sample Demographics Compared to Population (cont.)

Grade 7	Calibration Sample		Population		Difference
	N	%	N	%	%
All Students	60268		63200		
<b>Gender</b>					
Male	31073	51.56	32417	51.29	0.27
Female	29195	48.44	30783	48.71	-0.27
<b>Race/Ethnicity</b>					
White	42229	70.07	43856	69.39	0.68
Black	5782	9.59	6349	10.05	-0.45
Hispanic	7357	12.21	7770	12.29	-0.09
Asian/Pacific Islander	2248	3.73	2446	3.87	-0.14
American Indian	749	1.24	792	1.25	-0.01
Other	1902	3.16	1987	3.14	0.01
<b>LEP</b>					
No	57710	95.76	60447	95.64	0.11
Yes	2557	4.24	2753	4.36	-0.11
<b>Disability</b>					
No	53592	88.92	56158	88.86	0.07
Yes	6675	11.08	7042	11.14	-0.07
<b>SES Disadvantaged</b>					
No	38638	64.11	40223	63.64	0.47
Yes	21629	35.89	22977	36.36	-0.47
<b>Grade 8</b>	<b>N</b>	<b>%</b>	<b>N</b>	<b>%</b>	<b>%</b>
All Students	60097		62175		
<b>Gender</b>					
Male	30936	51.48	31956	51.40	0.08
Female	29161	48.52	30219	48.60	-0.08
<b>Race/Ethnicity</b>					
White	42351	70.47	43477	69.93	0.54
Black	5636	9.38	6181	9.94	-0.56
Hispanic	7365	12.26	7629	12.27	-0.02
Asian/Pacific Islander	2298	3.82	2355	3.79	0.04
American Indian	744	1.24	771	1.24	0.00
Other	1703	2.83	1762	2.83	0.00
<b>LEP</b>					
No	57478	95.64	59479	95.66	-0.02
Yes	2619	4.36	2696	4.34	0.02
<b>Disability</b>					
No	53578	89.15	55338	89.00	0.15
Yes	6519	10.85	6837	11.00	-0.15
<b>SES Disadvantaged</b>					
No	39131	65.11	40214	64.68	0.43
Yes	20966	34.89	21961	35.32	-0.43



Table 6-3 Science Calibration Sample Demographics Compared to Population

Grade 4	Calibration Sample		Population		Difference
	N	%	N	%	%
All Students	62121		64520		
<b>Gender</b>					
Male	31763	51.13	33022	51.18	-0.05
Female	30358	48.87	31498	48.82	0.05
<b>Race/Ethnicity</b>					
White	41530	66.85	43028	66.69	0.16
Black	6617	10.65	7032	10.90	-0.25
Hispanic	8318	13.39	8591	13.32	0.07
Asian/Pacific Islander	2457	3.96	2547	3.95	0.01
American Indian	813	1.31	839	1.30	0.01
Other	2386	3.84	2483	3.85	-0.01
<b>LEP</b>					
No	57288	92.22	59546	92.29	-0.07
Yes	4833	7.78	4974	7.71	0.07
<b>Disability</b>					
No	55090	88.68	57143	88.57	0.12
Yes	7031	11.32	7377	11.43	-0.12
<b>SES Disadvantaged</b>					
No	37015	59.59	38378	59.48	0.10
Yes	25106	40.41	26142	40.52	-0.10
<b>Grade 8</b>	<b>N</b>	<b>%</b>	<b>N</b>	<b>%</b>	<b>%</b>
All Students	60461		62113		
<b>Gender</b>					
Male	31079	51.40	31921	51.39	0.01
Female	29382	48.60	30192	48.61	-0.01
<b>Race/Ethnicity</b>					
White	42536	70.35	43469	69.98	0.37
Black	5711	9.45	6139	9.88	-0.44
Hispanic	7444	12.31	7627	12.28	0.03
Asian/Pacific Islander	2326	3.85	2353	3.79	0.06
American Indian	748	1.24	768	1.24	0.00
Other	1696	2.81	1757	2.83	-0.02
<b>LEP</b>					
No	57821	95.63	59421	95.67	-0.03
Yes	2640	4.37	2692	4.33	0.03
<b>Disability</b>					
No	53951	89.23	55292	89.02	0.21
Yes	6510	10.77	6821	10.98	-0.21
<b>SES Disadvantaged</b>					
No	39359	65.10	40217	64.75	0.35
Yes	21102	34.90	21896	35.25	-0.35

Table 6-4 Social Studies Calibration Sample Demographics Compared to Population

Grade 4	Calibration Sample		Population		Difference
	N	%	N	%	%
All Students	61471		64512		
<b>Gender</b>					
Male	31464	51.19	33016	51.18	0.01
Female	30007	48.81	31496	48.82	-0.01
<b>Race/Ethnicity</b>					
White	41118	66.89	43022	66.69	0.20
Black	6481	10.54	7029	10.90	-0.35
Hispanic	8262	13.44	8589	13.31	0.13
Asian/Pacific Islander	2429	3.95	2546	3.95	0.00
American Indian	802	1.30	838	1.30	0.01
Other	2379	3.87	2488	3.86	0.01
<b>LEP</b>					
No	56650	92.16	59540	92.29	-0.14
Yes	4821	7.84	4972	7.71	0.14
<b>Disability</b>					
No	54500	88.66	57139	88.57	0.09
Yes	6971	11.34	7373	11.43	-0.09
<b>SES Disadvantaged</b>					
No	36528	59.42	38384	59.50	-0.08
Yes	24943	40.58	26128	40.50	0.08
<b>Grade 8</b>	<b>N</b>	<b>%</b>	<b>N</b>	<b>%</b>	<b>%</b>
All Students	59995		62079		
<b>Gender</b>					
Male	30820	51.37	31900	51.39	-0.02
Female	29175	48.63	30179	48.61	0.02
<b>Race/Ethnicity</b>					
White	42184	70.31	43457	70.00	0.31
Black	5653	9.42	6129	9.87	-0.45
Hispanic	7407	12.35	7611	12.26	0.09
Asian/Pacific Islander	2316	3.86	2355	3.79	0.07
American Indian	749	1.25	769	1.24	0.01
Other	1686	2.81	1758	2.83	-0.02
<b>LEP</b>					
No	57372	95.63	59390	95.67	-0.04
Yes	2623	4.37	2689	4.33	0.04
<b>Disability</b>					
No	53520	89.21	55262	89.02	0.19
Yes	6475	10.79	6817	10.98	-0.19
<b>SES Disadvantaged</b>					
No	39033	65.06	40205	64.76	0.30
Yes	20962	34.94	21874	35.24	-0.30

Table 6-4 Social Studies Calibration Sample Demographics Compared to Population (cont.)

Grade 10	Calibration Sample		Population		Difference
	N	%	N	%	%
All Students	59307		63764		
<b>Gender</b>					
Male	30272	51.04	32530	51.02	0.03
Female	29035	48.96	31234	48.98	-0.03
<b>Race/Ethnicity</b>					
White	43616	73.54	46533	72.98	0.57
Black	4702	7.93	5459	8.56	-0.63
Hispanic	6522	11.00	7021	11.01	-0.01
Asian/Pacific Islander	2338	3.94	2443	3.83	0.11
American Indian	630	1.06	709	1.11	-0.05
Other	1498	2.53	1599	2.51	0.02
<b>LEP</b>					
No	57369	96.73	61666	96.71	0.02
Yes	1937	3.27	2098	3.29	-0.02
<b>Disability</b>					
No	53365	89.98	57164	89.65	0.33
Yes	5941	10.02	6600	10.35	-0.33
<b>SES Disadvantaged</b>					
No	40794	68.79	43595	68.37	0.42
Yes	18512	31.21	20169	31.63	-0.42

Table 6-5 Item Flagged Based on Yen’s Q1

Content	Grade	Item Number in Calibration	Type	N	Z	Critical Z
ELA	4	36	CR	64248	242.75	171.33
	5	25	CR	62898	293.23	167.73
	5	32	CR	62898	484.14	167.73
	7	8*	MC	62890	172.21	167.71
	7	9	CR	62890	515.77	167.71
	7	32	CR	62890	306.25	167.71
	8	2	CR	61851	704.14	164.94
	8	32*	CR	61851	376.02	164.94
Math	3	34*	CR	60121	189.79	160.32
	7	27	CR	60221	186.48	160.59
	8	25*	MC	60068	163.36	160.18
Social Studies	8	31	MC	59143	166.35	157.71

Note: An asterisk (\*) indicates an anchor item.

Table 6-6 Equating Evaluation Results, Stocking and Lord Method

Content Area	Grade	Number of Anchors	Stocking and Lord TCC Method Results						Equating Constants	
			TCC Results		Parameter Comparison Statistics				A	B
			# Iterations	F Value	a-Parameter		b-Parameter			
					Corr	# RMSD Outliers	Corr	# RMSD Outliers		
ELA	3	12	10	0.2426	0.98	0	1.00	0	0.9462	-1.1245
	4	13	10	0.2926	0.98	0	0.97	0	1.0491	-0.537
	5	14	7	0.2305	0.98	1	0.99	0	1.0215	-0.1329
	6	15	7	0.1203	0.99	0	1.00	0	0.9853	0.1256
	7	17	12	0.2939	0.97	0	0.98	2	1.1626	0.4168
	8	22	6	0.1960	0.97	1	0.99	1	1.2351	0.6321
Math	3	31	4	0.0506	0.97	1	0.99	2	0.8850	-1.1844
	4	38	17	0.1441	0.97	3	0.99	1	0.9338	-0.7351
	5	22	19	0.1607	0.96	0	0.99	1	0.8703	-0.1828
	6	27	15	0.1544	0.97	1	0.99	1	1.0054	0.0707
	7	30	24	0.1674	0.98	0	1.00	1	1.0123	0.4604
	8	22	32	0.0460	0.99	1	0.97	1	0.9538	0.8401
Science	4	12	26	0.0566	1.00	0	0.98	1	1.0336	-0.0294
	8	12	45	0.0798	0.99	0	1.00	0	1.0402	-0.1327
Social Studies	4*	13	20	0.0475	0.97	0	0.94	1	1.0384	-0.2576
	<b>4**</b>	<b>12</b>	<b>22</b>	<b>0.0769</b>	<b>0.98</b>	<b>0</b>	<b>0.99</b>	<b>1</b>	<b>1.0626</b>	<b>-0.1902</b>
	8	13	18	0.0487	0.98	0	0.99	0	1.0605	-0.0511
	10	17	9	0.0888	0.98	1	0.99	1	1.067	0.0001

\* Equating run with all anchor items included

\*\* Equating run with the flagged anchor item excluded (final).

Table 6-7 Statistics Comparing IRT Item-Ability Regression Curves for Flagged Anchor Items

Content Area	Grade	Anchor Item Position	UnWtd RMSD	UnWtd Mean Abs	Max Abs	UnWtd Mean	Wtd RMSD	Wtd Mean Abs	Wtd Mean
Social Studies	4	24	0.0902	0.0702	0.1702	-0.0702	0.1355	0.129	-0.129

Table 6-8 Scale Transformation Constants

Content Area	Grade	Scale Transformation Constants	
		M1	M2
ELA	3-8	43.7445	610.4987
Mathematics	3-8	46.4684	612.0818
Science	4	42.5532	401.7021
	8	39.5570	603.5601
Social Studies	4	40.1929	405.2251
	8	42.2297	600.8446
	10	42.8817	703.8594

Table 6-9 Scoring Table for English Language Arts Grade 3

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	330	84	31	576	13
1	330	84	32	580	13
2	330	84	33	585	14
3	330	84	34	590	14
4	358	62	35	595	14
5	401	39	36	600	14
6	424	31	37	606	15
7	440	27	38	612	15
8	453	23	39	618	16
9	463	21	<b>40</b>	<b>625</b>	<b>17</b>
10	472	20	41	632	18
11	480	18	42	641	19
12	487	17	43	650	20
13	493	17	44	660	21
14	499	16	45	672	23
15	504	15	46	685	24
16	510	15	47	700	26
17	515	14	48	717	28
18	519	14	49	738	30
<b>19</b>	<b>524</b>	<b>14</b>	50	762	32
20	528	14	51	790	35
21	533	13	52	829	45
22	537	13	53	900	94
23	541	13			
24	546	13			
25	550	13			
26	554	13			
27	558	13			
28	563	13			
29	567	13			
<b>30</b>	<b>571</b>	<b>13</b>			

Note: **Bold** represents CSEM around cut score (or the next higher scale score if the cut score value is not in the table).

Table 6-10 Scoring Table for English Language Arts Grade 4

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	340	72	31	588	13
1	340	72	<b>32</b>	<b>592</b>	<b>13</b>
2	340	72	33	596	13
3	340	72	34	601	13
4	340	72	35	605	13
5	340	72	36	610	14
6	372	56	37	615	14
7	411	41	38	620	14
8	435	34	39	625	14
9	453	29	40	630	15
10	467	26	41	636	15
11	478	24	42	641	15
12	488	22	43	647	16
13	497	20	<b>44</b>	<b>654</b>	<b>16</b>
14	504	19	45	661	17
15	511	18	46	669	18
16	518	17	47	677	19
17	524	17	48	687	21
18	529	16	49	699	23
19	534	15	50	713	27
20	539	15	51	732	32
21	544	15	52	760	39
<b>22</b>	<b>549</b>	<b>14</b>	53	798	44
23	554	14	54	842	43
24	558	14	55	890	49
25	562	13	56	930	66
26	567	13			
27	571	13			
28	575	13			
29	579	13			
30	583	13			

Note: **Bold** represents CSEM around cut score (or the next higher scale score if the cut score value is not in the table).

Table 6-11 Scoring Table for English Language Arts Grade 5

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	350	90	<b>31</b>	<b>611</b>	<b>14</b>
1	350	90	32	616	14
2	350	90	33	621	14
3	350	90	34	626	14
4	350	90	35	631	14
5	350	90	36	637	15
6	407	51	37	642	15
7	439	37	38	648	15
8	459	30	39	654	16
9	474	27	40	660	16
10	486	24	41	667	17
11	496	22	<b>42</b>	<b>674</b>	<b>18</b>
12	505	21	43	682	18
13	513	20	44	691	19
14	520	19	45	700	20
15	527	18	46	710	21
16	534	17	47	722	23
17	540	17	48	734	23
18	546	17	49	748	24
19	552	16	50	762	25
20	557	16	51	778	26
21	562	16	52	795	27
<b>22</b>	<b>568</b>	<b>15</b>	53	815	29
23	573	15	54	839	33
24	578	15	55	875	46
25	583	14	56	940	94
26	588	14			
27	592	14			
28	597	14			
29	602	14			
30	607	14			

Note: **Bold** represents CSEM around cut score (or the next higher scale score if the cut score value is not in the table).



Table 6-12 Scoring Table for English Language Arts Grade 6

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	360	76	31	613	14
1	360	76	32	618	14
2	360	76	<b>33</b>	<b>623</b>	<b>14</b>
3	360	76	34	628	15
4	360	76	35	633	15
5	415	45	36	638	15
6	443	35	37	643	15
7	462	29	38	649	15
8	476	26	39	654	16
9	487	24	40	660	16
10	497	22	41	666	16
11	506	21	<b>42</b>	<b>673</b>	<b>17</b>
12	514	19	43	680	18
13	521	18	44	687	18
14	528	18	45	695	19
15	534	17	46	704	20
16	540	16	47	713	21
17	546	16	48	724	22
18	551	16	49	735	23
19	556	15	50	748	25
20	561	15	51	762	27
21	566	15	52	779	29
22	571	15	53	800	34
<b>23</b>	<b>576</b>	<b>14</b>	54	828	42
24	580	14	55	876	64
25	585	14	56	950	124
26	590	14			
27	594	14			
28	599	14			
29	604	14			
30	608	14			

Note: **Bold** represents CSEM around cut score (or the next higher scale score if the cut score value is not in the table).

Table 6-13 Scoring Table for English Language Arts Grade 7

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	370	69	31	621	14
1	370	69	32	625	15
2	370	69	33	630	15
3	370	69	34	634	15
4	370	69	<b>35</b>	<b>639</b>	<b>15</b>
5	370	69	36	644	15
6	393	58	37	648	15
7	432	43	38	653	15
8	457	37	39	658	15
9	476	32	40	664	15
10	491	30	41	669	16
11	504	27	42	675	16
12	515	26	43	681	17
13	525	24	44	688	17
14	533	22	45	695	18
15	541	21	<b>46</b>	<b>704</b>	<b>19</b>
16	548	20	47	712	21
17	554	19	48	722	22
18	560	18	49	734	24
19	566	17	50	747	26
20	571	17	51	761	28
21	576	16	52	779	31
22	581	16	53	800	36
<b>23</b>	<b>586</b>	<b>16</b>	54	829	44
24	591	15	55	879	65
25	595	15	56	960	123
26	599	15			
27	604	15			
28	608	15			
29	612	15			
30	617	15			

Note: **Bold** represents CSEM around cut score (or the next higher scale score if the cut score value is not in the table).

Table 6-14 Scoring Table for English Language Arts Grade 8

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	380	61	31	635	15
1	380	61	32	640	15
2	380	61	33	644	15
3	380	61	34	649	15
4	380	61	<b>35</b>	<b>654</b>	<b>16</b>
5	380	61	36	660	16
6	400	51	37	665	16
7	436	40	38	671	16
8	460	35	39	676	17
9	478	32	40	682	17
10	494	30	41	689	18
11	507	28	42	695	18
12	519	26	43	702	19
13	529	24	<b>44</b>	<b>709</b>	<b>19</b>
14	538	23	45	717	20
15	546	22	46	726	21
16	554	21	47	735	22
17	561	20	48	745	23
18	568	19	49	756	25
19	574	18	50	768	27
20	580	18	51	782	29
21	585	17	52	799	32
22	591	17	53	820	36
<b>23</b>	<b>596</b>	<b>16</b>	54	847	44
24	601	16	55	893	62
25	606	16	56	970	117
26	611	15			
27	616	15			
28	620	15			
29	625	15			
30	630	15			

Note: **Bold** represents CSEM around cut score (or the next higher scale score if the cut score value is not in the table).

Table 6-15 Scoring Table for Mathematics Grade 3

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	360	103	<b>26</b>	<b>563</b>	<b>10</b>
1	360	103	27	567	11
2	360	103	28	571	11
3	360	103	29	575	11
4	360	103	30	579	11
5	360	103	31	584	11
6	413	52	32	588	11
7	447	31	33	593	12
8	464	25	34	599	12
9	477	21	35	605	13
10	486	18	<b>36</b>	<b>611</b>	<b>13</b>
11	494	17	37	619	14
12	501	15	38	628	16
13	507	14	39	639	19
14	513	13	40	656	24
<b>15</b>	<b>518</b>	<b>13</b>	41	686	40
16	523	12	42	760	103
17	527	12			
18	531	12			
19	536	11			
20	540	11			
21	544	11			
22	547	11			
23	551	11			
24	555	10			
25	559	10			

Note: **Bold** represents CSEM around cut score (or the next higher scale score if the cut score value is not in the table).

Table 6-16 Scoring Table for Mathematics Grade 4

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	405	116	26	594	9
1	405	116	27	597	9
2	405	116	28	600	9
3	405	116	29	603	9
4	405	116	30	607	9
5	405	116	31	610	9
6	405	116	32	613	9
7	405	116	33	616	9
8	471	51	34	620	9
9	497	33	35	623	9
10	513	25	36	627	10
11	524	21	37	631	10
12	533	18	<b>38</b>	<b>636</b>	<b>11</b>
<b>13</b>	<b>540</b>	<b>17</b>	39	641	11
14	546	15	40	646	12
15	552	14	41	652	13
16	557	13	42	660	15
17	562	12	43	670	17
18	566	12	44	684	22
19	570	11	45	709	33
20	574	11	46	800	111
21	578	10			
22	581	10			
23	584	10			
<b>24</b>	<b>588</b>	<b>10</b>			
25	591	9			

Note: **Bold** represents CSEM around cut score (or the next higher scale score if the cut score value is not in the table).

Table 6-17 Scoring Table for Mathematics Grade 5

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	430	109	26	627	9
1	430	109	27	630	9
2	430	109	28	634	9
3	430	109	29	637	9
4	430	109	30	640	9
5	430	109	31	644	9
6	430	109	32	647	10
7	498	45	33	651	10
8	526	30	34	655	10
9	542	24	<b>35</b>	<b>659</b>	<b>10</b>
10	553	20	36	663	10
11	562	17	37	668	11
12	569	16	38	673	11
<b>13</b>	<b>576</b>	<b>14</b>	39	678	12
14	581	13	40	684	12
15	586	12	41	691	14
16	591	12	42	700	16
17	595	11	43	710	18
18	599	11	44	726	23
19	603	10	45	752	34
20	607	10	46	830	97
21	610	10			
<b>22</b>	<b>614</b>	<b>10</b>			
23	617	9			
24	620	9			
25	624	9			

Note: **Bold** represents CSEM around cut score (or the next higher scale score if the cut score value is not in the table).

Table 6-18 Scoring Table for Mathematics Grade 6

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	440	98	26	640	10
1	440	98	27	644	10
2	440	98	28	648	10
3	440	98	29	651	10
4	440	98	30	655	10
5	440	98	31	659	10
6	440	98	32	663	10
7	475	63	33	667	10
8	511	33	34	671	10
9	530	24	35	675	10
10	543	21	36	680	11
11	554	19	37	685	11
12	564	18	<b>38</b>	<b>690</b>	<b>12</b>
13	572	17	39	695	12
14	580	16	40	702	13
<b>15</b>	<b>587</b>	<b>16</b>	41	709	14
16	593	15	42	718	16
17	599	14	43	730	19
18	605	13	44	749	27
19	610	13	45	788	49
20	615	12	46	870	113
21	619	12			
22	624	11			
<b>23</b>	<b>628</b>	<b>11</b>			
24	632	11			
25	636	11			

Note: **Bold** represents CSEM around cut score (or the next higher scale score if the cut score value is not in the table).

Table 6-19 Scoring Table for Mathematics Grade 7

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	450	132	26	667	10
1	450	132	27	671	10
2	450	132	28	675	10
3	450	132	29	678	10
4	450	132	30	682	10
5	450	132	31	686	10
6	450	132	32	689	10
7	529	53	33	693	10
8	557	32	34	697	11
9	574	24	35	702	11
10	586	20	36	706	11
11	595	18	37	711	12
12	603	16	<b>38</b>	<b>716</b>	<b>12</b>
<b>13</b>	<b>610</b>	<b>15</b>	39	722	13
14	616	14	40	728	14
15	621	13	41	736	15
16	627	13	42	745	17
17	632	12	43	757	20
18	636	12	44	773	25
19	640	11	45	802	38
20	645	11	46	880	105
<b>21</b>	<b>649</b>	<b>11</b>			
22	653	11			
23	656	10			
24	660	10			
25	664	10			

Note: **Bold** represents CSEM around cut score (or the next higher scale score if the cut score value is not in the table).



Table 6-20 Scoring Table for Mathematics Grade 8

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	470	125	26	679	10
1	470	125	27	683	11
2	470	125	28	687	11
3	470	125	29	691	11
4	470	125	30	695	11
5	470	125	31	699	11
6	470	125	32	703	11
7	551	45	33	708	11
8	575	31	34	712	11
9	590	24	35	717	11
10	601	21	<b>36</b>	<b>722</b>	<b>12</b>
11	610	18	37	727	12
12	617	16	38	732	12
<b>13</b>	<b>624</b>	<b>15</b>	39	738	13
14	629	14	40	744	14
15	634	13	41	752	15
16	639	12	42	760	16
17	644	12	43	771	19
18	648	11	44	786	24
19	652	11	45	812	35
20	656	11	46	890	102
21	660	11			
22	664	11			
<b>23</b>	<b>668</b>	<b>10</b>			
24	672	10			
25	675	10			

Note: **Bold** represents CSEM around cut score (or the next higher scale score if the cut score value is not in the table).

Table 6-21 Scoring Table for Science Grade 4

Raw Score	Scale Score	SEM
0	190	105
1	190	105
2	190	105
3	190	105
4	190	105
5	190	105
6	190	105
7	190	105
8	212	83
9	249	48
10	269	35
11	284	28
12	296	24
13	305	21
14	314	19
15	321	18
16	328	17
17	334	16
18	340	15
19	346	15
<b>20</b>	<b>351</b>	<b>15</b>
21	357	14
22	362	14
23	367	14
24	372	14
25	377	14
26	383	14
27	388	14
28	393	14
<b>29</b>	<b>399</b>	<b>14</b>
30	405	14
31	411	15
32	418	15
33	426	16
34	435	18
35	445	20
<b>36</b>	<b>457</b>	<b>23</b>
37	473	27
38	496	33
39	535	51
40	600	102

Note: **Bold** represents CSEM around cut score (or the next higher scale score if the cut score value is not in the table).

Table 6-22 Scoring Table for Science Grade 8

Raw Score	Scale Score	SEM
0	390	117
1	390	117
2	390	117
3	390	117
4	390	117
5	390	117
6	390	117
7	390	117
8	396	111
9	450	57
10	472	37
11	487	29
12	498	24
13	507	21
14	515	19
15	522	17
16	528	16
17	534	15
18	539	14
19	544	14
20	549	13
<b>21</b>	<b>554</b>	<b>13</b>
22	559	13
23	564	13
24	569	13
25	573	13
26	578	13
27	583	13
28	589	13
29	594	13
<b>30</b>	<b>600</b>	<b>14</b>
31	606	14
32	613	15
33	620	16
34	628	17
35	638	18
<b>36</b>	<b>649</b>	<b>20</b>
37	663	23
38	681	28
39	713	41
40	770	83

Note: **Bold** represents CSEM around cut score (or the next higher scale score if the cut score value is not in the table).

Table 6-23 Scoring Table for Social Studies Grade 4

Raw Score	Scale Score	SEM
0	200	113
1	200	113
2	200	113
3	200	113
4	200	113
5	200	113
6	200	113
7	200	113
8	244	69
9	274	40
10	290	29
11	302	24
12	312	20
13	319	18
14	326	17
15	333	16
16	339	15
17	344	14
18	349	14
19	354	14
20	359	13
<b>21</b>	<b>364</b>	<b>13</b>
22	369	13
23	374	13
24	379	13
25	384	13
26	390	13
27	395	14
<b>28</b>	<b>401</b>	<b>14</b>
29	407	14
30	413	15
31	420	15
32	428	16
<b>33</b>	<b>436</b>	<b>17</b>
34	446	19
35	459	22
36	476	27
37	505	39
38	570	90

Note: **Bold** represents CSEM around cut score (or the next higher scale score if the cut score value is not in the table).

Table 6-24 Scoring Table for Social Studies Grade 8

Raw Score	Scale Score	SEM
0	420	99
1	420	99
2	420	99
3	420	99
4	420	99
5	420	99
6	420	99
7	420	99
8	454	65
9	480	39
10	496	28
11	507	23
12	516	20
13	523	18
14	530	16
15	536	15
16	541	15
17	546	14
18	551	14
19	556	13
20	560	13
<b>21</b>	<b>565</b>	<b>12</b>
22	569	12
23	574	12
24	578	12
25	582	12
26	587	12
27	591	12
28	596	12
<b>29</b>	<b>600</b>	<b>12</b>
30	605	13
31	611	13
32	616	14
33	622	14
34	629	15
35	637	16
<b>36</b>	<b>647</b>	<b>18</b>
37	659	21
38	675	25
39	701	36
40	780	101

Note: **Bold** represents CSEM around cut score (or the next higher scale score if the cut score value is not in the table).

Table 6-25 Scoring Table for Social Studies Grade 10

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	490	129	<b>26</b>	<b>673</b>	<b>12</b>
1	490	129	27	677	12
2	490	129	28	681	12
3	490	129	29	685	12
4	490	129	30	688	12
5	490	129	31	692	12
6	490	129	32	696	12
7	490	129	33	700	12
8	490	129	<b>34</b>	<b>705</b>	<b>12</b>
9	490	129	35	709	12
10	490	129	36	713	12
11	529	90	37	718	12
12	567	52	38	722	13
13	588	37	39	727	13
14	601	30	40	732	13
15	612	25	41	738	14
16	621	22	<b>42</b>	<b>744</b>	<b>14</b>
17	628	20	43	750	15
18	635	18	44	758	16
19	641	17	45	766	17
20	646	16	46	776	19
21	651	15	47	788	21
22	656	14	48	805	26
23	660	13	49	833	37
24	665	13	50	890	78
25	669	12			

Note: **Bold** represents CSEM around cut score (or the next higher scale score if the cut score value is not in the table).

Table 6-26 The Number and Percentage of Students at LOSS and HOSS

Content	Grade	LOSS	N	Percent	HOSS	N	Percent
ELA	3	330	7	.01	900	0	.00
	4	340	12	.02	930	0	.00
	5	350	19	.03	940	0	.00
	6	360	5	.01	950	0	.00
	7	370	30	.05	960	21	.03
	8	380	21	.03	970	12	.02
Math	3	360	518	.81	760	151	.24
	4	405	2028	3.14	800	139	.22
	5	430	1181	1.87	830	32	.05
	6	440	1356	2.16	870	46	.07
	7	450	1953	3.09	880	42	.07
	8	470	2905	4.67	890	26	.04
Science	4	190	133	.21	600	357	.55
	8	390	365	.59	770	295	.47
Social Studies	4	200	316	.49	570	658	1.02
	8	420	516	.83	780	871	1.40
	10	490	624	.98	890	219	.34

Figure 6-1 Anchor Set TCCs: ELA Grade 3

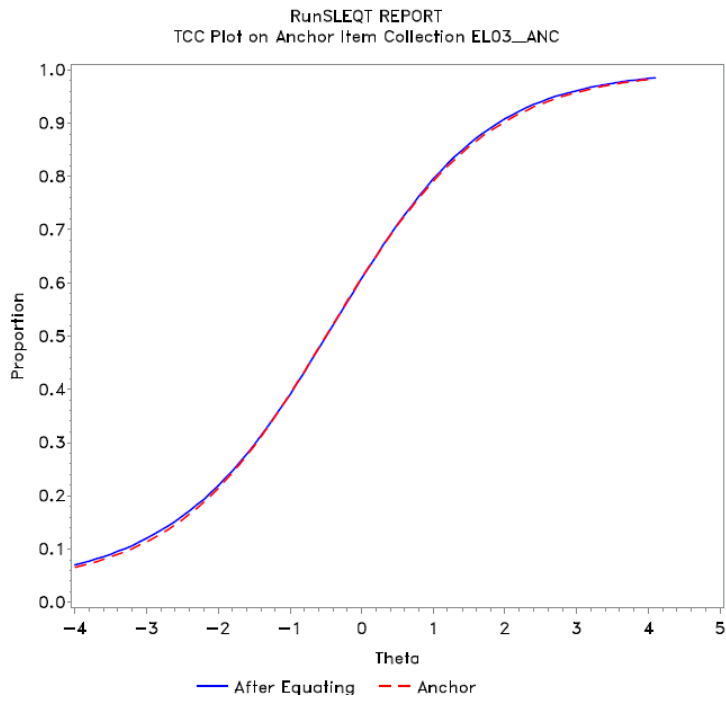


Figure 6-2 Anchor Set TCCs: ELA Grade 4

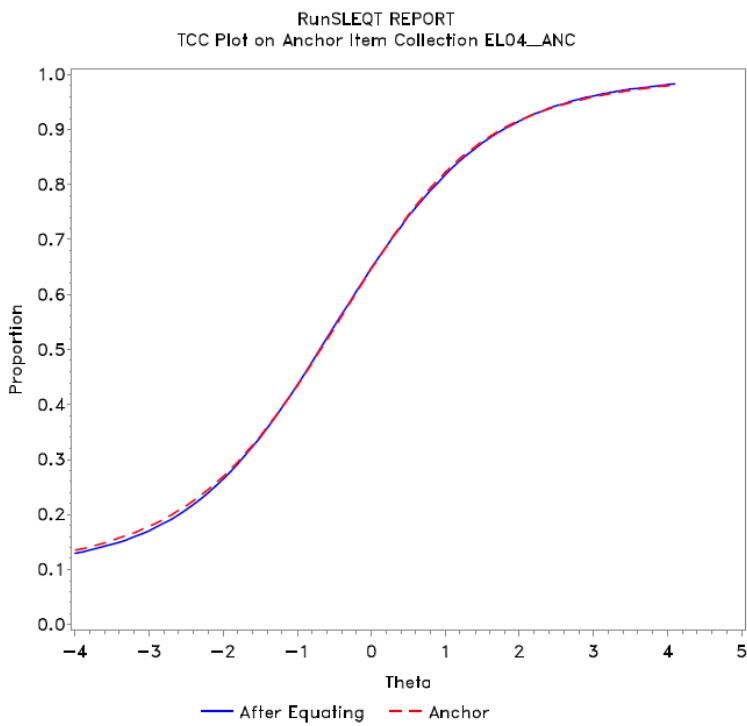




Figure 6-3 Anchor Set TCCs: ELA Grade 5

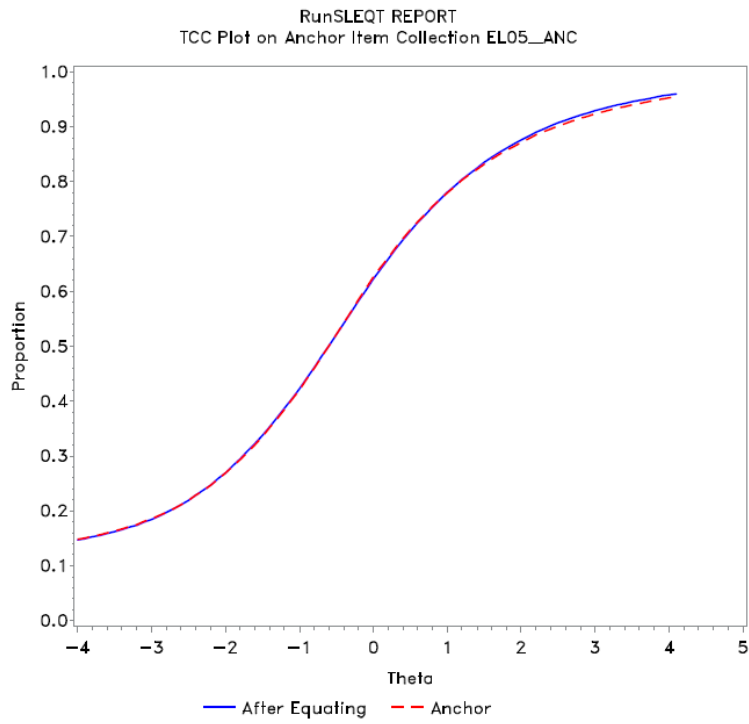


Figure 6-4 Anchor Set TCCs: ELA Grade 6

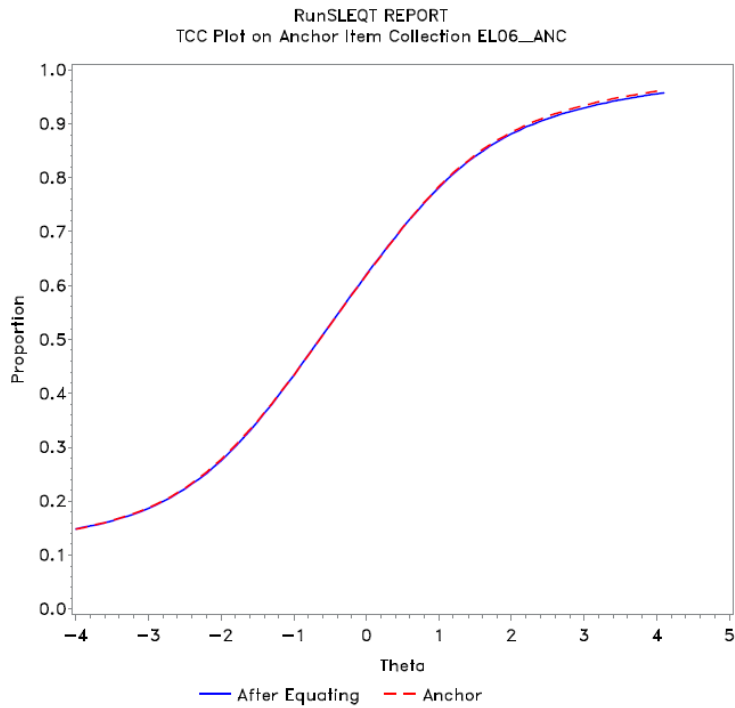


Figure 6-5 Anchor Set TCCs: ELA Grade 7

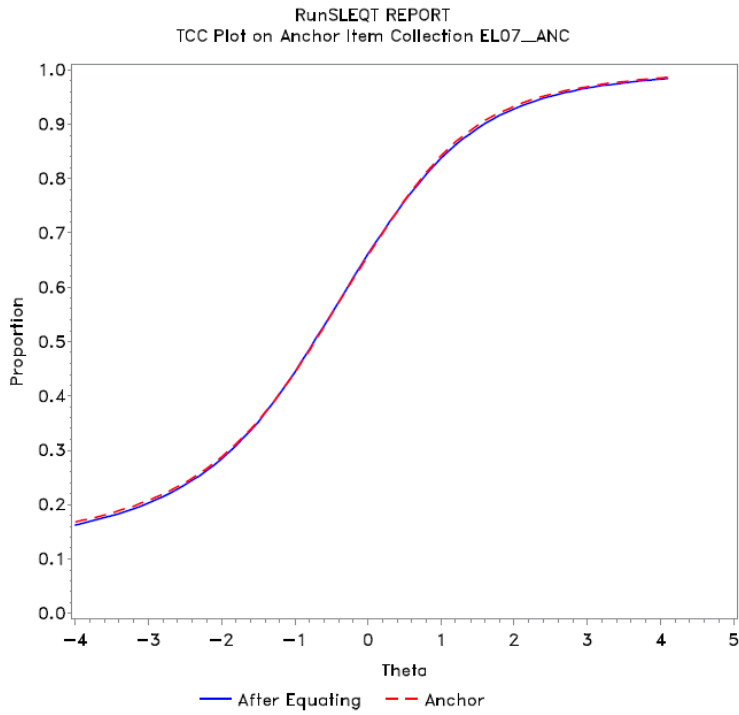


Figure 6-6 Anchor Set TCCs: ELA Grade 8

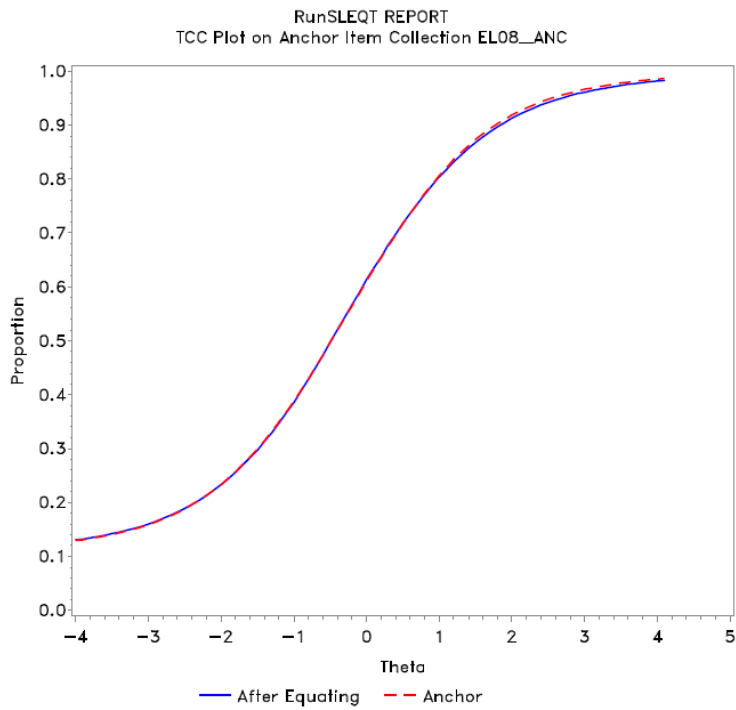


Figure 6-7 Anchor Set TCCs: Mathematics Grade 3

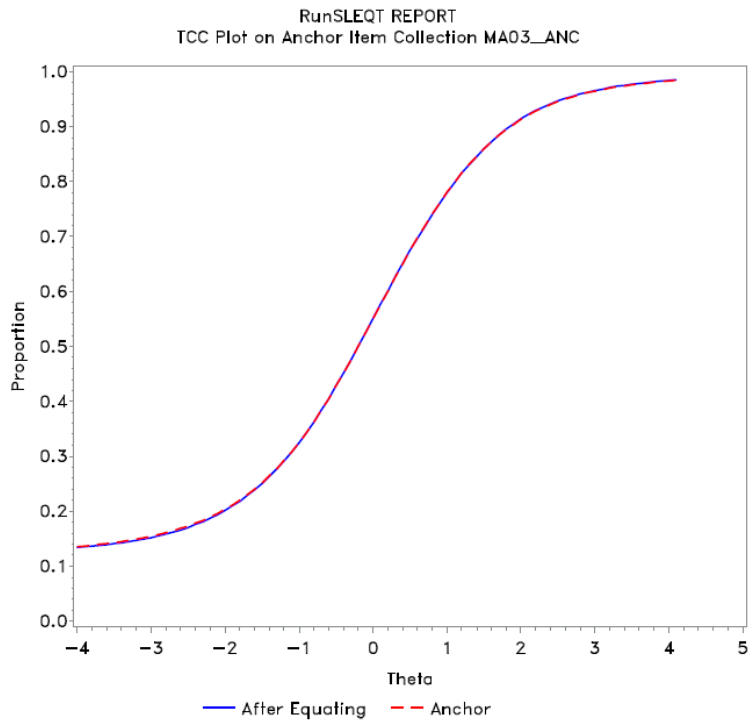


Figure 6-8 Anchor Set TCCs: Mathematics Grade 4

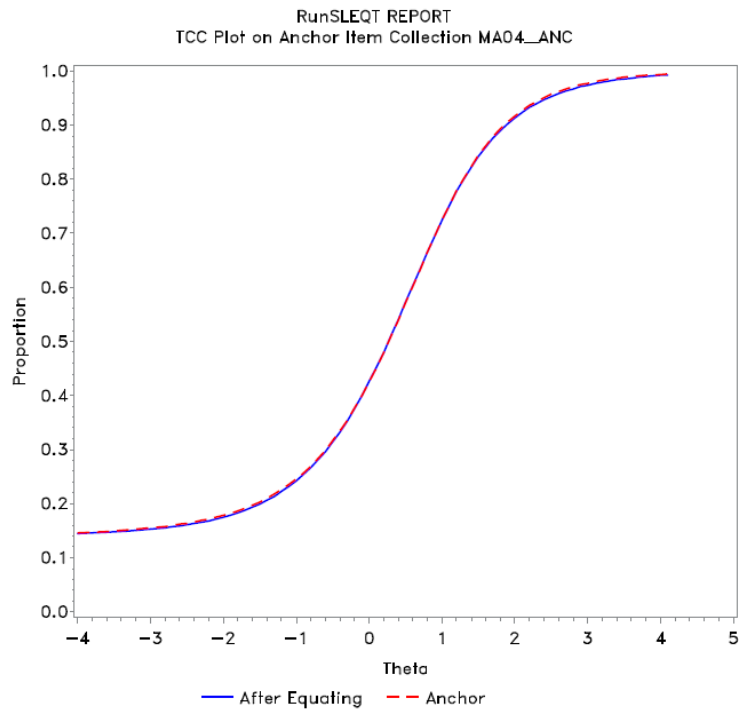


Figure 6-9 Anchor Set TCCs: Mathematics Grade 5

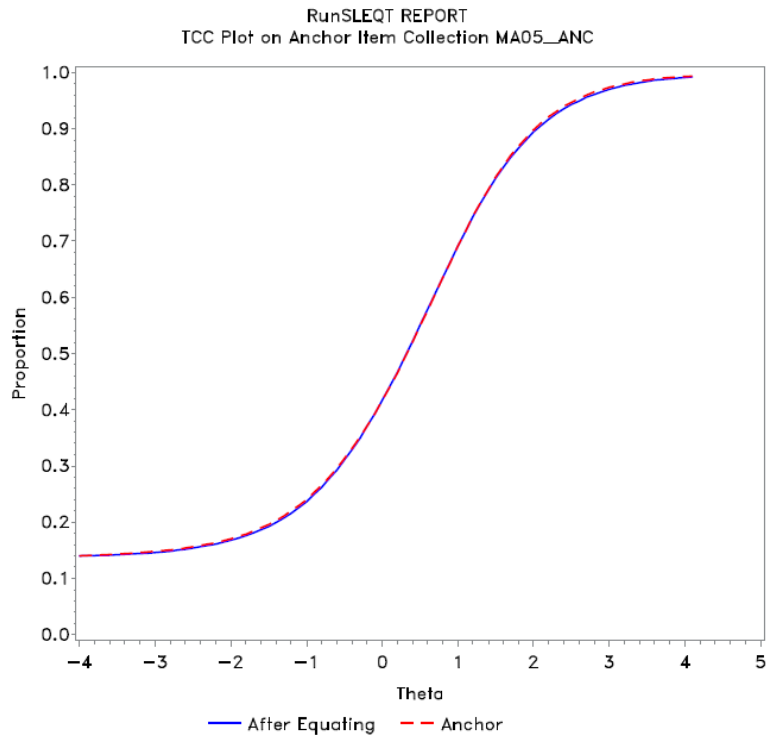


Figure 6-10 Anchor Set TCCs: Mathematics Grade 6

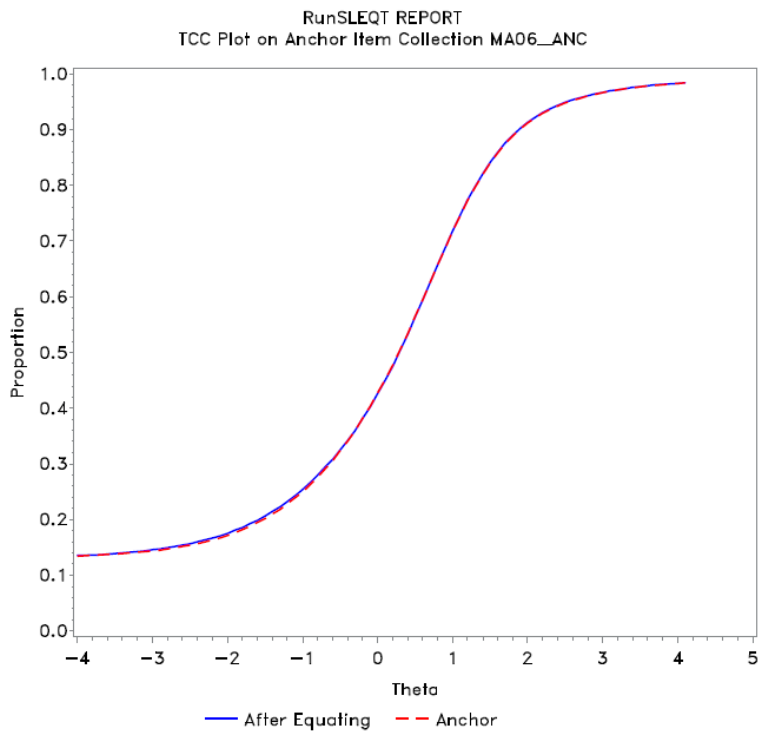


Figure 6-11 Anchor Set TCCs: Mathematics Grade 7

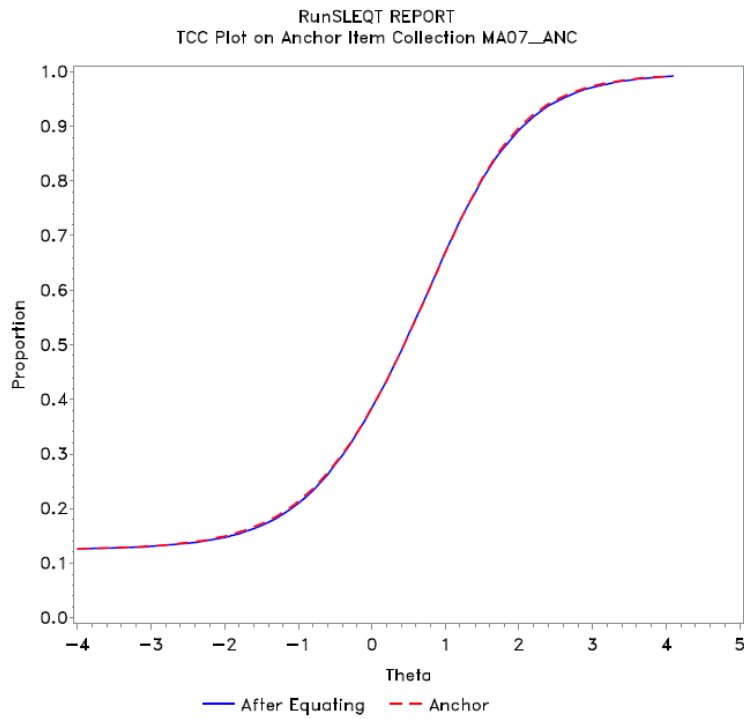


Figure 6-12 Anchor Set TCCs: Mathematics Grade 8

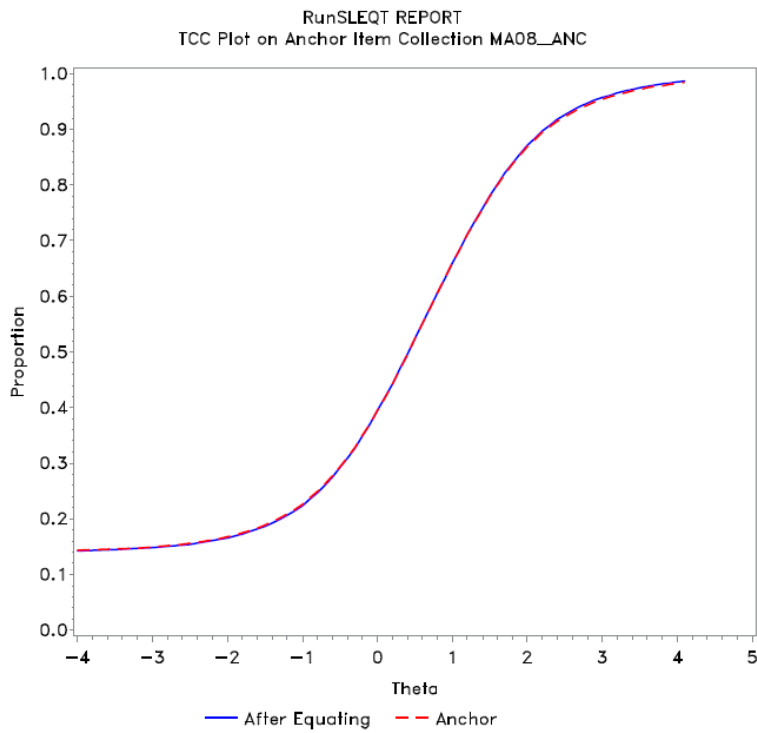


Figure 6-13 Anchor Set TCCs: Science Grade 4

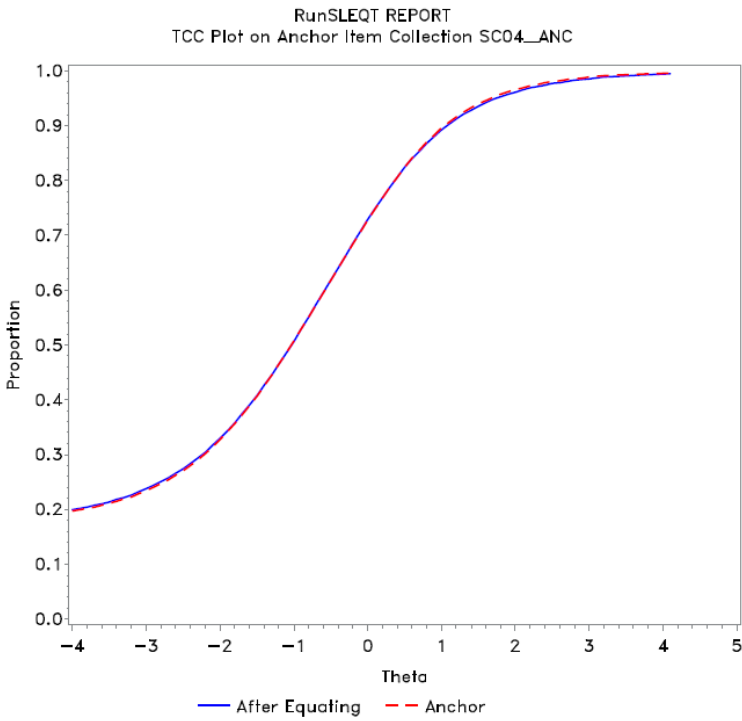


Figure 6-14 Anchor Set TCCs: Science Grade 8

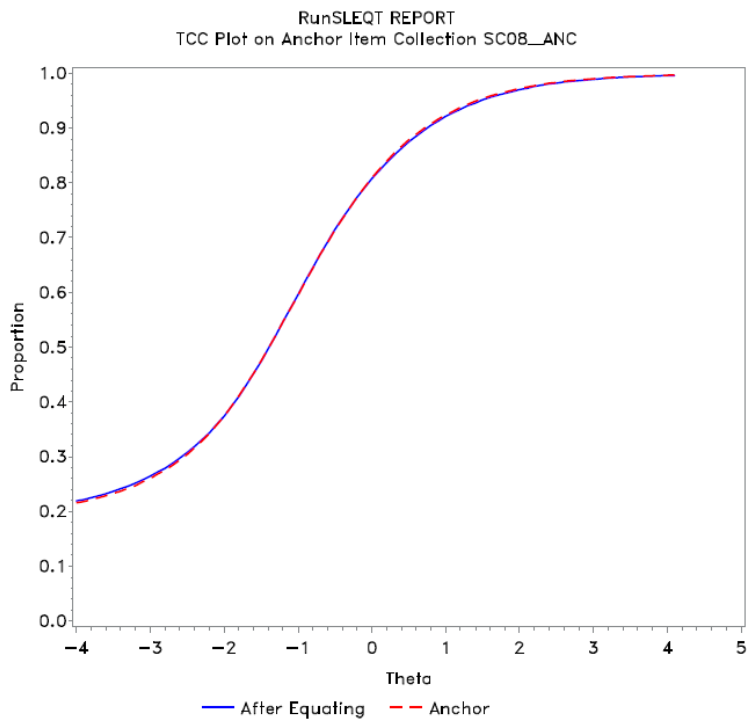


Figure 6-15 Anchor Set TCCs: Social Studies Grade 4

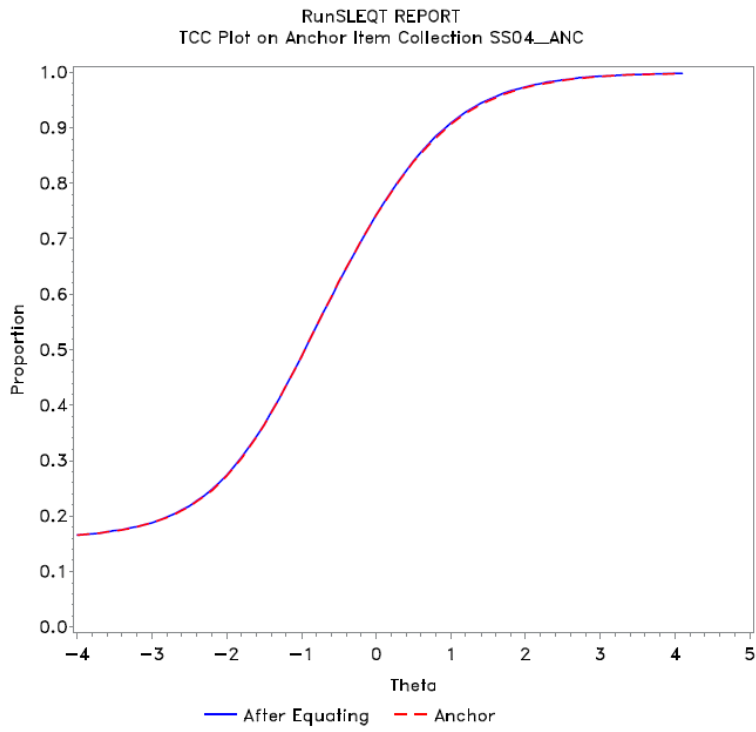


Figure 6-16 Anchor Set TCCs: Social Studies Grade 8

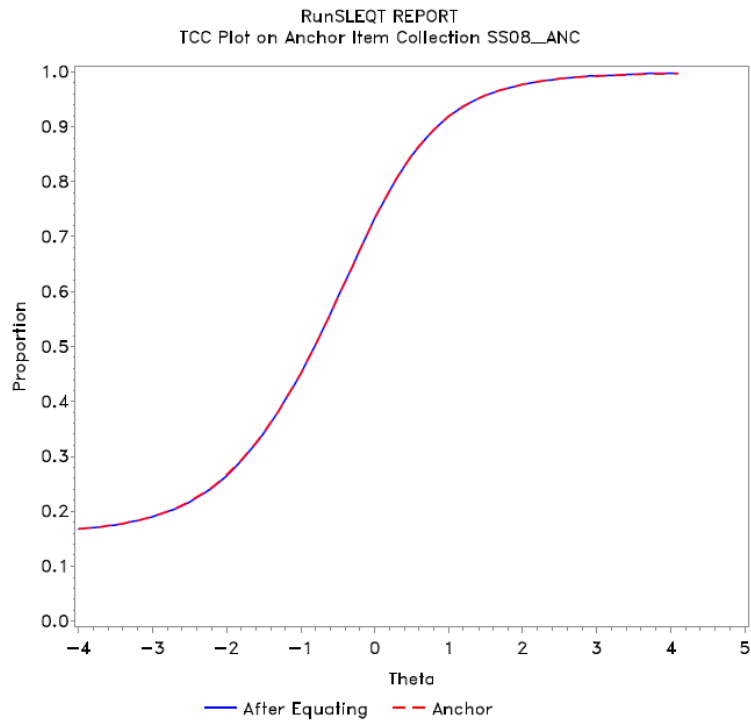


Figure 6-17 Anchor Set TCCs: Social Studies Grade 10

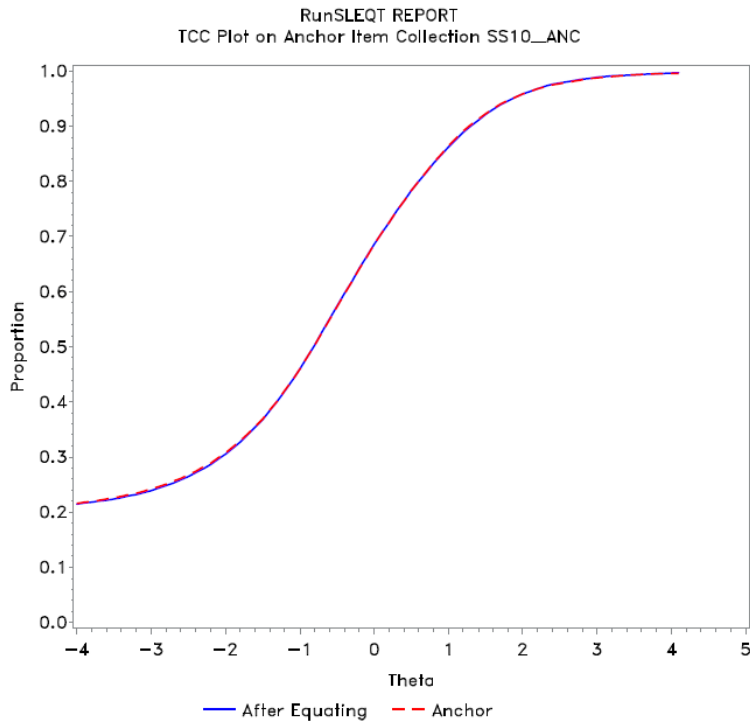


Figure 6-18 Item Characteristic Curves for the Flagged Social Studies Grade 4 Anchor

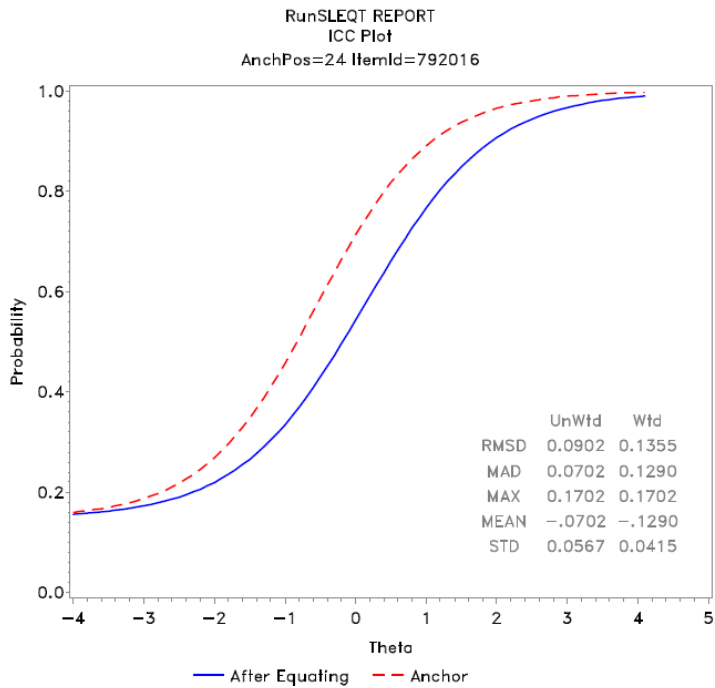




Figure 6-19 English Language Arts Test Characteristic Curves

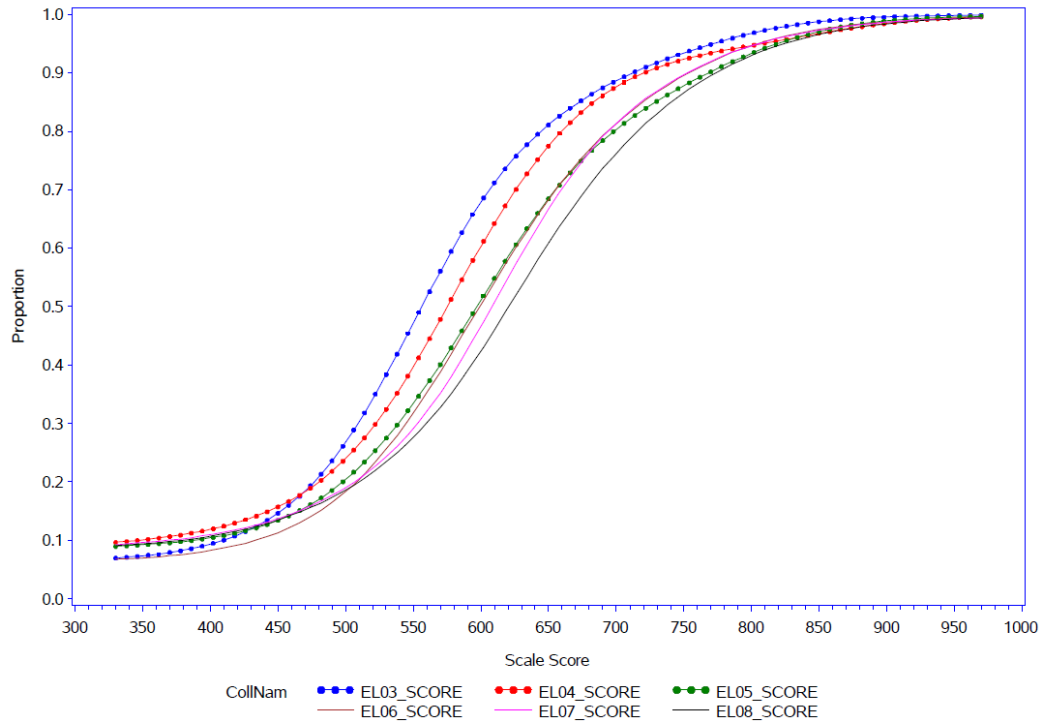


Figure 6-20 English Language Arts Standard Error Curves

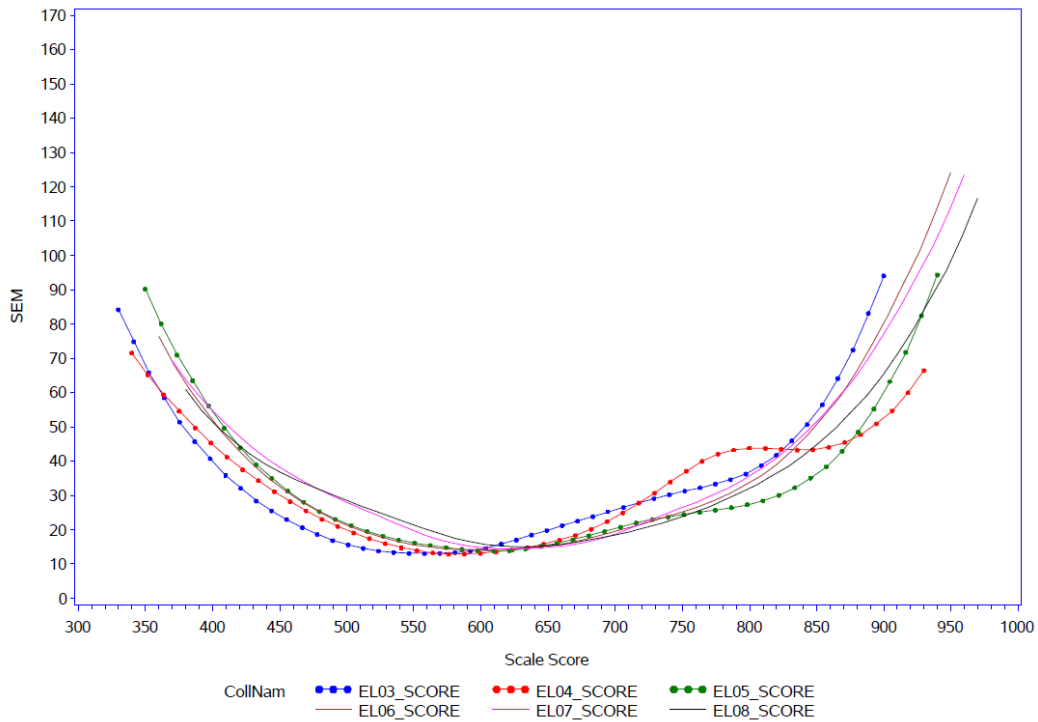


Figure 6-21 English Language Arts Growth at Quartiles

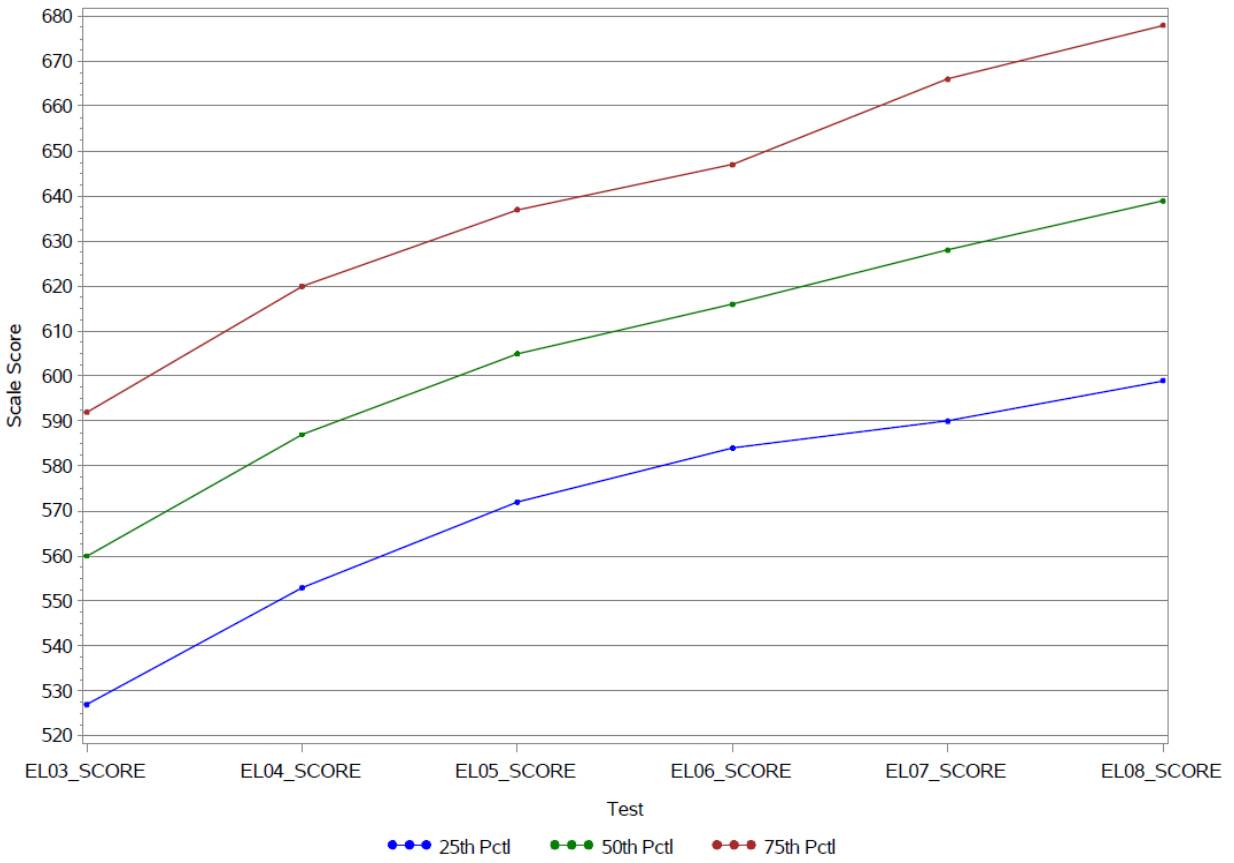


Figure 6-22 Mathematics Test Characteristic Curves

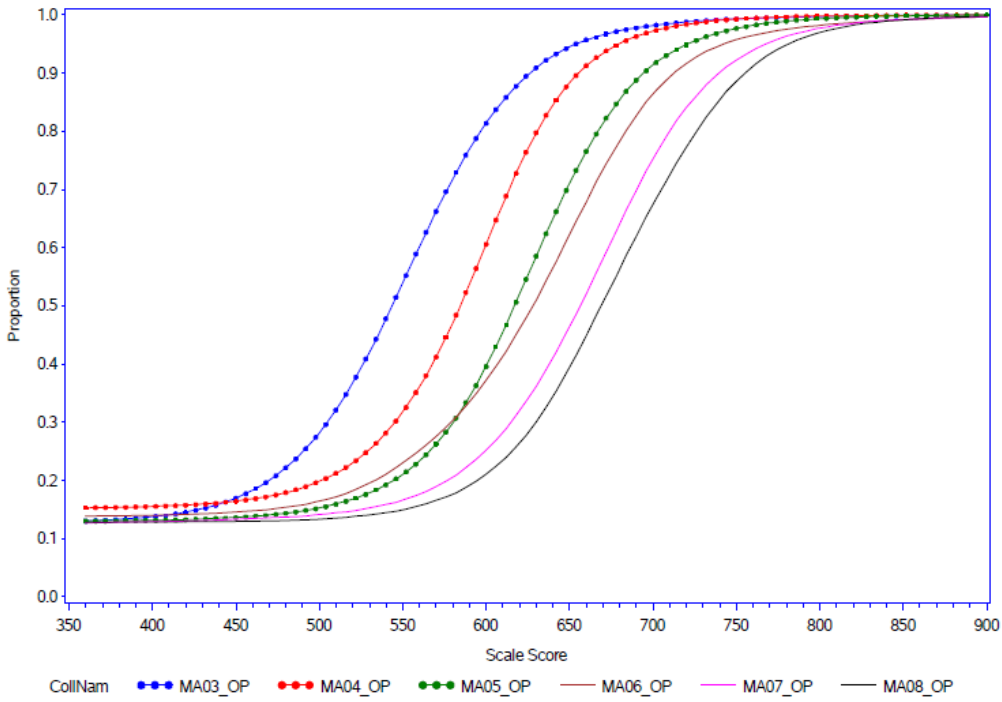


Figure 6-23 Mathematics Standard Error Curves

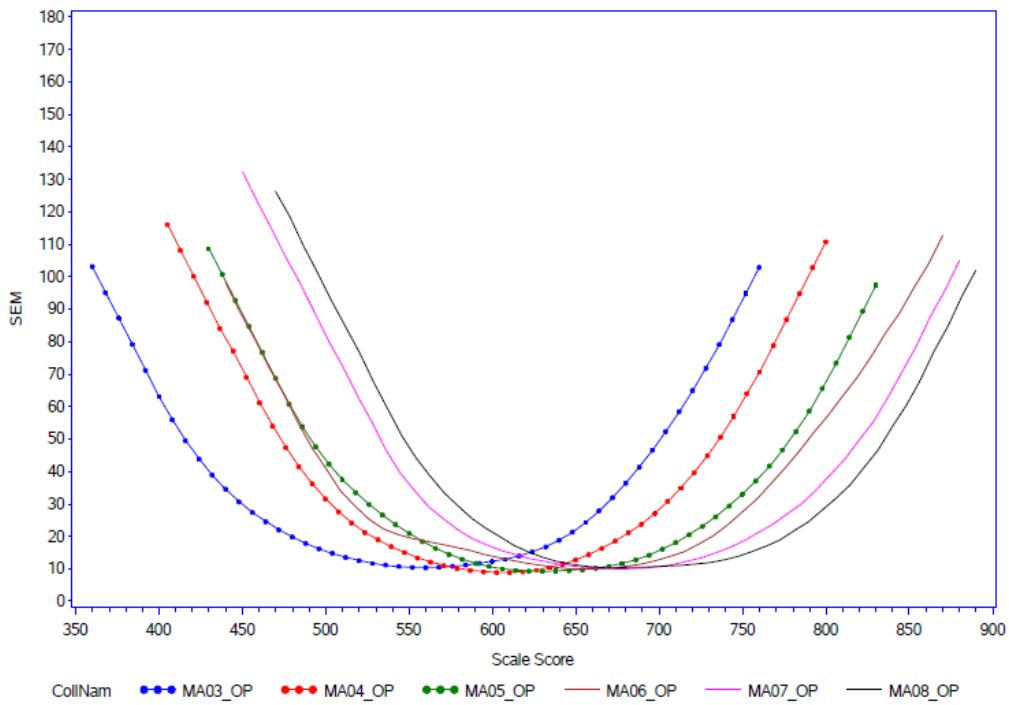


Figure 6-24 Mathematics Growth at Quartiles

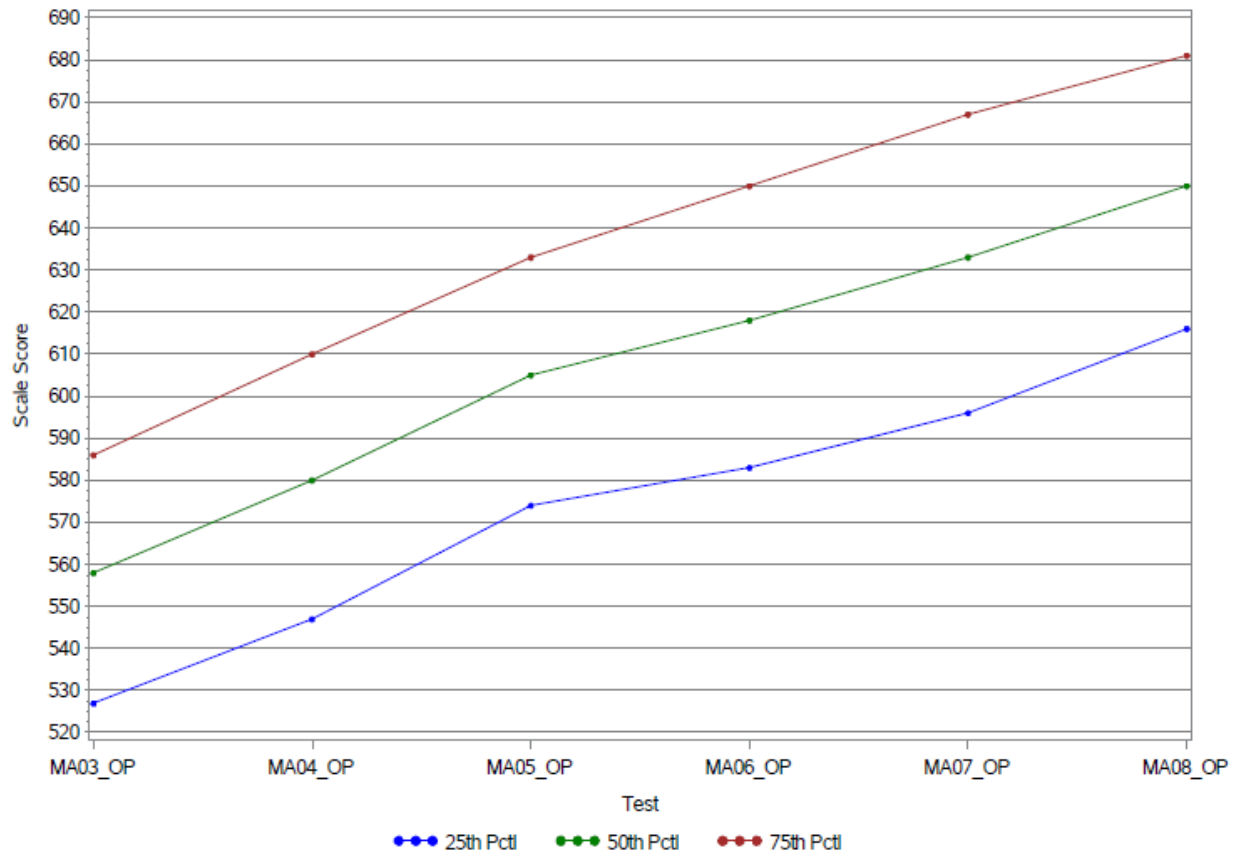


Figure 6-25 Science Test Characteristic Curves

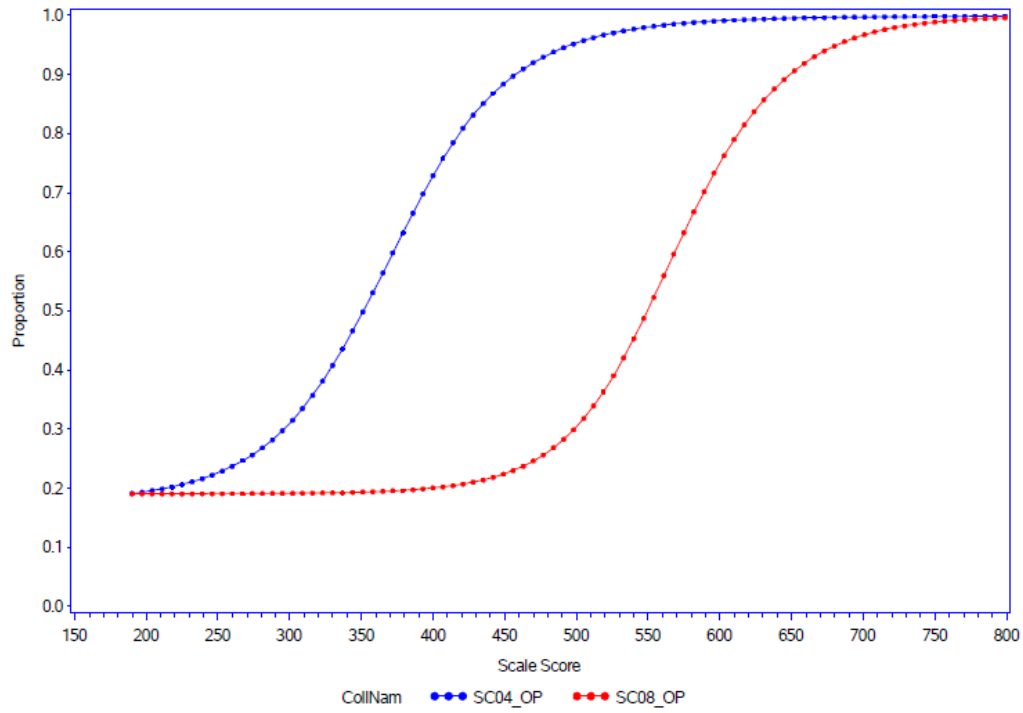


Figure 6-26 Science Standard Error Curves

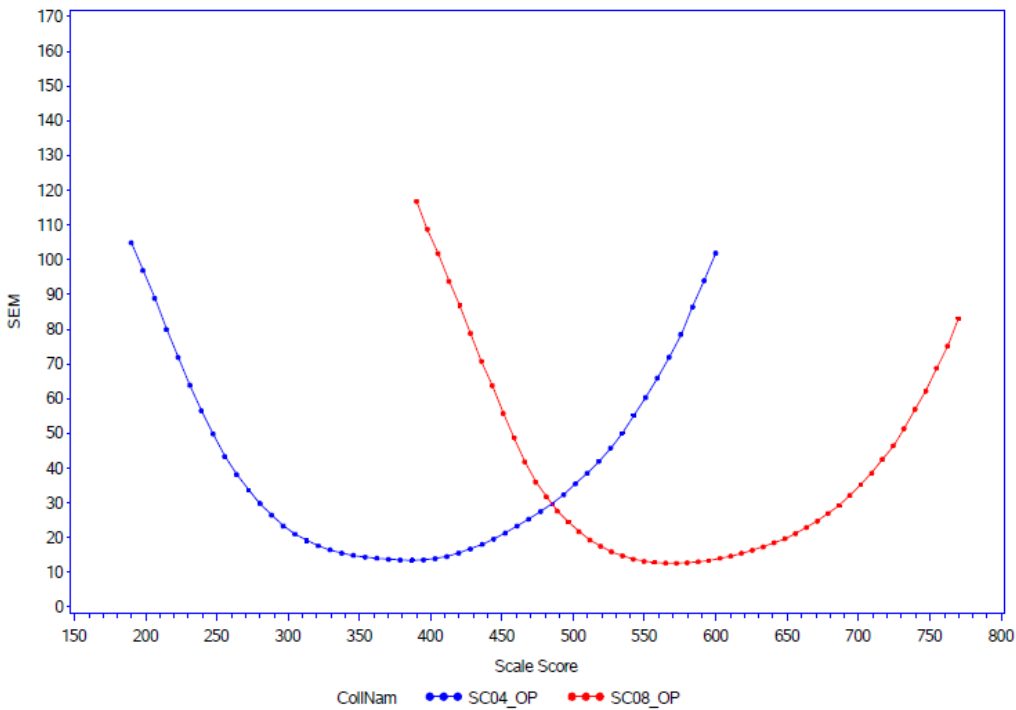


Figure 6-27 Science Growth at Quartiles

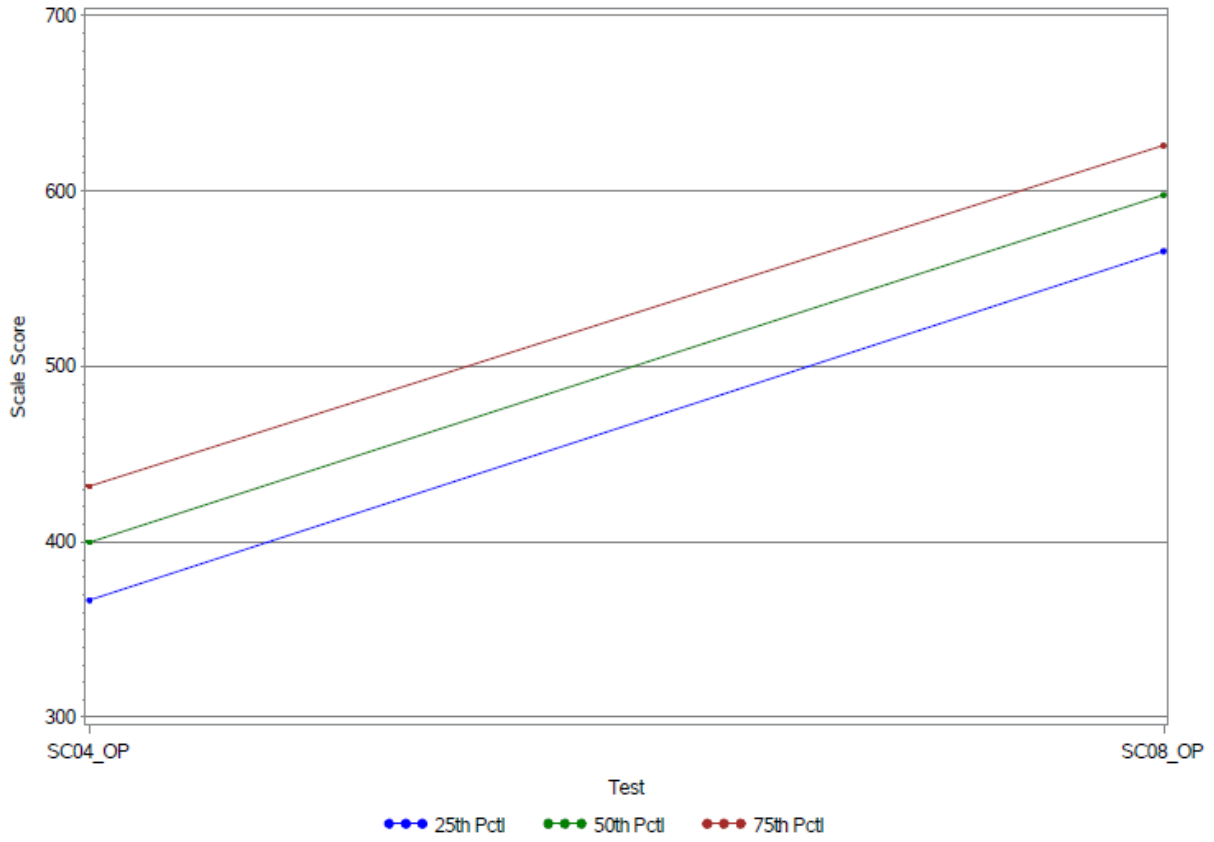


Figure 6-28 Social Studies Test Characteristic Curves

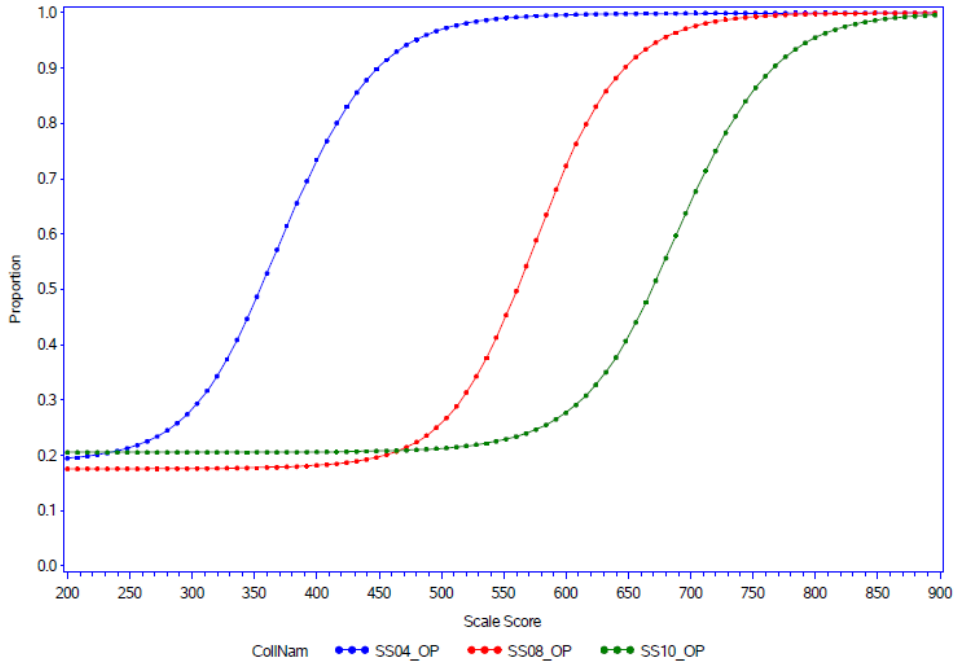


Figure 6-29 Social Studies Standard Error Curves

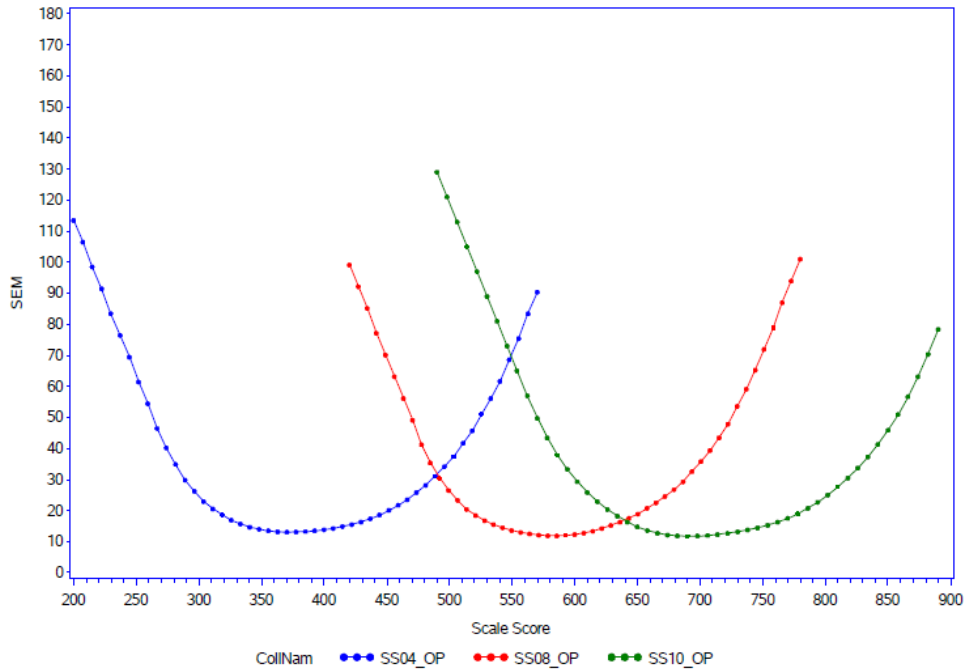
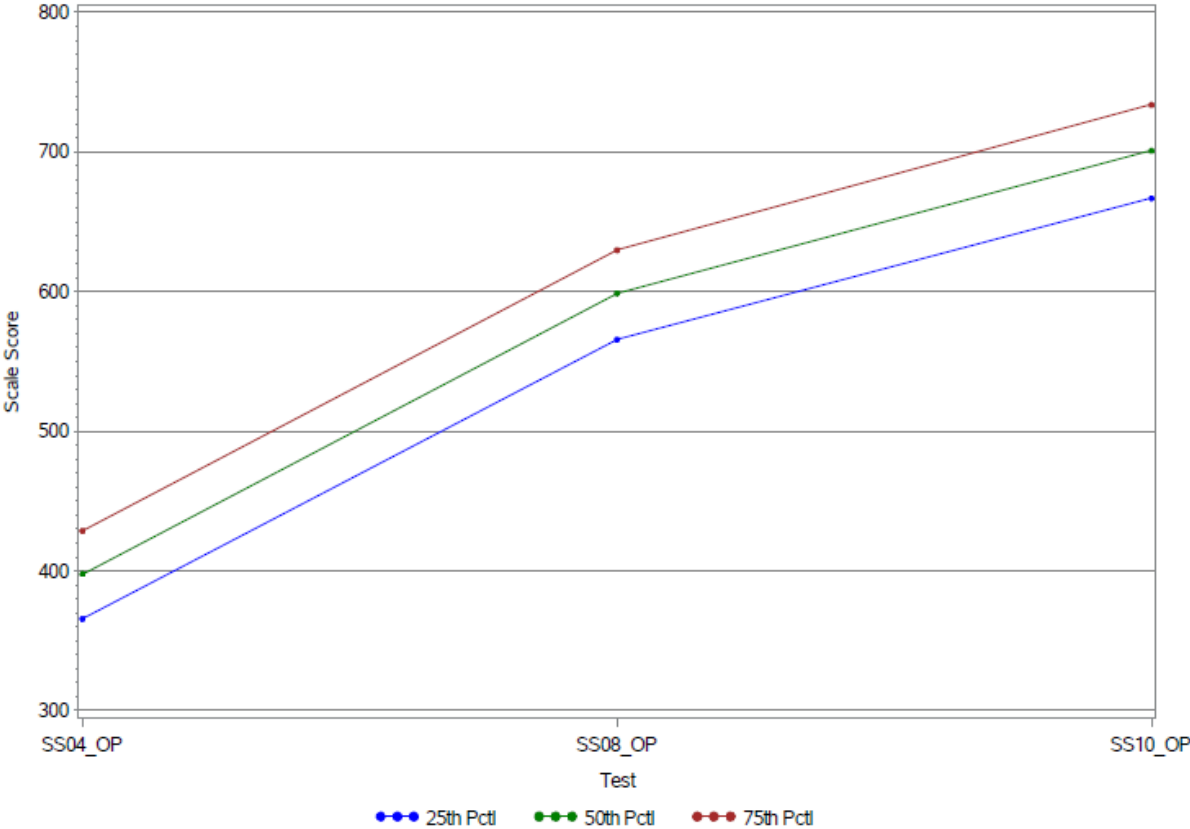


Figure 6-30 Social Studies Growth at Quartiles





## Part 7: Standard Setting

---

In this chapter, we briefly describe the Wisconsin Forward Exam standard setting, and we present the cut scores established and the performance level descriptors derived from the standard setting. The information in this chapter comes from the *Wisconsin Standard Setting 2016 Final Technical Report* submitted to DPI and available at <http://dpi.wi.gov/assessment/forward/resources>.

### 7.1 Background Information

Several changes were made to Wisconsin's statewide tests, especially for English Language Arts (ELA) and Mathematics, in recent years. In the 2014–15 school year, the Wisconsin Badger Exam measured students' abilities in ELA and Mathematics using assessments developed by the Smarter Balanced Assessment Consortium (SBAC). Cut scores for the Wisconsin Badger Exam were taken from the national SBAC standard setting, conducted in 2014. For Science and Social Studies, the Wisconsin Knowledge and Concepts Examination (WKCE) was administered. Cut scores for the WKCE were established in 2005.

In the 2015–16 school year, DPI consolidated the Wisconsin Badger Exam and the WKCE into a unified program, the Wisconsin Forward Exam. At the inception of the Wisconsin Forward Exam, DPI indicated that they would no longer use SBAC items or test scales for ELA and Mathematics and that new test scales would be established for the Wisconsin Forward Exam. New test scales were established for all four content areas using data from the Spring 2016 administration of the Wisconsin Forward Exam.

On June 14–17, 2016, DPI and DRC conducted the Wisconsin Forward Exam Standard Setting for grades 3–8 in ELA and Mathematics, grades 4 and 8 in Science, and grades 4, 8, and 10 in Social Studies. The purpose of the standard setting was to develop performance standards for the Wisconsin Forward Exam, including the development of *cut scores* that divide students into four performance levels: *Below Basic*, *Basic*, *Proficient*, and *Advanced*. During this benchmarked standard setting, DPI developed cut scores on the Wisconsin Forward Exam that reflected these content-based expectations on the tests, as informed by test data from well-respected measures of student achievement.

A total of 59 Wisconsin educators and stakeholders worked individually and in committees to recommend performance standards associated with four performance levels: *Below Basic*, *Basic*, *Proficient*, and *Advanced*. This process yielded performance standards for the 17 tests of the Wisconsin Forward Exam program. The performance standards were approved by the Superintendent of Public Instruction in July 2016.

The process of the standard setting adhered to the AERA, APA, & NCME (2014) Standards 5.21 and 5.22, which state the following:

**Standard 5.21** When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be documented clearly. (107)

**Standard 5.22** When cut scores defining pass-fail or proficiency levels are based on direct judgments about the adequacy of item or test performances, the judgmental process should be designed so that the participants providing the judgments can bring their knowledge and experience to bear in a reasonable way. (108)

## **7.2 Standard Setting Methodology**

Prior to the standard setting workshop, DPI worked in collaboration with DRC and its other technical advisors to select the methodology to be used at the standard setting. In recognition of its use in Wisconsin and widespread use across the country, DPI selected the Bookmark Standard Setting Procedure (BSSP) for the Wisconsin Forward Exam. The BSSP was well suited for standard setting for these assessments because (a) the tests are composed of both multiple-choice and constructed-response items, (b) the items are scaled and can be mapped using item mapping techniques, and (c) the BSSP allows participants to focus on the knowledge, skills, and abilities expected of students in each performance level. The BSSP has been well documented in the standard setting literature. Developed in 1996, the BSSP has been implemented in over half of the states in the United States and abroad by DRC and by other major testing firms, making it the most widely used standard setting procedure in K–12 education (Karantonis & Sireci, 2006; Cizek & Bunch, 2007).

## **7.3 Performance Level Descriptors**

In terms of the validity of the Wisconsin Forward Exam scores, it is essential to understand that descriptors and cut scores are established in a collaborative and participatory process. The descriptors clearly establish, in plain language, the proper frame of reference for understanding how to interpret test scores, particularly cut scores. Performance level descriptors (PLDs) summarize the knowledge, skills, and abilities expected of students in each performance level. DPI provided policy PLDs for the Wisconsin Forward Exam. These brief descriptors, shown in Table 7-1, described DPI’s vision for each performance level. At the standard setting, Wisconsin used the policy PLDs in conjunction with the content standards to consider the content-based expectations for students in each performance level on each test in the Wisconsin Forward Exam program.

## **7.4 Cut Scores**

Table 7-2 shows the cut scores for all grades and content areas. The cut scores reflect the content-based expectations for students and policy-based decisions (i.e., the impact of the cut scores on Wisconsin students as shown through the impact data). The cut scores established after Spring 2016 test administration remained unchanged for the Spring 2017 assessments.

## 7.5 Summary

Part 7 presented a brief overview of the standard setting process used for establishing the Wisconsin Forward Exam cut scores after the Spring 2016 test administration. These procedures are addressed in more detail in the *Wisconsin Standard Setting 2016 Final Technical Report*. The standard settings undertaken by DPI and facilitated by DRC support Standards 5.21 and 5.22 from the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014).

Table 7-1 Policy Performance Level Descriptors for the Wisconsin Forward Exam

Level	Performance Level Descriptor
<i>Below Basic</i>	Student demonstrates <b>minimal understanding</b> of and ability to apply the knowledge and skills for his or her grade level that are associated with college content-readiness.
<i>Basic</i>	Student demonstrates <b>partial understanding</b> of and ability to apply the knowledge and skills for his or her grade level that are associated with college content-readiness.
<i>Proficient</i>	Student demonstrates <b>adequate understanding</b> of and ability to apply the knowledge and skills for his or her grade level that are associated with college content-readiness.
<i>Advanced</i>	Student demonstrates <b>thorough understanding</b> of and ability to apply the knowledge and skills for his or her grade level that are associated with college content-readiness.

Table 7-2 Wisconsin Forward Exam Cut Scores

Content	Grade	Basic	Proficient	Advanced
ELA	3	522	570	624
	4	546	592	650
	5	564	610	670
	6	572	622	671
	7	585	638	697
	8	592	652	708
Mathematics	3	517	560	611
	4	536	588	633
	5	574	611	658
	6	582	626	688
	7	606	647	712
	8	620	667	718
Science	4	348	399	447
	8	552	600	645
Social Studies	4	363	396	436
	8	563	599	640
	10	670	703	741

## Part 8: Test Results

---

Part 8 presents a classical item analysis and summary of student results for the Spring 2017 Wisconsin Forward Exam. The summary results are presented for all Wisconsin students and cover four types of scores: raw scores; scale scores; performance level results; and scores based on each of the content standards within each content area, which are called standard performance index (SPI) scores. Combined, the classical item analysis and the four forms of scores offer the reader several vantage points from which to understand and evaluate the Wisconsin Forward Exam testing program. The AERA, APA, & NCME (2014) standards addressed in Part 8 include 1.8, 4.14, 5.1, 5.21, 7.0, and 7.1. These standards are cited below:

**Standard 1.8** The composition of any sample of test takers from which validity evidence is obtained should be described in as much detail as is practical and permissible, including major relevant socio-demographic and developmental characteristics. (25)

**Standard 4.14** For a test that has a time limit, test development research should examine the degree to which scores include a speed component and should evaluate the appropriateness of that component, given the domain the test is designed to measure. (90)

**Standard 5.1** Test users should be provided with clear explanations of the characteristics, meaning, and intended interpretation of scale scores, as well as their limitations. (102)

**Standard 5.21** When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be documented clearly. (107)

**Standard 7.0** Information relating to tests should be clearly documented so that those who use tests can make informed decisions regarding which test to use for a specific purpose, how to administer the chosen test, and how to interpret test scores. (125)

**Standard 7.1** The rationale for a test, recommended uses of the test, support for such uses, and information that assists in score interpretation should be documented. When particular misuses of a test can be reasonably anticipated, cautions against such misuses should be specified. (125)

### 8.1 Classical Item Analysis: Item Level Statistics

Three statistics are frequently used in item analysis at the item level: the proportion correct ( $p$ -value), the item-total correlation coefficient, and the omit rate for the item.

The  $p$ -value is an indication of the difficulty of an item. The  $p$ -value for an MC item represents the proportion of students who answered the item correctly. If all students answered a given MC item correctly, its  $p$ -value would be 1.0. If only 30% of students answered the question correctly, the  $p$ -value would be 0.30. The lower the  $p$ -value is, the more difficult the item. Item  $p$ -value is a good indication of difficulty, as it takes student performance into account

and it makes comparing items in terms of a common statistic very simple. A test made up of items well distributed across the range of item difficulty levels is desirable because it supports the assessment of students at all ability levels.

The  $p$ -value for a CR item represents the mean proportion of possible raw score points that students actually obtained for the item. A  $p$ -value of 0.33 for a given CR item would indicate that, on average, students obtained one-third of the possible points for the item. If a  $p$ -value were 0.75, this would indicate a much easier item where, on average, students scored 75% of the maximum possible points for the item. As such, the  $p$ -value indicates difficulty for CR items as well, with lower  $p$ -values indicating more difficult items.

The item-total correlation indicates the extent to which individual test items provide reliable measurement of the construct being measured by the total test, and it is an index of the item's ability to discriminate between high-ability and low-ability students. For dichotomously scored MC items, the item-total correlations are computed as point-biserial correlations between the score on the item and the score on the remaining items in the test. For CR items, the item-total correlations are computed as Pearson product-moment correlations between the score on the item and the score on the remaining items in the test.<sup>1</sup> The item-total correlation coefficients can range from -1.0 to +1.0. A large positive value (such as 0.40) indicates a strong relationship between a score on an individual item and the total score, with students who earn high scores on the total test tending to score higher on the item than students with low scores on the total test. A low positive value (such as 0.10) indicates a weak relationship between scores on the item and the total score, while a negative value indicates that students who do well on the total test tend to score lower on the item than students who do poorly on the total test.

For MC items, the point-biserial correlation between each distractor and the total score was also calculated. In most cases, items will have negative correlations for each distractor and the total score. However, a weak positive correlation for a distractor does not necessarily mean that the item is defective, provided that the distractor correlation is substantially smaller than the item-total correlation for the correct response. In some cases, it may simply mean that the particular distractor is attractive to moderate-ability students and unattractive to low-ability students.

The omit rate is also computed for each item, reflecting the percentage of students who did not respond to the item. A high omit rate can indicate an especially difficult item or, if located near the end of the test, it can indicate what is referred to as a "speeded" test, where students have insufficient time to respond to all items.

For the Spring 2017 Wisconsin Forward Exam, items were flagged for further investigation according to the following rules:

- The  $p$ -value was less than 0.20. Such a  $p$ -value indicates a difficult item, where fewer than 20% of students obtained the correct answer.

---

<sup>1</sup> For both the point-biserial and the Pearson correlations, the studied item is excluded from the computation of the total score so as to not artificially inflate the correlation statistic. This effect would be most noticeable for CR items worth several points.

- The item-total correlation was less than 0.15 for the correct answer. A low value may indicate that the item is not providing a high degree of discrimination between high-ability and low-ability students, and, in addition, it may be an indication that the correct answer is in question.
- A distractor had a positive correlation with the total test score.
- The omit rate was greater than 5%.

Flagging an item for investigation is just one aspect of a complete evaluation of an item, and flagged items are not necessarily defective. It is desirable to include a small number of items with very high  $p$ -values (especially easy items) or very low  $p$ -values (especially difficult items) in order to provide more reliable measurement at the extreme high and low levels of ability and to fully represent the range of difficulty for particular content standards. In this case, the flagging of  $p$ -values is a useful way of verifying that the number of extremely easy or difficult items is relatively small and consistent with the purposes of the test. Thus, flagged items do not necessarily indicate a challenge to test validity, because items have been found to be appropriate during item reviews.

Omit rates may reflect a number of different properties, and an item that is omitted by more than 5% of the students (the Wisconsin Forward Exam flagging criterion) is not necessarily problematic. Omit rates are typically higher for CR items than for MC items because students who are fairly certain they do not know the answer may be inclined to simply skip the item altogether rather than taking the time to form a response. Items with high omit rates are referred to content specialists for further review to ensure there is no unintended ambiguity in the items. If these flagged items are judged to be clear and provide a valid measurement of the intended knowledge, skill, or ability, then they are retained on the test.

Items flagged for a low item-total correlation or for a positive distractor-total test correlation are more troublesome because these statistics show the relationship of each option to the construct being measured. In determining whether these items should be retained or removed from scoring, it is important to consider the relative magnitude of the correlation between the correct response and the total score and between the distractor and the total score. In most cases, removing an item with a modest item-total correlation and negative correlations for all of the distractors will actually lower the reliability of the total test, so it is generally preferable to retain these items. The same is true of an item with a small positive correlation for one of the distractors and a much larger positive correlation for the correct response. However, an item that exhibits a low correlation for the correct response in combination with a positive correlation for one or more distractors is likely to degrade the measurement and lower the reliability of the test. Such items should be removed from scoring.

Overall, 56 operational items were flagged on the Spring 2017 Wisconsin Forward Exam operational tests as meeting the investigational criteria bulleted above.

Table 8-A shows the number of scored items in the Spring 2017 Wisconsin Forward Exam operational tests flagged for these conditions by grade and content area. Because some items were flagged for more than one condition, the number of flags may be greater than the number of flagged items.

The flagged items were referred to DRC’s content specialists for further review to ensure that the items were unambiguous and the answer keys were correct. As part of this review, DRC’s content experts also evaluated each flagged item against the Wisconsin Forward Exam depth-of-knowledge criteria to ensure that the cognitive demands of the item reflected the skills and knowledge that the item was designed to measure. Tables 8-B, 8-C, and 8-D provide more information about the flagged items.

### **8.1.1 Flagging for a Positive Distractor Correlation**

In Tables 8-B through 8-D, the distractor correlation coefficients are provided for items that were flagged because of positive distractor correlations. The distractor correlations tend to be small and are generally much smaller than the item-total correlations for the correct answer key. The majority of items flagged for a positive distractor-total test correlation had a distractor-total test correlation close to 0 and an acceptable item-total test correlation for the correct answer. All flagged items were judged to be acceptable based on their other statistics and were retained in order to meet the Wisconsin Forward Exam test blueprints.

### **8.1.2 Flagging for the Item-Total Correlation**

Three items were flagged for item-total correlations  $<0.15$  for Mathematics and one item was flagged for an item-total correlation  $<0.15$  for Social Studies. All of the flagged items had item-total test correlations of at least 0.12.

### **8.1.3 Flagging for $p$ -Value**

Twenty-four items were flagged for  $p$ -values  $<0.20$  in Mathematics assessments, and all flagged items had  $p$ -values between 0.05 and 0.18. While these statistics indicate items that were very difficult, the number of items flagged for difficulty was very small. No operational items were flagged for difficulty in ELA, Science, or Social Studies.

### **8.1.4 Flagging for Omit Rate**

No operational items on the Wisconsin Forward Exam were flagged for an omit rate higher than 5%. Most of the items had an omit rate less than 1%.

### **8.1.5 Speededness**

The degree to which a test is speeded can be evaluated by examining the percentage of students who fail to respond to the final items on a test or the last items in a timed section. One criterion of test speededness currently in use in the testing industry is a rule introduced by Educational Testing Services, which formulates that at least 80% of the test takers should be able to answer all of the items and all of the test takers should be able to answer at least 75% of the items (Swineford, 1956). However, a more stringent requirement is often applied, considering tests to be unspeeded only if at least 95% of the examinees attempt the final item. As shown in



Table 8-E, the Wisconsin Forward Exam satisfies this more stringent requirement, with more than 99% of the examinees attempting the final item in each of the four content areas.

### **8.1.6 Supplemental Tables on Classical Item Analysis**

Tables 8-1 through 8-17 present more comprehensive results from the classical item analysis for all of the items retained in each grade and content area. In those tables, the item-total test correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the total test score, the omit rate is flagged when it is above 5%, and the *p*-value is flagged when it is below 0.20.

The item analysis tables show the item number, which can be used to understand the location of test items as students actually encountered them on the test. The item analysis tables also indicate item type (e.g., MC, ESR). Items removed from the scoring of these tests are not included in these tables.

The number of flagged items across grade and content areas are summarized in Table 8-A. As indicated above, relatively few items were flagged. The item analysis indicated that the *p*-values of the items in the operational tests were well distributed throughout the range of difficulty levels, with point-biserial correlations reasonably high for most items. Detailed item analysis results including distractor statistics for MC items and score point distribution for non-multiple choice items are included in Appendix G.

## **8.2 Raw Score Results**

Raw score results based on all students who took the Spring 2017 Wisconsin Forward Exam are presented in Table 8-18. To facilitate interpretation of the raw score results, Table 8-18 provides the maximum possible score, the number of students, a measure of test difficulty, the standard deviation (SD) of raw scores, the skewness of the raw score distribution, kurtosis, the minimum obtained score, the maximum obtained score, the reliability (Cronbach's alpha), and the standard error of measurement (SEM) for raw scores. These measurements are further explained below. Readers can refer to Tables 3-1 through 3-4 for a count of the number of items in the test and the number of raw score points corresponding to each item.

The mean raw score varies by grade and content area and, specifically, in the context of the maximum possible score points. In ELA, for example, the maximum possible raw score is either 53 or 56, and it is either 42 or 46 in Mathematics.

Test difficulty is computed as the mean raw score divided by the maximum possible score points. Test difficulty ranges from 0 to 1.0. A larger test difficulty value indicates a mean raw score that is closer to the maximum possible score and, therefore, indicates an easier test. A smaller test difficulty value indicates a mean raw score that is further from the maximum possible score and, therefore, indicates a more difficult test. Consider an example: A test difficulty statistic would be 0.90 if a mean score of 45 were obtained on a test with a maximum possible score of 50. This would be considered an easier test. On the other hand, test difficulty

would be 0.50 if a mean raw score of 25 were obtained on the same test. This would then be considered a more difficult test. For example, the Mathematics grade 3 test mean raw score is 24.09 and the maximum possible score is 42, resulting in the test mean  $p$ -value of approximately 0.58. Note that this computation formula will not apply to ELA results. The mean  $p$ -value for ELA was computed using unweighted item scores, while the mean raw score was computed with weighted TDA items.

Table 8-18 also shows the skewness and kurtosis statistics for each distribution of raw scores. Skewness and kurtosis describe the shape of a distribution. When a distribution is perfectly normal, skewness is zero. A negative skew indicates a long tail on the left side of the distribution because of the presence of some low scores and (because the mean is sensitive to extreme scores) that most student scores are clustered on the high end of the scale. A positive skew indicates a distribution with some extreme high scores and a corresponding increase in the number of scores below the mean. Kurtosis describes a distribution in terms of its shape relative to a perfectly normal distribution. When a distribution is perfectly normal, kurtosis is zero. A negative kurtosis statistic indicates a distribution that is flatter than a perfectly normal curve, and a positive kurtosis statistic indicates a distribution that has more scores in the center of the score distribution (making it peaked) than a perfectly normal curve. Table 8-18 reveals that, in most cases, Wisconsin Forward Exam students are not normally distributed along the test scale in each grade and content area. Although this has implications for practitioners who wish to use Wisconsin Forward Exam raw scores in statistical analyses (normality of the data cannot be assumed), from a criterion-referenced testing standpoint, it indicates that students on the whole are mastering the Wisconsin Academic Standards for ELA and Wisconsin's Model Academic Standards for Science and Social Studies. The Mathematics assessments in grades 4 through 8 tend to be more difficult, however, showing most of the scores clustered below the mean (as indicated by positively skewed score distributions).

In addition, Table 8-18 shows that the minimum obtained scores in nine out of seventeen areas/grades are zero, meaning that at least one student failed all items for each of those tests. The table also shows that, except for ELA grades 3 through 6, the maximum obtained scores are equal to the maximum number of points possible on the test, meaning that at least one student obtained the full score for all items on each of those tests. For example, as displayed in Table 8-18, in Mathematics grade 3, there is at least one student who failed all items and at least one student who obtained a perfect raw score of 42.

A reliable test is one with high reliability, as represented by statistics such as Cronbach's alpha, and a low SEM. When interpreting reliability statistics, readers should note that test length (number of items and score points) is one of the important factors that influence reliability statistics and SEM. These concepts are described further in Part 9. For present purposes, the reader should note that measurement error is associated with every test score. A student's true score is the hypothetical average score that would result if the test could be administered repeatedly without the effects of practice or fatigue. Obtained scores should not be regarded as absolute but as one point within a range that, with a certain degree of probability, includes a student's true score.

The raw score results for each content area are summarized and discussed below using the measurements described above.

### **English Language Arts**

- Test difficulty ranged from 0.58 to 0.61.
- Standard deviations ranged from 9.01 to 10.59 raw score points.
- Alpha was relatively high in every grade (0.87 to 0.89).
- SEM ranged from 3.25 to 3.56.

### **Mathematics**

- Test difficulty ranged from 0.42 to 0.58, with generally lower difficulty in lower grades and higher difficulty in higher grades.
- Standard deviations ranged from 9.04 to 10.28 raw score points.
- Alpha was relatively high in every grade (0.91 to 0.92).
- SEM ranged from 2.69 to 2.86.

### **Science**

- Test difficulty was 0.70 for both grades.
- Standard deviations were 7.33 and 7.44 raw score points for grades 4 and 8, respectively.
- Alpha was 0.88 for both grades.
- SEM was 2.54 and 2.53 for grades 4 and 8, respectively.

### **Social Studies**

- Test difficulty ranged from 0.64 to 0.69.
- Standard deviations ranged from 7.33 to 10.19 raw score points.
- Alpha ranged from 0.88 to 0.91.
- SEM ranged from 2.49 to 3.01.

### **Subgroup Performance Patterns in Raw Score Results**

In the previous section, the raw score results were discussed with reference to the total student population. In this section, subgroup comparisons are made based on gender, race/ethnicity, socioeconomic status, disability status, and English language proficiency. These subgroup comparisons draw from Tables 8-19 through 8-26.

Overall, the raw score results show some consistent performance patterns by subgroups, that is, in terms of gender, race/ethnicity, socioeconomic status, disability status, and English language proficiency.

Regarding scores by gender, in ELA, the tests were slightly easier for female students as a group than for male students as a group in each grade level, with test difficulty differences ranging from 0.03 in grades 3 and 4 to 0.06 in grade 8. In Mathematics, the test difficulties were very similar between male and female students in grades 3, 5, 6, and 7 (differences of 0.1 or 0.0 in test  $p$ -value). At grade 4, the test was slightly easier for male students than for female students, with a difference of 0.03. At grade 8, the Mathematics test was slightly easier for female students, with a difference of 0.02. In Science, the test difficulties were very similar between male and female students in grades 4 and 8, with differences of 0.01 for grade 4 and 0.02 for grade 8. In Social Studies, there was no difference in the overall test difficulty between genders at grade 4, and the differences in test difficulty between genders were very small (at 0.02) for grades 8 and 10, with grade 8 female students performing slightly better than male students and grade 10 male students performing slightly better than female students.

In all grades and content areas, the raw score results showed consistent performance patterns by ethnicity. In every grade and content area, the test was generally the easiest for White students, followed by Asian students, American Indian students and Hispanic students, and African-American students. American Indian students had similar or slightly lower mean raw scores than Hispanic students. Differences in test difficulty between American Indian and Hispanic students were 0.00 or 0.01 in most grades and content areas.

In every grade and content area, the test was easier for students who were not economically disadvantaged than for those who were economically disadvantaged. The difference in test difficulty between the two groups ranged from 0.12 in ELA grade 3 to 0.16 in Mathematics grade 4.

There were also differences in test difficulty between students with disabilities and those without disabilities in all grades and content areas. The test was consistently easier for students without disabilities than for students with disabilities, with differences ranging from 0.13 in ELA grade 3 and Science grade 4 to 0.22 in Social Studies grade 8. Larger differences in student performance were observed for higher grade levels compared to lower grade levels.

In every grade and content area, the test was markedly easier for students who were fully English proficient than for students who were limited English proficient. Differences in test difficulty ranged from 0.12 in ELA grade 3 to 0.24 in Social Studies grade 10. Larger differences in student performance were observed for higher grade levels compared to lower grade levels.

### **8.3 Summary Statistics for Scale Scores**

The Wisconsin Forward Exam program reports scale scores as well as raw scores. The scale score of a student in a given content area represents the student's level of performance in that content area. Higher scale scores indicate higher levels of performance, and lower scale scores indicate lower levels of performance. Scale scores are based on the entire set of scored operational items per grade and content area.

Summary descriptive statistics based on the scale score results are described below. Table 8-27 is the summary scale score table based on the census data. The table shows the mean scale score, the standard deviation of the scale scores, skewness and kurtosis, the minimum and maximum obtained scale scores, and the lowest and highest obtainable scale scores (LOSS and HOSS, respectively) for all content areas and grades based on the census data. The LOSS and HOSS, as discussed in Part 6, identify the lower and upper limits of the scale score range. These values were established when the current scales were developed and do not change from one administration to another.

### **English Language Arts**

- Mean scale score increased as grade level increased, ranging from 559.12 for grade 3 to 637.69 for grade 8. This mean scale score pattern supports the ELA vertical scale properties.
- Standard deviations ranged from 46.93 to 61.61 scale score points.
- In grades 7 and 8, student scores spanned the full-scale score range from the LOSS to the HOSS. The HOSS was not obtained in grades 3 through 6.

### **Mathematics**

- Mean scale score increased by grade level, ranging from 555.03 for grade 3 to 641.11 for grade 8. This mean scale score pattern supports the Mathematics vertical scale properties.
- Standard deviations ranged from 48.63 to 59.36 scale score points.
- In each grade level, student scores spanned the full-scale score range from the LOSS to the HOSS.

### **Science**

- Mean scale scores were 399.27 and 594.12 for grades 4 and 8, respectively.
- Standard deviations were 53.16 and 51.25 scale score points for grades 4 and 8, respectively.
- In each grade level, student scores spanned the full-scale score range from the LOSS to the HOSS.

### **Social Studies**

- Mean scale scores were 397.05, 597.60, and 696.92 for grades 4, 8, and 10, respectively.
- Standard deviations ranged from 51.71 to 56.56 scale score points.
- In each grade level, student scores spanned the full-scale score range from the LOSS to the HOSS.

## **Subgroup Performance Patterns in Scale Score Results**

The scale score results, like the raw score results, showed some consistent performance patterns in terms of subgroups. The results for gender, race/ethnicity, socioeconomic status, disability status, and English language proficiency are drawn from Tables 8-28 through 8-35.

### **Gender**

- In terms of gender, male students as a group showed lower mean scale scores in ELA than female students as a group in each grade level. The difference ranged from 8.58 scale score points in grade 3 to 22.22 scale score points in grade 8.
- In Mathematics, male students as a group showed slightly higher mean scale scores in grades 3 through 5 and lower mean scale scores in grades 6 through 8 than female students. The differences between genders ranged from 0.53 scale score points in grade 5 to 6.63 scale score points in grade 8.
- In Science, the mean scale scores between genders were very similar, with a difference of 1.18 scale score points in grade 4 and a difference of 3.93 scale score points in grade 8. Male students performed slightly better than female students in grade 4, and female students performed better than male students in grade 8.
- There were very small differences between mean scale scores by gender in Social Studies, from 1.74 scale score points to 4.45 scale score points. Female students performed better than male students in grades 4 and 8. Male students performed better than female students in grade 10.

### **Race/Ethnicity**

- The scale score results showed some consistent performance differences by ethnicity.
- In every grade and content area, White students as a group had the highest mean scale scores, followed by Asian students, Hispanic students and American Indian students, and African-American students.
- As was noted in the context of the raw score results, the differences in mean scale scores for American Indian students and Hispanic students were often very small.

### **Socioeconomic Status**

- Economically disadvantaged students as a group scored lower than students who were not economically disadvantaged as a group across all grades and content areas. Differences ranged from 31.90 scale score points in ELA grade 3 to 42.95 scale score points in Mathematics grade 8.
- For every grade and content area, the mean scale score of students who were economically disadvantaged was more than two-thirds standard deviation lower than the mean scale score of students who were not economically disadvantaged.

## Disability Status

- Students with disabilities and students without disabilities showed consistent and large differences in mean scale scores by group. Differences ranged from 33.42 scale score points in ELA grade 3 to 68.40 scale score points in ELA grade 8.
- For every grade and content area, the mean scale scores of students with disabilities were lower than the mean scale scores of students without disabilities by about two-thirds to over one standard deviation.

## English Language Proficiency

- Students who were fully English proficient and students who were limited English proficient showed consistent and large differences in mean scale scores by group. Differences ranged from 27.22 scale score points in ELA grade 3 to 64.26 scale score points in Social Studies grade 10.
- For every grade and content area, the mean scale scores of limited English proficient students were lower than the mean scale scores of fully English proficient students by about half to over one standard deviation.

## 8.4 Cut Scores and Performance Level Classifications

Student performance on the Wisconsin Forward Exam is reported in terms of four performance categories: *Below Basic*, *Basic*, *Proficient*, and *Advanced*. These performance categories are established through cut scores.

Standard 5.21 of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) indicates that “when proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be documented clearly” (107).

In terms of the validity of the Wisconsin Forward Exam, it is essential to understand that cut scores and performance level descriptors (PLDs) are established in a collaborative and participatory process. The descriptors clearly establish, in plain language, the proper frame of reference for understanding how to interpret test scores, particularly cut scores. PLDs summarize the knowledge, skills, and abilities expected of students in each achievement level. As stated in Part 7, DPI provided policy PLDs for the Wisconsin Forward Exam assessments. At the standard setting, Wisconsin used the policy PLDs in conjunction with the content standards to consider the content-based expectations for students in each achievement level on each test in the Wisconsin Forward Exam program.

Table 8-36 shows the cut scores for each content and grade level. For ease of reference, Tables 8-37 through 8-40 provide the scale score ranges that define performance levels together with the percentage of students in each performance level. The results for each content area and grade are summarized below.

## English Language Arts

- Between approximately 41% (grade 8) and 47% (grade 4) of students were either *Proficient* or *Advanced* in ELA.
- Between 8% and 12% of students were classified as *Advanced*, depending on the grade level.
- Across all grade levels, more than 50% of students were below *Proficient*. These percentages ranged from approximately 53% below *Proficient* in grade 4 to 59% below *Proficient* in grade 8.

## Mathematics

- Between approximately 35% (grade 8) and 48% (grade 3) of students were either *Proficient* or *Advanced* in Mathematics.
- The proportion of students who were *Advanced* was between approximately 5% and 11%, depending on the grade level.
- Across all grade levels, more than 50% of students were below *Proficient*. These percentages ranged from approximately 52% below *Proficient* in grade 3 to 65% below *Proficient* in grade 8.

## Science

- Approximately 51% of students were either *Proficient* or *Advanced* in grade 4 and about 48% of students were either *Proficient* or *Advanced* in grade 8.
- The percentage of students classified as *Advanced* was approximately 16% in grade 4 and close to 14% in grade 8.
- The proportion of students classified as below *Proficient* was approximately 49% in grade 4 and 52% in grade 8.

## Social Studies

- About half or more of the total students in each grade level were either *Proficient* or *Advanced* in Social Studies grades 4 and 8. The percentage of *Proficient* or *Advanced* students was approximately 52% in grade 4, 50% in grade 8, and 48% in grade 10.
- Approximately 20% of students were *Advanced* in grades 4 and 10, and about 19% of students were classified as *Advanced* in grade 8.
- The percentage of students classified as below *Proficient* was approximately 48% in grade 4, 50% in grade 8, and 52% in grade 10.

## Subgroup Patterns in Performance Level Results

The performance level results varied by subgroup: gender, race/ethnicity, socioeconomic status, disability status, and English language proficiency. The main subgroup performance patterns are described below. These comparisons are based on Tables 8-41 through 8-44.



In terms of gender, the percentages of both genders were generally similar in *Proficient* or above performance levels for Mathematics, Science, and Social Studies across all grades. The differences in the percentages of male and female students in *Proficient* or above categories for these content areas were, on average, less than 5%. For ELA, more female students than male students were classified as *Proficient* or above (with the differences between genders being about 10%) in all grades.

There were some consistent patterns in performance by ethnicity across grades and content areas. In terms of the *Proficient* or above categories, the prevailing tendency was that there were higher percentages of White students as a group, followed by Asian students, American Indian students and Hispanic students, and African-American students. The inverse sequence was found at the *Below Basic* performance level.

There were consistent differences in performance between economically disadvantaged students and not economically disadvantaged students. In every grade and content area, there were much higher percentages of students who were not economically disadvantaged classified as *Proficient* or above. There were much higher percentages of students who were economically disadvantaged who were classified in the lowest performance category.

Performance level results showed that there were higher percentages of students without disabilities who were classified as *Proficient* or above, and there were much higher percentages of students without disabilities in the reporting category *Advanced*. There were also much lower percentages of students without disabilities in the lowest performance level than students with disabilities. This pattern was evident in all grades and all content areas.

Performance level results showed a similar pattern in comparisons of students who were fully English proficient with students who were limited English proficient. In every grade and content area, there were generally higher percentages of students who were fully English proficient classified as *Proficient* and much higher percentages of students who were fully English proficient classified as *Advanced*. There were much lower percentages of fully English proficient students who were classified in the lowest performance level in all grades and content areas.

## **8.5 Standard Performance Index for Content Standards**

In addition to raw scores and scale scores, teachers and educational decision-makers frequently need diagnostic information to inform instructional strategies. Diagnostic information also helps to identify individual student strengths and needs. This kind of information can be derived from scores on subsets of test items that estimate how much a student knows in a clearly defined skill domain. These skill domains are called content standards (or standards or objectives). Scores on subsets of test items at the content standard level are called standard performance index (SPI) scores. The purpose of reporting SPI scores on the Wisconsin Forward Exam is to show the relationship between the overall achievement being measured (represented by the test score) and the skills within each of the content standards associated with the overall content area. Teachers may use the SPI scores for individual students as indicators of strengths

and weaknesses, but the SPI scores are best corroborated by other evidence, such as homework, class participation, diagnostic test scores, or observation. District and school administrators may compare their results by content standard and grade level with the state mean percentage to better understand their strengths and weaknesses within a particular content area and grade level.

An SPI score can be interpreted as an estimate of the number of items a student would be expected to answer correctly if there had been 100 similar items for a given reporting category. For example, an SPI score of 77 for a given reporting category means that, if the student were given 100 similar items, the student would be expected to answer 77 of them correctly. This is a criterion-referenced score, in that it estimates how much a student knows in a clearly defined skill domain (i.e., the criterion). Technical readers can refer to Appendix H of this report for more details.

This approach, identifying student proficiency on each content standard, relates to the ELA and Mathematics Wisconsin Academic Standards and Wisconsin’s Model Academic Standards for Science and Social Studies. SPI scores provide a more reliable estimate of student achievement on each content standard than is possible by simply reporting percent correct. However, *SPI scores should be used for low-stakes purposes because these scores cannot be considered stable for any content standard with a small number of items.*

Readers should note that the average difficulty of items will vary across content standards and grades. Content standards vary in their complexity, level of abstraction, and cognitive demand. Some standards may be intrinsically more difficult than others, and the difficulty of individual items is determined, in part, by the difficulty of the content domain being measured. The current test blueprints do not specify the average difficulty level of items for each content standard within grades or across grades. If the difficulty of the items varies across years, grades, and content standards, the mean SPI scores will be affected by differences in item difficulty as well as differences in student ability. *Thus, differences in SPI scores across years, grades, or content standards should not be seen as reliable indicators of differences in student ability, since these differences may be explained in whole or in part by differences in the difficulty of the items themselves.* However, comparisons across years, grades, or content standards are appropriate for assessing the relative difficulty of the items, and comparisons of individual student scores or of group mean scores on a single SPI score can provide useful information about the *relative* strengths and needs of individual students or groups on these standards.

Tables 8-45 through 8-48 identify the content standards/domain, the number of MC and CR items within each standard/domain, the total number of possible points per standard/domain, the mean raw score, the mean *p*-value, the standard deviation of the raw scores, the mean SPI score, and the standard deviation of SPI scores for all content areas across grades. The results from Tables 8-45 through 8-48 are summarized below. Tables 8-49 through 8-52 identify the SPI cut scores for each content area reporting category and grade level.

## **English Language Arts**

Tables 8-45a and 8-45b present mean *p*-values and SPI scores for ELA across content standards/domains and grades. Results show that the mean ELA SPI scores across grades ranged

from 53.18 to 60.27 for content standards and from 52.47 to 58.57 for domains, indicating that the items were moderately difficult for examinees. In general, content standard D (Writing/Language—Text Types and Purposes) was the most difficult in grade 3, and content standard E (Writing/Language—Research) was the most difficult in grades 4 and above. These content standards contained the TDA item, which was generally difficult for students. The Listening domain was easier than other domains for students in grades 3, 4, 5, and 8, and the Writing domain was the most difficult for students in all grades except for grade 5 (where the Reading domain was just slightly more difficult than the Writing domain).

## **Mathematics**

Table 8-46 presents Mathematics *p*-values and SPI scores across grades and content standards. Results show that the mean *p*-values and SPI scores varied across standards in all grades. Mean SPI scores, across all content standards, ranged from 57.80 for grade 3 to 41.90 for grade 7 and 42.17 for grade 8, indicating that the Mathematics items were more challenging for higher grades than lower grades. There was no consistent pattern in regard to the content standard difficulty across grade levels.

Content standard D (Measurement and Data) was the most difficult in grade 3, and content standard C (Number and Operations—Fractions) was the most difficult in grade 4. Content standard E (Geometry) was the most difficult in grades 5 and 6. Content standard H (Expressions and Equations) was the most difficult in grade 7, and content standard G (The Number System) was the most difficult in grade 8.

## **Science**

Table 8-47 presents Science *p*-values and SPI scores across grades and content standards. The mean Science SPI scores across all content standards were 69.03 for grade 4 and 69.45 for grade 8, indicating that the test items were relatively easy. SPI scores indicated that content standard E (Earth and Space Science) was the most difficult in both grades.

## **Social Studies**

Table 8-48 presents Social Studies *p*-values and SPI scores across grades and content standards. The mean Social Studies SPI scores across all content standards ranged from 68.10 for grade 8 to 63.07 for grade 10, indicating that the test items were relatively easy. The mean SPI scores indicated that the most difficult content standard varied between the three Social Studies grades. In grades 4 and 10, the most difficult standard was content standard D (Economics), and in grade 8, the most difficult standard was content standard B (History).

## **Summary of Student Performance Indicator Results**

Overall, the mean SPI scores across grades and content standards range in difficulty. The content standards with SPI mean scores  $>75$  were the following:

- Grade 6 ELA content standard D (Writing/Language—Text Types and Purposes)

- Grade 4 Science content standard G/H (Science Applications and Personal Social Perspectives)
- Grade 8 Science content standard A/B (Science Connections and Nature of Science) and content standard D (Physical Science)

The one content standard with an SPI mean score <35 was the grade 7 Mathematics content standard H (Expressions and Equations).

It is important to note that some variation in difficulty of the items across content standards within and across grades and test forms is inevitable and that some of that variation is independent of any intrinsic differences in the difficulty of the standards themselves (e.g., variations in the difficulty of the particular items that were selected for the test forms). For this reason, SPI scores should be interpreted with caution and should not be used to make comparisons of student performance across testing years or grade levels.

## 8.6 Longitudinal Comparisons of Test Scores

It is often desirable to examine the scores of students across time and monitor group performance. This is possible if the test content and the construct measured by the test are comparable from year to year and if the scores are reported on the same scale in multiple years.

For the Wisconsin Forward Exam assessments, two years of the test scores on the same reporting scales are available, and the state-level scale score means and standard deviations for 2016 and 2017 administration years are presented for ELA, Mathematics, Science, and Social Studies in Tables 8-53 through 8-56. The statistics presented in these tables are based on the total population of Wisconsin students, including students attending public, choice, and private schools. (Note that the Spring 2016 student performance data presented in this section of the report differs from the Spring 2016 student performance data presented in the *Wisconsin Forward Exam Spring 2016 Technical Report*, in which the summary of student performance was based on the public school data only).

It was observed that the mean scale score for ELA increased for all grade levels except for grade 3. The score increase for grades 4 through 8 ranged from approximately half a score point for grade 8 to over 4 scale score points for grade 6. The scale score decrease for grade 3 was small and just over 1 scale score point.

The mean scale score for Mathematics increased by less than 1 scale score point for grades 3 through 6 and grade 8 between the last two test administrations. The mean scale score for grade 7 Mathematics did not, practically, change between Spring 2016 and 2017 administrations.

For Science, the mean scale score increased by approximately half a score point for grade 4 and decreased by approximately 4 scale score points for grade 8 between the 2016 and 2017 test administrations.

For Social Studies, the mean scale scores decreased between the 2016 and 2017 test administrations for all grade levels. The decrease in the scale score means was small and ranged from approximately half a point for grade 8 to less than 2 points for grade 10.

Tables 8-57 through 8-60 show the percentage of students in each achievement level in Spring 2016 and 2017 test administrations for ELA, Mathematics, Science, and Social Studies. The results presented in these tables are based on the total population of Wisconsin students, including students attending public, choice, and private schools.

For ELA, an increase in the percentage of students at or above *Proficient* was observed for grades 4 through 7, ranging from less than 2% for grade 7 to close to 4% for grade 5. A small decrease in the percentage of students at or above *Proficient* was observed for grade 3 (a decrease of approximately 1%) and for grade 8 (a decrease of less than 1%).

For Mathematics, a small increase in the percentage of students at or above *Proficient* was observed for grades 5, 6, and 8. This increase was less than 1%. A small decrease in the percentage of students at or above *Proficient* was observed for grade 4 (less than 1%). There was no practical change in the percentage of students at or above *Proficient* for grades 3 and 7.

For Science, less than a 1% decrease in the percentage of students at or above *Proficient* was observed for grade 4, and about a 2% decrease in the percentage of students at or above *Proficient* was observed for grade 8.

For Social Studies, less than a 1% decrease in the percentage of students at or above *Proficient* was observed for grades 4 and 10, and less than a 1% increase in the percentage of students at or above *Proficient* was seen for grade 8.

Overall, the percentages of students classified in any of the four performance level categories were found to be comparable between the Spring 2016 and 2017 test administrations across all grade levels and content areas. With a few exceptions, the changes between the percentage of students in Spring 2016 and Spring 2017 in any performance level category, grade, or content were less than 2%.

## **8.7 Summary**

In the Wisconsin Forward Exam, the purpose of the ELA, Mathematics, Science, and Social Studies assessments is to demonstrate student achievement through test scores in the respective content areas. The results presented in Part 8, together with the reliability and validity evidence, indicate that the scale scores and performance levels reported in the Wisconsin Forward Exam program are valid and reliable evidence of student achievement in the tested content areas and grades. As such, test scores and performance levels can be used to classify students, schools, districts, and the state with respect to how much achievement is shown for each content area. Classroom teachers may use these scores as evidence of student achievement in these content areas. District and school administrators may use this information for activities such as planning curricula. At the state level, the overall results, including the longitudinal test results, can be drawn upon for accountability and reporting purposes.

Table 8-A Summary of Flagged Operational Items on the Spring 2017 Wisconsin Forward Exam

Content	Grade	# of Items Flagged	Number of Flags			
			Correlation <0.15	Distractor Correlation >0	Omit >5%	p-Value <0.20
ELA	3	1		1		
	4	2		2		
	5	1		1		
	6	1		1		
	7	1		1		
	8	0				
MA	3	2		1		1
	4	1	1	1		
	5	12	1	4		7
	6	9	1	4		4
	7	9		3		6
	8	9		3		6
SC	4	2		2		
	8	1		1		
SS	4	1		1		
	8	2		2		
	10	2	1	2		
<b>Total</b>		<b>56</b>	<b>4</b>	<b>30</b>	<b>0</b>	<b>24</b>

Note: The number of flags may be greater than the number of flagged items.

Table 8-B English Language Arts Items Flagged for Classical Item Analysis Statistics

Grade	Content	Item	Item Type	<i>p</i> -Value	Corr	Percent Omit	Flags				
							Corr	Distractor	Omit	<i>p</i> -Value	
3	ELA	33	MC	0.37	0.24	0.27		+	0.04		
4	ELA	16	MC	0.59	0.19	0.13		+	0.03		
	ELA	31	MC	0.47	0.30	0.21		+	0.02		
5	ELA	22	MC	0.33	0.27	0.07		+	0.03		
6	ELA	5	MC	0.82	0.17	0.05		+	0.00		
7	ELA	26	MC	0.50	0.21	0.26		+	0.05		

Table 8-C Mathematics Items Flagged for Classical Item Analysis Statistics

Grade	Content	Item	Item Type	p-Value	Corr	Percent Omit	Flags				
							Corr	Distractor	Omit	p-Value	
3	MA	16	MC	0.17	0.25	0.15					+
	MA	30	MC	0.36	0.18	0.19		+	0.02		
4	MA	43	MC	0.26	0.14	0.17	+	+	0.12		
5	MA	3	MC	0.59	0.19	0.13		+	0.02		
	MA	11	MC	0.64	0.12	0.16	+				
	MA	12	TE	0.16	0.48	0.50					+
	MA	25	MC	0.26	0.17	0.11		+	0.11		
	MA	26	ESR	0.17	0.46	0.13					+
	MA	27	SA	0.15	0.45	0.44					+
	MA	28	SA	0.14	0.42	0.20					+
	MA	37	ESR	0.10	0.26	0.19					+
	MA	38	MC	0.39	0.25	0.33		+	0.12		
	MA	39	MC	0.14	0.30	0.45					+
	MA	41	MC	0.27	0.39	0.19		+	0.02		
	MA	46	ESR	0.07	0.26	0.25					+
6	MA	9	TE	0.12	0.44	0.35					+
	MA	10	MC	0.33	0.15	0.10		+	0.25		
	MA	13	SA	0.18	0.48	0.17					+
	MA	28	MC	0.32	0.33	0.36		+	0.04		
	MA	29	ESR	0.06	0.22	0.43					+
	MA	33	SA	0.11	0.50	0.51					+
	MA	41	MC	0.34	0.12	0.42	+				
	MA	43	MC	0.34	0.22	0.27		+	0.02		
MA	44	MC	0.37	0.25	0.35		+	0.05			
7	MA	7	SA	0.14	0.49	0.40					+
	MA	11	MC	0.33	0.25	0.12		+	0.07		
	MA	15	TE	0.15	0.39	0.73					+
	MA	21	ESR	0.18	0.42	0.44					+
	MA	27	TE	0.09	0.34	0.53					+
	MA	29	ESR	0.14	0.35	0.40					+
	MA	37	TE	0.13	0.51	0.85					+
	MA	39	MC	0.30	0.22	0.61		+	0.05		
	MA	45	MC	0.30	0.35	0.48		+	0.10		



Table 8-C Mathematics Items Flagged for Classical Item Analysis Statistics (cont.)

Grade	Content	Item	Item Type	p-Value	Corr	Percent Omit	Flags				
							Corr	Distractor	Omit	p-Value	
8	MA	5	MC	0.48	0.50	0.12		+	0.01		
	MA	6	SA	0.07	0.36	0.47					+
	MA	11	SA	0.05	0.31	0.75					+
	MA	13	SA	0.17	0.50	0.59					+
	MA	18	SA	0.10	0.40	0.45					+
	MA	21	TE	0.10	0.40	0.48					+
	MA	25	MC	0.23	0.17	0.31		+	0.02		
	MA	40	ESR	0.17	0.43	0.55					+
	MA	41	MC	0.36	0.28	0.57		+	0.02		

Table 8-D Science and Social Studies Items Flagged for Classical Item Analysis Statistics

Grade	Content	Item	Item Type	p-Value	Corr	Percent Omit	Flags				
							Corr	Distractor	Omit	p-Value	
4	SC	9	MC	0.35	0.21	0.40		+	0.02		
	SC	25	MC	0.47	0.16	0.10		+	0.08		
8	SC	33	MC	0.36	0.23	0.23		+	0.02		
4	SS	14	MC	0.42	0.23	0.11		+	0.01		
8	SS	17	MC	0.49	0.34	0.15		+	0.02		
	SS	22	MC	0.45	0.36	0.13		+	0.04		
10	SS	18	MC	0.23	0.13	0.27	+	+	0.02		
	SS	47	MC	0.45	0.24	0.46		+	0.02		

Table 8-E Percentage of Students Attempting Last Operational Item in Test

Content	Grade						
	3	4	5	6	7	8	10
English Language Arts	99.73	99.80	99.80	99.77	99.75	99.77	
Mathematics	99.81	99.77	99.75	99.72	99.44	99.54	
Science		99.87				99.79	
Social Studies		99.87				99.80	99.51

Table 8-1 Item Analysis, Grade 3 English Language Arts

Item	Item Type	<i>p</i> -Value	Corr	Percent Omit	Flags			
					Corr	Distractor	Omit	<i>p</i> -Value
1	MC	0.64	0.37	0.12				
2	MC	0.83	0.28	0.11				
3	MC	0.57	0.31	0.12				
4	TDA	0.33	0.45	0.26				
5	MC	0.67	0.30	0.09				
6	TE	0.58	0.34	0.56				
7	TE	0.74	0.49	0.11				
8	TE	0.43	0.21	0.75				
9	MC	0.81	0.43	0.18				
10	TE	0.51	0.44	0.15				
11	MC	0.64	0.32	0.15				
12	MC	0.60	0.41	0.20				
13	MC	0.53	0.30	0.18				
14	MC	0.49	0.30	0.19				
15	TE	0.40	0.29	0.36				
16	ESR	0.71	0.43	0.06				
17	MC	0.69	0.49	0.15				
18	MC	0.88	0.35	0.13				
19	ESR	0.57	0.51	0.12				
20	MC	0.56	0.28	0.20				
21	MC	0.56	0.28	0.12				
22	ESR	0.65	0.53	0.09				
23	TE	0.49	0.58	0.32				
24	MC	0.50	0.33	0.25				
25	MC	0.64	0.44	0.25				
26	ESR	0.40	0.46	0.14				
27	MC	0.61	0.45	0.23				
28	MC	0.52	0.44	0.24				
29	MC	0.64	0.37	0.23				
30	TE	0.69	0.52	0.36				
31	MC	0.41	0.33	0.29				
32	MC	0.51	0.36	0.27				
33	MC	0.37	0.24	0.27		+		

Table 8-2 Item Analysis, Grade 4 English Language Arts

Item	Item Type	<i>p</i> -Value	Corr	Percent Omit	Flags			
					Corr	Distractor	Omit	<i>p</i> -Value
1	MC	0.49	0.43	0.08				
2	MC	0.80	0.47	0.11				
3	TE	0.57	0.56	0.13				
4	TE	0.26	0.44	0.43				
5	MC	0.61	0.42	0.11				
6	TDA	0.37	0.54	0.30				
7	TE	0.34	0.29	0.12				
8	TE	0.65	0.46	0.14				
9	MC	0.83	0.36	0.09				
10	MC	0.65	0.32	0.14				
11	MC	0.72	0.37	0.14				
12	MC	0.45	0.26	0.15				
13	MC	0.60	0.45	0.16				
14	TE	0.54	0.47	0.14				
15	MC	0.71	0.46	0.14				
16	MC	0.59	0.19	0.13		+		
17	MC	0.72	0.47	0.16				
18	TE	0.70	0.38	0.54				
19	TE	0.80	0.35	0.16				
20	TE	0.69	0.48	0.07				
21	MC	0.60	0.36	0.14				
22	MC	0.56	0.31	0.14				
23	ESR	0.61	0.44	0.10				
24	MC	0.73	0.40	0.19				
25	MC	0.38	0.30	0.19				
26	MC	0.57	0.36	0.14				
27	MC	0.89	0.40	0.15				
28	MC	0.81	0.42	0.14				
29	MC	0.75	0.49	0.20				
30	TE	0.28	0.47	0.44				
31	MC	0.47	0.30	0.21		+		
32	TE	0.74	0.50	0.16				
33	MC	0.56	0.42	0.21				
34	MC	0.55	0.35	0.25				
35	MC	0.69	0.50	0.24				
36	ESR	0.30	0.39	0.16				
37	MC	0.42	0.32	0.20				

Table 8-3 Item Analysis, Grade 5 English Language Arts

Item	Item Type	<i>p</i> -Value	Corr	Percent Omit	Flags			
					Corr	Distractor	Omit	<i>p</i> -Value
1	MC	0.84	0.37	0.05				
2	MC	0.80	0.37	0.05				
3	MC	0.78	0.43	0.07				
4	TDA	0.33	0.49	0.12				
5	MC	0.66	0.22	0.05				
6	MC	0.70	0.28	0.10				
7	MC	0.84	0.42	0.09				
8	MC	0.68	0.33	0.10				
9	MC	0.68	0.45	0.11				
10	MC	0.76	0.32	0.11				
11	MC	0.82	0.42	0.10				
12	TE	0.59	0.46	0.08				
13	MC	0.85	0.38	0.08				
14	MC	0.49	0.29	0.13				
15	MC	0.76	0.44	0.16				
16	MC	0.61	0.31	0.12				
17	MC	0.64	0.40	0.11				
18	TE	0.71	0.38	0.09				
19	TE	0.61	0.42	0.05				
20	MC	0.62	0.34	0.10				
21	MC	0.79	0.42	0.07				
22	MC	0.33	0.27	0.07		+		
23	MC	0.56	0.33	0.12				
24	ESR	0.39	0.45	0.07				
25	TE	0.42	0.40	0.08				
26	MC	0.55	0.33	0.16				
27	MC	0.44	0.35	0.23				
28	MC	0.56	0.46	0.19				
29	ESR	0.23	0.30	0.08				
30	MC	0.61	0.39	0.21				
31	MC	0.66	0.53	0.22				
32	ESR	0.41	0.32	0.09				
33	MC	0.51	0.29	0.26				
34	MC	0.53	0.45	0.24				
35	TE	0.45	0.38	0.19				
36	MC	0.55	0.31	0.19				
37	MC	0.40	0.21	0.20				

Table 8-4 Item Analysis, Grade 6 English Language Arts

Item	Item Type	<i>p</i> -Value	Corr	Percent Omit	Flags			
					Corr	Distractor	Omit	<i>p</i> -Value
1	TE	0.72	0.46	0.05				
2	MC	0.73	0.39	0.07				
3	MC	0.48	0.34	0.08				
4	TDA	0.39	0.57	0.24				
5	MC	0.82	0.17	0.05		+		
6	MC	0.88	0.39	0.07				
7	MC	0.76	0.39	0.09				
8	MC	0.68	0.26	0.12				
9	TE	0.54	0.30	0.23				
10	TE	0.69	0.41	0.11				
11	TE	0.33	0.20	0.20				
12	TE	0.55	0.49	0.23				
13	MC	0.81	0.39	0.10				
14	TE	0.25	0.23	0.17				
15	TE	0.40	0.41	0.26				
16	MC	0.56	0.37	0.15				
17	MC	0.82	0.37	0.07				
18	MC	0.85	0.38	0.16				
19	TE	0.63	0.44	0.11				
20	ESR	0.37	0.25	0.07				
21	MC	0.50	0.32	0.14				
22	MC	0.65	0.40	0.17				
23	TE	0.69	0.48	0.35				
24	MC	0.54	0.29	0.23				
25	MC	0.51	0.38	0.27				
26	MC	0.46	0.35	0.26				
27	MC	0.40	0.33	0.18				
28	MC	0.79	0.42	0.18				
29	MC	0.78	0.40	0.22				
30	TE	0.59	0.44	0.36				
31	MC	0.75	0.48	0.28				
32	MC	0.72	0.42	0.27				
33	TE	0.68	0.47	0.25				
34	TE	0.58	0.36	0.29				
35	MC	0.51	0.33	0.23				

Table 8-5 Item Analysis, Grade 7 English Language Arts

Item	Item Type	<i>p</i> -Value	Corr	Percent Omit	Flags			
					Corr	Distractor	Omit	<i>p</i> -Value
1	MC	0.70	0.37	0.06				
2	MC	0.75	0.46	0.07				
3	MC	0.79	0.47	0.08				
4	TDA	0.45	0.60	0.29				
5	TE	0.65	0.28	0.22				
6	MC	0.53	0.23	0.15				
7	MC	0.47	0.36	0.15				
8	MC	0.72	0.30	0.14				
9	ESR	0.57	0.46	0.11				
10	TE	0.65	0.31	0.29				
11	TE	0.46	0.25	0.22				
12	MC	0.80	0.37	0.13				
13	MC	0.51	0.39	0.15				
14	MC	0.64	0.35	0.18				
15	MC	0.59	0.26	0.22				
16	ESR	0.50	0.47	0.13				
17	MC	0.76	0.38	0.07				
18	ESR	0.70	0.55	0.05				
19	MC	0.44	0.43	0.13				
20	ESR	0.60	0.43	0.08				
21	TE	0.60	0.54	0.15				
22	MC	0.64	0.34	0.08				
23	MC	0.65	0.38	0.21				
24	MC	0.80	0.47	0.17				
25	ESR	0.65	0.49	0.09				
26	MC	0.50	0.21	0.26		+		
27	MC	0.56	0.45	0.22				
28	MC	0.46	0.27	0.40				
29	TE	0.63	0.43	2.58				
30	MC	0.46	0.32	0.37				
31	MC	0.58	0.38	0.26				
32	ESR	0.55	0.54	0.20				
33	MC	0.58	0.50	0.28				
34	MC	0.78	0.43	0.27				
35	MC	0.67	0.45	0.26				
36	MC	0.54	0.48	0.25				

Table 8-6 Item Analysis, Grade 8 English Language Arts

Item	Item Type	<i>p</i> -Value	Corr	Percent Omit	Flags			
					Corr	Distractor	Omit	<i>p</i> -Value
1	TE	0.48	0.36	0.48				
2	ESR	0.53	0.37	0.04				
3	MC	0.66	0.45	0.09				
4	MC	0.67	0.20	0.09				
5	TDA	0.42	0.59	0.69				
6	MC	0.62	0.34	0.05				
7	MC	0.62	0.39	0.12				
8	MC	0.49	0.19	0.14				
9	MC	0.66	0.37	0.12				
10	MC	0.47	0.43	0.10				
11	TE	0.52	0.48	0.63				
12	MC	0.75	0.45	0.12				
13	MC	0.70	0.36	0.10				
14	TE	0.77	0.51	0.20				
15	MC	0.59	0.41	0.14				
16	TE	0.49	0.35	0.46				
17	MC	0.54	0.27	0.16				
18	TE	0.52	0.41	0.14				
19	MC	0.78	0.31	0.06				
20	MC	0.69	0.36	0.14				
21	ESR	0.47	0.38	0.05				
22	MC	0.49	0.41	0.14				
23	MC	0.74	0.42	0.17				
24	ESR	0.66	0.49	0.09				
25	MC	0.73	0.52	0.09				
26	MC	0.48	0.25	0.16				
27	MC	0.65	0.42	0.18				
28	MC	0.67	0.50	0.23				
29	MC	0.63	0.50	0.18				
30	TE	0.56	0.39	0.48				
31	TE	0.50	0.45	0.21				
32	TE	0.73	0.58	0.25				
33	MC	0.64	0.35	0.24				
34	TE	0.58	0.40	0.23				
35	MC	0.77	0.48	0.25				
36	MC	0.61	0.42	0.23				

Table 8-7 Item Analysis, Grade 3 Mathematics

Item	Item Type	<i>p</i> -Value	Corr	Percent Omit	Flags			
					Corr	Distractor	Omit	<i>p</i> -Value
1	MC	0.65	0.46	0.10				
2	MC	0.71	0.44	0.08				
3	MC	0.77	0.47	0.12				
4	MC	0.76	0.45	0.10				
5	MC	0.52	0.46	0.10				
6	TE	0.43	0.58	0.29				
7	MC	0.82	0.43	0.40				
8	SA	0.52	0.38	0.18				
9	MC	0.65	0.39	0.14				
10	SA	0.60	0.58	0.11				
11	MC	0.46	0.34	0.16				
12	SA	0.21	0.45	0.14				
13	MC	0.64	0.41	0.41				
14	MC	0.44	0.29	0.53				
15	MC	0.40	0.30	0.22				
16	MC	0.17	0.25	0.15				+
17	MC	0.40	0.44	0.13				
18	SA	0.29	0.38	0.15				
19	MC	0.72	0.49	0.13				
20	TE	0.85	0.37	0.98				
21	MC	0.80	0.43	0.15				
22	MC	0.46	0.36	0.13				
23	MC	0.45	0.39	0.14				
24	MC	0.62	0.43	0.21				
25	MC	0.60	0.37	0.18				
26	SA	0.59	0.58	0.15				
27	TE	0.61	0.32	0.33				
28	MC	0.65	0.43	0.53				
29	MC	0.77	0.43	0.16				
30	MC	0.36	0.18	0.19		+		
31	MC	0.56	0.43	0.13				
32	MC	0.41	0.36	0.21				
33	SA	0.86	0.39	0.14				
34	SA	0.25	0.31	0.38				
35	MC	0.63	0.50	0.47				



Table 8-7 Item Analysis, Grade 3 Mathematics (cont.)

Item	Item Type	<i>p</i> -Value	Corr	Percent Omit	Flags			
					Corr	Distractor	Omit	<i>p</i> -Value
36	SA	0.46	0.57	0.20				
37	MC	0.67	0.48	0.21				
38	SA	0.67	0.57	0.21				
39	MC	0.79	0.42	0.17				
40	MC	0.76	0.38	0.18				
41	MC	0.50	0.55	0.23				
42	MC	0.67	0.48	0.19				

Table 8-8 Item Analysis, Grade 4 Mathematics

Item	Item Type	<i>p</i> -Value	Corr	Percent Omit	Flags			
					Corr	Distractor	Omit	<i>p</i> -Value
1	MC	0.30	0.49	0.28				
2	MC	0.44	0.32	0.09				
3	MC	0.56	0.49	0.07				
4	MC	0.56	0.54	0.13				
5	MC	0.72	0.46	0.09				
6	SA	0.26	0.53	0.22				
7	MC	0.63	0.43	0.14				
8	MC	0.63	0.50	0.24				
9	MC	0.75	0.39	0.09				
10	MC	0.35	0.31	0.14				
11	MC	0.59	0.37	0.16				
12	TE	0.70	0.37	0.12				
13	MC	0.47	0.57	0.12				
14	MC	0.41	0.26	0.16				
15	MC	0.38	0.31	0.45				
16	SA	0.25	0.43	0.29				
17	SA	0.40	0.58	0.15				
18	MC	0.38	0.42	0.16				
19	MC	0.58	0.33	0.14				
20	TE	0.24	0.55	0.52				
21	MC	0.59	0.40	0.13				
22	MC	0.82	0.29	0.13				
23	MC	0.43	0.45	0.23				
24	MC	0.81	0.40	0.27				
25	MC	0.84	0.35	0.09				
26	MC	0.47	0.40	0.09				
27	MC	0.48	0.49	0.18				
28	MC	0.35	0.58	0.15				
29	TE	0.26	0.45	0.17				
30	SA	0.58	0.48	0.19				
31	MC	0.34	0.40	0.34				
32	MC	0.41	0.24	0.19				
33	MC	0.52	0.60	0.16				
34	MC	0.63	0.47	0.21				
35	MC	0.61	0.48	0.13				

Table 8-8 Item Analysis, Grade 4 Mathematics (cont.)

Item	Item Type	<i>p</i> -Value	Corr	Percent Omit	Flags			
					Corr	Distractor	Omit	<i>p</i> -Value
36	MC	0.72	0.41	0.14				
37	SA	0.41	0.49	0.22				
38	MC	0.66	0.44	0.32				
39	MC	0.40	0.53	0.33				
40	MC	0.46	0.48	0.17				
41	MC	0.29	0.51	0.19				
42	SA	0.41	0.58	0.24				
43	MC	0.26	0.14	0.17	+	+		
44	SA	0.38	0.52	0.21				
45	MC	0.55	0.39	0.20				
46	MC	0.53	0.41	0.23				

Table 8-9 Item Analysis, Grade 5 Mathematics

Item	Item Type	<i>p</i> -Value	Corr	Percent Omit	Flags			
					Corr	Distractor	Omit	<i>p</i> -Value
1	MC	0.81	0.33	0.27				
2	MC	0.46	0.51	0.08				
3	MC	0.59	0.19	0.13		+		
4	MC	0.32	0.39	0.14				
5	MC	0.67	0.41	0.09				
6	TE	0.34	0.59	0.34				
7	SA	0.78	0.36	0.14				
8	MC	0.42	0.35	0.32				
9	MC	0.68	0.45	0.12				
10	MC	0.77	0.46	0.14				
11	MC	0.64	0.12	0.16	+			
12	TE	0.16	0.48	0.50				+
13	SA	0.38	0.55	0.19				
14	SA	0.43	0.53	0.44				
15	MC	0.53	0.43	0.32				
16	SA	0.28	0.55	0.33				
17	MC	0.47	0.32	0.20				
18	MC	0.43	0.47	0.26				
19	MC	0.43	0.46	0.20				
20	TE	0.60	0.46	0.25				
21	MC	0.53	0.56	0.19				
22	SA	0.45	0.58	0.28				
23	MC	0.65	0.40	0.34				
24	MC	0.55	0.32	0.24				
25	MC	0.26	0.17	0.11		+		
26	ESR	0.17	0.46	0.13				+
27	SA	0.15	0.45	0.44				+
28	SA	0.14	0.42	0.20				+
29	TE	0.46	0.49	0.31				
30	SA	0.54	0.52	0.29				
31	MC	0.56	0.31	0.28				
32	MC	0.51	0.46	0.16				
33	MC	0.51	0.58	0.14				
34	SA	0.34	0.50	0.21				
35	TE	0.23	0.50	1.23				

Table 8-9 Item Analysis, Grade 5 Mathematics (cont.)

Item	Item Type	<i>p</i> -Value	Corr	Percent Omit	Flags			
					Corr	Distractor	Omit	<i>p</i> -Value
36	MC	0.51	0.48	0.23				
37	ESR	0.10	0.26	0.19				+
38	MC	0.39	0.25	0.33		+		
39	MC	0.14	0.30	0.45				+
40	MC	0.60	0.45	0.19				
41	MC	0.27	0.39	0.19		+		
42	MC	0.43	0.41	0.17				
43	MC	0.55	0.46	0.21				
44	MC	0.80	0.24	0.18				
45	MC	0.65	0.37	0.21				
46	ESR	0.07	0.26	0.25				+

Table 8-10 Item Analysis, Grade 6 Mathematics

Item	Item Type	<i>p</i> -Value	Corr	Percent Omit	Flags			
					Corr	Distractor	Omit	<i>p</i> -Value
1	MC	0.51	0.46	0.10				
2	ESR	0.63	0.54	0.11				
3	MC	0.84	0.38	0.11				
4	TE	0.58	0.51	0.41				
5	MC	0.55	0.34	0.10				
6	SA	0.39	0.61	0.25				
7	MC	0.92	0.36	0.10				
8	MC	0.54	0.31	0.14				
9	TE	0.12	0.44	0.35				+
10	MC	0.33	0.15	0.10		+		
11	SA	0.41	0.50	0.13				
12	MC	0.39	0.27	0.18				
13	SA	0.18	0.48	0.17				+
14	MC	0.49	0.36	0.19				
15	MC	0.65	0.33	0.16				
16	MC	0.66	0.48	0.17				
17	MC	0.77	0.40	0.20				
18	MC	0.69	0.38	0.20				
19	SA	0.71	0.50	0.36				
20	MC	0.35	0.50	0.14				
21	MC	0.71	0.30	0.17				
22	MC	0.54	0.53	0.22				
23	MC	0.52	0.42	0.24				
24	SA	0.25	0.61	0.21				
25	MC	0.91	0.32	0.20				
26	MC	0.52	0.24	0.32				
27	TE	0.20	0.50	0.29				
28	MC	0.32	0.33	0.36		+		
29	ESR	0.06	0.22	0.43				+
30	ESR	0.41	0.38	0.14				
31	MC	0.66	0.49	0.19				
32	MC	0.63	0.46	0.22				
33	SA	0.11	0.50	0.51				+
34	MC	0.43	0.47	0.25				
35	MC	0.62	0.48	0.22				

Table 8-10 Item Analysis, Grade 6 Mathematics (cont.)

Item	Item Type	<i>p</i> -Value	Corr	Percent Omit	Flags			
					Corr	Distractor	Omit	<i>p</i> -Value
36	MC	0.44	0.45	0.22				
37	TE	0.39	0.50	0.55				
38	MC	0.39	0.24	0.35				
39	MC	0.40	0.24	0.39				
40	MC	0.35	0.53	0.47				
41	MC	0.34	0.12	0.42	+			
42	SA	0.24	0.53	0.51				
43	MC	0.34	0.22	0.27		+		
44	MC	0.37	0.25	0.35		+		
45	MC	0.57	0.45	0.29				
46	MC	0.35	0.49	0.28				

Table 8-11 Item Analysis, Grade 7 Mathematics

Item	Item Type	<i>p</i> -Value	Corr	Percent Omit	Flags			
					Corr	Distractor	Omit	<i>p</i> -Value
1	MC	0.23	0.41	0.06				
2	MC	0.46	0.55	0.06				
3	SA	0.57	0.60	0.12				
4	TE	0.21	0.44	0.11				
5	MC	0.49	0.42	0.09				
6	MC	0.38	0.44	0.09				
7	SA	0.14	0.49	0.40				+
8	MC	0.49	0.37	0.07				
9	MC	0.52	0.33	0.10				
10	MC	0.50	0.39	0.11				
11	MC	0.33	0.25	0.12		+		
12	MC	0.50	0.42	0.15				
13	MC	0.63	0.48	0.28				
14	SA	0.62	0.53	0.40				
15	TE	0.15	0.39	0.73				+
16	MC	0.29	0.52	0.32				
17	MC	0.36	0.47	0.27				
18	MC	0.43	0.22	0.22				
19	SA	0.26	0.62	0.71				
20	TE	0.64	0.18	0.79				
21	ESR	0.18	0.42	0.44				+
22	MC	0.29	0.22	0.36				
23	MC	0.65	0.33	0.43				
24	MC	0.78	0.43	0.45				
25	MC	0.34	0.31	0.45				
26	MC	0.65	0.39	0.36				
27	TE	0.09	0.34	0.53				+
28	MC	0.51	0.46	0.37				
29	ESR	0.14	0.35	0.40				+
30	SA	0.26	0.57	0.58				
31	MC	0.63	0.55	0.38				
32	MC	0.56	0.18	0.41				
33	MC	0.68	0.37	0.42				
34	MC	0.51	0.49	0.49				
35	SA	0.38	0.55	0.58				



Table 8-11 Item Analysis, Grade 7 Mathematics (cont.)

Item	Item Type	<i>p</i> -Value	Corr	Percent Omit	Flags			
					Corr	Distractor	Omit	<i>p</i> -Value
36	MC	0.56	0.34	0.53				
37	TE	0.13	0.51	0.85				+
38	MC	0.53	0.22	0.60				
39	MC	0.30	0.22	0.61		+		
40	TE	0.30	0.62	0.65				
41	SA	0.49	0.24	0.82				
42	MC	0.37	0.45	0.54				
43	MC	0.54	0.51	0.52				
44	SA	0.33	0.62	0.85				
45	MC	0.30	0.35	0.48		+		
46	MC	0.44	0.32	0.56				

Table 8-12 Item Analysis, Grade 8 Mathematics

Item	Item Type	<i>p</i> -Value	Corr	Percent Omit	Flags			
					Corr	Distractor	Omit	<i>p</i> -Value
1	MC	0.22	0.16	0.11				
2	SA	0.38	0.61	0.55				
3	MC	0.52	0.39	0.13				
4	MC	0.46	0.20	0.13				
5	MC	0.48	0.50	0.12		+		
6	SA	0.07	0.36	0.47				+
7	MC	0.37	0.27	0.12				
8	MC	0.50	0.49	0.09				
9	MC	0.45	0.31	0.14				
10	MC	0.42	0.29	0.18				
11	SA	0.05	0.31	0.75				+
12	MC	0.44	0.37	0.14				
13	SA	0.17	0.50	0.59				+
14	MC	0.37	0.40	0.19				
15	MC	0.63	0.28	0.28				
16	TE	0.49	0.50	0.49				
17	MC	0.59	0.46	0.30				
18	SA	0.10	0.40	0.45				+
19	SA	0.46	0.59	0.69				
20	MC	0.56	0.38	0.27				
21	TE	0.10	0.40	0.48				+
22	MC	0.41	0.43	0.28				
23	MC	0.35	0.26	0.26				
24	MC	0.65	0.41	0.38				
25	MC	0.23	0.17	0.31		+		
26	SA	0.20	0.48	1.57				
27	TE	0.57	0.50	1.25				
28	MC	0.55	0.57	0.43				
29	MC	0.48	0.24	0.31				
30	ESR	0.31	0.38	0.31				
31	SA	0.30	0.58	1.21				
32	TE	0.34	0.49	0.90				
33	MC	0.66	0.31	0.36				
34	MC	0.51	0.31	0.41				
35	MC	0.73	0.46	0.46				

Table 8-12 Item Analysis, Grade 8 Mathematics (cont.)

Item	Item Type	<i>p</i> -Value	Corr	Percent Omit	Flags			
					Corr	Distractor	Omit	<i>p</i> -Value
36	MC	0.55	0.46	0.53				
37	TE	0.35	0.37	0.94				
38	MC	0.76	0.44	0.55				
39	MC	0.40	0.36	0.64				
40	ESR	0.17	0.43	0.55				+
41	MC	0.36	0.28	0.57		+		
42	SA	0.52	0.43	1.09				
43	MC	0.66	0.54	0.49				
44	SA	0.23	0.56	1.30				
45	MC	0.66	0.39	0.43				
46	MC	0.57	0.48	0.46				

Table 8-13 Item Analysis, Grade 4 Science

Item	Item Type	<i>p</i> -Value	Corr	Percent Omit	Flags			
					Corr	Distractor	Omit	<i>p</i> -Value
1	MC	0.76	0.39	0.13				
2	MC	0.87	0.46	0.07				
3	MC	0.92	0.18	0.07				
4	MC	0.94	0.32	0.07				
5	MC	0.57	0.41	0.09				
6	MC	0.79	0.44	0.12				
7	MC	0.78	0.47	0.09				
8	MC	0.73	0.46	0.13				
9	MC	0.35	0.21	0.40		+		
10	MC	0.66	0.50	0.14				
11	TE	0.43	0.24	0.50				
12	MC	0.61	0.49	0.12				
13	MC	0.89	0.47	0.08				
14	MC	0.61	0.37	0.11				
15	MC	0.87	0.28	0.07				
16	MC	0.79	0.38	0.09				
17	MC	0.79	0.45	0.23				
18	MC	0.58	0.37	0.14				
19	MC	0.77	0.29	0.11				
20	MC	0.53	0.42	0.12				
21	MC	0.85	0.37	0.12				
22	MC	0.91	0.25	0.10				
23	MC	0.84	0.42	0.07				
24	MC	0.63	0.35	0.07				
25	MC	0.47	0.16	0.10		+		
26	MC	0.60	0.34	0.12				
27	MC	0.64	0.39	0.18				
28	MC	0.56	0.29	0.08				
29	MC	0.76	0.31	0.14				
30	MC	0.67	0.41	0.29				
31	MC	0.71	0.36	0.14				
32	MC	0.75	0.49	0.14				
33	MC	0.71	0.48	0.16				
34	MC	0.78	0.45	0.10				
35	MC	0.82	0.46	0.13				

Table 8-13 Item Analysis, Grade 4 Science (cont.)

Item	Item Type	<i>p</i> -Value	Corr	Percent Omit	Flags			
					Corr	Distractor	Omit	<i>p</i> -Value
36	MC	0.41	0.20	0.09				
37	MC	0.69	0.35	0.10				
38	MC	0.44	0.33	0.24				
39	MC	0.56	0.45	0.18				
40	MC	0.77	0.47	0.13				

Table 8-14 Item Analysis, Grade 8 Science

Item	Item Type	<i>p</i> -Value	Corr	Percent Omit	Flags			
					Corr	Distractor	Omit	<i>p</i> -Value
1	MC	0.86	0.41	0.05				
2	MC	0.87	0.40	0.06				
3	MC	0.91	0.40	0.08				
4	MC	0.78	0.38	0.09				
5	MC	0.86	0.30	0.08				
6	MC	0.85	0.40	0.08				
7	MC	0.77	0.44	0.07				
8	MC	0.84	0.42	0.13				
9	MC	0.62	0.45	0.41				
10	MC	0.45	0.19	0.20				
11	MC	0.66	0.43	0.18				
12	MC	0.45	0.24	0.20				
13	MC	0.76	0.29	0.07				
14	MC	0.91	0.33	0.12				
15	MC	0.75	0.35	0.09				
16	MC	0.75	0.36	0.06				
17	MC	0.62	0.29	0.18				
18	MC	0.67	0.42	0.14				
19	MC	0.86	0.51	0.14				
20	MC	0.66	0.45	0.15				
21	MC	0.77	0.35	0.13				
22	MC	0.71	0.36	0.12				
23	MC	0.74	0.48	0.18				
24	MC	0.54	0.24	0.15				
25	MC	0.54	0.27	0.17				
26	MC	0.78	0.50	0.15				
27	MC	0.62	0.33	0.15				
28	MC	0.64	0.46	0.12				
29	MC	0.71	0.45	0.17				
30	MC	0.66	0.51	0.36				
31	MC	0.49	0.32	0.29				
32	MC	0.29	0.17	0.24				
33	MC	0.36	0.23	0.23		+		
34	MC	0.85	0.46	0.18				
35	MC	0.72	0.42	0.18				

Table 8-14 Item Analysis, Grade 8 Science (cont.)

Item	Item Type	<i>p</i> -Value	Corr	Percent Omit	Flags			
					Corr	Distractor	Omit	<i>p</i> -Value
36	MC	0.66	0.49	0.18				
37	MC	0.84	0.52	0.17				
38	MC	0.67	0.36	0.27				
39	MC	0.70	0.46	0.21				
40	MC	0.74	0.45	0.21				

Table 8-15 Item Analysis, Grade 4 Social Studies

Item	Item Type	<i>p</i> -Value	Corr	Percent Omit	Flags			
					Corr	Distractor	Omit	<i>p</i> -Value
1	MC	0.89	0.34	0.04				
2	MC	0.81	0.44	0.13				
3	MC	0.86	0.33	0.06				
4	MC	0.57	0.40	0.15				
5	MC	0.52	0.39	0.15				
6	MC	0.75	0.38	0.14				
7	MC	0.84	0.42	0.09				
8	MC	0.84	0.35	0.30				
9	MC	0.81	0.45	0.08				
10	MC	0.69	0.41	0.11				
11	MC	0.79	0.45	0.09				
12	MC	0.44	0.24	0.12				
13	MC	0.55	0.25	0.09				
14	MC	0.42	0.23	0.11		+		
15	MC	0.51	0.36	0.10				
16	MC	0.74	0.45	0.27				
17	MC	0.78	0.50	0.15				
18	MC	0.86	0.46	0.09				
19	MC	0.74	0.31	0.10				
20	MC	0.81	0.31	0.05				
21	MC	0.83	0.45	0.09				
22	MC	0.59	0.36	0.10				
23	MC	0.73	0.46	0.10				
24	MC	0.54	0.36	0.12				
25	MC	0.59	0.39	0.39				
26	MC	0.56	0.46	0.14				
27	MC	0.81	0.42	0.17				
28	MC	0.68	0.19	0.11				
29	MC	0.85	0.44	0.34				
30	MC	0.70	0.44	0.14				
31	MC	0.64	0.46	0.14				
32	MC	0.64	0.45	0.11				
33	MC	0.67	0.40	0.15				
34	MC	0.48	0.39	0.19				
35	MC	0.50	0.39	0.19				
36	MC	0.82	0.46	0.13				
37	MC	0.54	0.34	0.13				
38	MC	0.82	0.47	0.13				



Table 8-16 Item Analysis, Grade 8 Social Studies

Item	Item Type	<i>p</i> -Value	Corr	Percent Omit	Flags			
					Corr	Distractor	Omit	<i>p</i> -Value
1	MC	0.84	0.47	0.09				
2	MC	0.78	0.40	0.10				
3	MC	0.82	0.44	0.10				
4	MC	0.82	0.41	0.10				
5	MC	0.85	0.47	0.14				
6	MC	0.75	0.28	0.11				
7	MC	0.91	0.38	0.13				
8	MC	0.78	0.42	0.15				
9	MC	0.56	0.36	0.15				
10	MC	0.74	0.47	0.10				
11	MC	0.63	0.30	0.11				
12	MC	0.67	0.51	0.15				
13	MC	0.81	0.50	0.19				
14	MC	0.69	0.42	0.18				
15	MC	0.73	0.43	0.19				
16	MC	0.61	0.56	0.24				
17	MC	0.49	0.34	0.15		+		
18	MC	0.66	0.42	0.15				
19	MC	0.58	0.46	0.14				
20	MC	0.65	0.50	0.19				
21	MC	0.62	0.38	0.15				
22	MC	0.45	0.36	0.13		+		
23	MC	0.41	0.26	0.16				
24	MC	0.67	0.56	0.11				
25	MC	0.74	0.47	0.13				
26	MC	0.70	0.37	0.16				
27	MC	0.85	0.45	0.14				
28	MC	0.66	0.40	0.16				
29	MC	0.59	0.40	0.17				
30	MC	0.59	0.42	0.22				
31	MC	0.76	0.34	0.17				
32	MC	0.52	0.44	0.16				
33	MC	0.71	0.39	0.25				
34	MC	0.56	0.33	0.18				
35	MC	0.63	0.48	0.19				
36	MC	0.69	0.55	0.18				
37	MC	0.71	0.49	0.19				
38	MC	0.59	0.42	0.19				
39	MC	0.51	0.32	0.17				
40	MC	0.81	0.49	0.20				

Table 8-17 Item Analysis, Grade 10 Social Studies

Item	Item Type	<i>p</i> -Value	Corr	Percent Omit	Flags			
					Corr	Distractor	Omit	<i>p</i> -Value
1	MC	0.85	0.35	0.08				
2	MC	0.80	0.43	0.18				
3	MC	0.67	0.39	0.23				
4	MC	0.78	0.35	0.13				
5	MC	0.55	0.39	0.10				
6	MC	0.86	0.44	0.09				
7	MC	0.77	0.41	0.13				
8	MC	0.66	0.33	0.17				
9	MC	0.67	0.30	0.15				
10	MC	0.71	0.30	0.16				
11	MC	0.64	0.40	0.20				
12	MC	0.50	0.30	0.31				
13	MC	0.58	0.38	0.32				
14	MC	0.65	0.37	0.20				
15	MC	0.75	0.33	0.19				
16	MC	0.55	0.29	0.24				
17	MC	0.55	0.39	0.24				
18	MC	0.23	0.13	0.27	+	+		
19	MC	0.73	0.49	0.29				
20	MC	0.67	0.40	0.31				
21	MC	0.27	0.33	0.31				
22	MC	0.48	0.26	0.33				
23	MC	0.66	0.51	0.35				
24	MC	0.52	0.44	0.36				
25	MC	0.60	0.53	0.39				
26	MC	0.77	0.47	0.19				
27	MC	0.67	0.47	0.28				
28	MC	0.85	0.35	0.26				
29	MC	0.70	0.47	0.30				
30	MC	0.63	0.46	0.34				
31	MC	0.45	0.39	0.27				
32	MC	0.69	0.49	0.28				
33	MC	0.67	0.50	0.30				
34	MC	0.81	0.52	0.39				
35	MC	0.80	0.52	0.35				

Table 8-17 Item Analysis, Grade 10 Social Studies (cont.)

Item	Item Type	<i>p</i> -Value	Corr	Percent Omit	Flags			
					Corr	Distractor	Omit	<i>p</i> -Value
36	MC	0.73	0.34	0.36				
37	MC	0.51	0.41	0.42				
38	MC	0.66	0.43	0.41				
39	MC	0.67	0.50	0.37				
40	MC	0.53	0.34	0.40				
41	MC	0.56	0.42	0.57				
42	MC	0.51	0.25	0.48				
43	MC	0.58	0.46	0.56				
44	MC	0.53	0.35	0.56				
45	MC	0.67	0.47	0.46				
46	MC	0.71	0.43	0.49				
47	MC	0.45	0.24	0.46		+		
48	MC	0.68	0.48	0.46				
49	MC	0.71	0.44	0.46				
50	MC	0.80	0.38	0.49				

Table 8-18 Raw Score Descriptive Statistics

Content	Grade	N Count	Mean Raw Score	Test Difficulty	Raw Score SD	Skewness	Kurtosis	Min Obtained	Max Obtained	Max Possible	Alpha	SEM
<b>English Language Arts</b>	3	63946	26.89	0.58	9.31	-0.10	-0.84	1	50	53	0.87	3.31
	4	64423	30.26	0.60	10.09	-0.17	-0.75	0	55	56	0.89	3.33
	5	62995	29.20	0.60	9.01	-0.18	-0.66	1	53	56	0.87	3.25
	6	62754	30.89	0.61	9.15	-0.28	-0.44	2	55	56	0.87	3.31
	7	63091	31.82	0.61	10.59	-0.21	-0.79	1	56	56	0.89	3.56
	8	62109	31.19	0.61	10.40	-0.17	-0.73	0	56	56	0.89	3.49
<b>Mathematics</b>	3	64066	24.09	0.58	9.04	-0.14	-0.93	0	42	42	0.91	2.69
	4	64533	22.77	0.50	10.28	0.27	-0.96	0	46	46	0.92	2.86
	5	63152	20.64	0.45	9.56	0.40	-0.73	0	46	46	0.91	2.80
	6	62847	21.71	0.47	9.30	0.35	-0.70	1	46	46	0.91	2.82
	7	63200	19.05	0.42	9.52	0.57	-0.52	0	46	46	0.91	2.84
	8	62175	19.25	0.42	9.36	0.43	-0.63	0	46	46	0.91	2.84
<b>Science</b>	4	64520	27.76	0.70	7.33	-0.58	-0.47	1	40	40	0.88	2.54
	8	62113	27.88	0.70	7.44	-0.72	-0.22	0	40	40	0.88	2.53
<b>Social Studies</b>	4	64512	26.19	0.69	7.33	-0.58	-0.48	3	38	38	0.88	2.49
	8	62079	27.08	0.68	8.40	-0.52	-0.69	1	40	40	0.91	2.58
	10	63764	31.76	0.64	10.19	-0.33	-0.79	0	50	50	0.91	3.01

Table 8-19 Raw Score Descriptive Statistics by Gender

Content	Grade	Male					Female				
		N Count	Mean Raw Score	Test Difficulty	Raw Score SD	Alpha	N Count	Mean Raw Score	Test Difficulty	Raw Score SD	Alpha
English Language Arts	3	32521	26.10	0.57	9.21	0.87	31341	27.75	0.60	9.32	0.87
	4	32946	29.28	0.58	10.07	0.89	31415	31.30	0.61	9.99	0.89
	5	32281	28.06	0.58	8.91	0.87	30671	30.41	0.62	8.94	0.87
	6	32002	29.57	0.59	9.12	0.87	30696	32.29	0.63	8.96	0.87
	7	32317	30.34	0.59	10.57	0.89	30687	33.41	0.63	10.35	0.88
	8	31869	29.40	0.58	10.34	0.89	30143	33.13	0.64	10.07	0.88
Mathematics	3	32601	24.35	0.58	9.24	0.92	31411	23.82	0.57	8.81	0.91
	4	32997	23.43	0.51	10.55	0.93	31476	22.10	0.48	9.94	0.92
	5	32348	20.85	0.45	9.84	0.92	30762	20.43	0.45	9.25	0.91
	6	32032	21.73	0.47	9.49	0.91	30755	21.70	0.47	9.08	0.90
	7	32383	19.17	0.42	9.73	0.92	30739	18.95	0.41	9.29	0.91
	8	31916	18.91	0.41	9.57	0.91	30184	19.65	0.43	9.11	0.90
Science	4	32989	27.87	0.70	7.45	0.89	31473	27.65	0.69	7.19	0.88
	8	31888	27.61	0.69	7.74	0.89	30162	28.18	0.71	7.07	0.87
Social Studies	4	32985	26.05	0.69	7.45	0.89	31468	26.35	0.69	7.19	0.88
	8	31869	26.73	0.67	8.68	0.91	30146	27.48	0.69	8.06	0.90
	10	32446	32.04	0.65	10.55	0.92	31158	31.54	0.63	9.73	0.90

Table 8-20 Raw Score Descriptive Statistics for English Language Arts by Race/Ethnicity

Race/Ethnicity	Grade	N Count	Mean Raw Score	Test Difficulty	Raw Score SD	Alpha
White	3	42305	29.00	0.62	8.67	0.86
	4	42981	32.64	0.64	9.25	0.87
	5	42203	31.22	0.64	8.33	0.85
	6	43001	32.83	0.65	8.41	0.85
	7	43815	33.84	0.65	9.93	0.88
	8	43444	33.09	0.64	9.84	0.88
African American	3	7125	19.34	0.44	8.26	0.83
	4	7010	21.58	0.43	9.26	0.87
	5	6786	21.69	0.45	8.40	0.84
	6	6480	23.18	0.47	8.65	0.84
	7	6298	23.23	0.45	9.80	0.86
	8	6147	22.97	0.46	9.42	0.86
Hispanic	3	8537	23.40	0.51	8.66	0.85
	4	8527	26.21	0.51	9.43	0.87
	5	8382	25.91	0.53	8.37	0.84
	6	7873	27.22	0.54	8.71	0.85
	7	7697	27.85	0.53	10.07	0.87
	8	7569	27.43	0.53	9.77	0.87
Asian	3	2557	26.68	0.57	9.30	0.87
	4	2500	29.98	0.58	9.98	0.89
	5	2522	28.77	0.58	9.19	0.87
	6	2448	31.47	0.61	9.33	0.87
	7	2419	33.11	0.63	10.36	0.88
	8	2329	32.62	0.62	10.32	0.89
American Indian	3	767	22.75	0.50	8.63	0.85
	4	837	25.87	0.51	9.25	0.86
	5	803	24.99	0.52	8.13	0.84
	6	791	26.48	0.53	8.33	0.83
	7	791	26.70	0.52	10.05	0.87
	8	768	26.23	0.52	9.23	0.86
Two or More	3	2568	26.68	0.58	9.43	0.88
	4	2485	29.60	0.58	9.94	0.89
	5	2250	28.36	0.58	8.86	0.86
	6	2087	30.04	0.60	9.09	0.87
	7	1977	30.92	0.59	10.47	0.88
	8	1753	30.20	0.59	10.47	0.89

Table 8-21 Raw Score Descriptive Statistics for Mathematics by Race/Ethnicity

Race/Ethnicity	Grade	N Count	Mean Raw Score	Test Difficulty	Raw Score SD	Alpha
White	3	42324	26.25	0.63	8.36	0.90
	4	43002	25.29	0.55	9.81	0.91
	5	42218	22.87	0.50	9.29	0.91
	6	43011	23.74	0.52	8.99	0.90
	7	43819	21.07	0.46	9.36	0.91
	8	43449	21.16	0.46	9.12	0.90
African American	3	7145	16.60	0.40	7.72	0.87
	4	7018	14.25	0.31	7.33	0.86
	5	6825	12.88	0.28	6.63	0.84
	6	6498	14.02	0.31	6.63	0.83
	7	6329	11.30	0.25	5.95	0.80
	8	6147	11.53	0.25	6.37	0.83
Hispanic	3	8605	20.22	0.48	8.36	0.89
	4	8591	17.93	0.39	8.77	0.89
	5	8451	16.53	0.36	7.91	0.88
	6	7918	17.41	0.38	7.73	0.87
	7	7756	14.60	0.32	7.56	0.87
	8	7620	15.01	0.33	7.69	0.87
Asian	3	2599	24.39	0.58	9.37	0.92
	4	2540	22.93	0.50	11.06	0.94
	5	2561	21.31	0.46	10.05	0.92
	6	2475	22.99	0.50	9.86	0.92
	7	2445	20.16	0.44	10.22	0.92
	8	2355	20.85	0.45	10.20	0.92
American Indian	3	768	19.96	0.48	8.40	0.89
	4	836	17.71	0.39	8.86	0.90
	5	802	15.60	0.34	7.40	0.86
	6	793	17.31	0.38	7.46	0.86
	7	790	14.07	0.31	7.57	0.87
	8	768	14.33	0.31	7.53	0.87
Two or More	3	2569	23.28	0.56	9.15	0.91
	4	2484	21.64	0.47	10.04	0.92
	5	2251	19.09	0.42	9.40	0.91
	6	2090	20.50	0.45	9.05	0.90
	7	1980	17.64	0.39	9.18	0.91
	8	1758	18.02	0.39	9.16	0.90

Table 8-22 Raw Score Descriptive Statistics for Science by Race/Ethnicity

Race/Ethnicity	Grade	N Count	Mean Raw Score	Test Difficulty	Raw Score SD	Alpha
White	4	43008	29.70	0.74	6.37	0.85
	8	43445	29.62	0.74	6.50	0.86
African American	4	7009	20.77	0.52	7.18	0.85
	8	6109	20.47	0.52	7.54	0.86
Hispanic	4	8585	24.56	0.61	7.18	0.86
	8	7620	24.49	0.61	7.40	0.86
Asian	4	2541	26.83	0.67	7.37	0.88
	8	2353	27.66	0.69	7.30	0.88
American Indian	4	838	24.54	0.61	7.04	0.85
	8	765	24.99	0.63	7.31	0.86
Two or More	4	2480	27.11	0.68	7.15	0.87
	8	1755	27.23	0.68	7.52	0.88

Table 8-23 Raw Score Descriptive Statistics for Social Studies by Race/Ethnicity

Race/Ethnicity	Grade	N Count	Mean Raw Score	Test Difficulty	Raw Score SD	Alpha
White	4	42999	28.05	0.74	6.40	0.86
	8	43432	28.89	0.72	7.65	0.89
	10	46472	33.65	0.68	9.47	0.90
African American	4	7004	19.50	0.52	7.40	0.86
	8	6103	19.39	0.49	7.98	0.87
	10	5397	22.34	0.46	9.29	0.88
Hispanic	4	8583	23.17	0.61	7.22	0.86
	8	7603	23.63	0.59	8.06	0.88
	10	6993	27.31	0.55	9.65	0.89
Asian	4	2541	25.29	0.67	7.43	0.88
	8	2354	27.19	0.68	8.31	0.91
	10	2438	32.11	0.65	9.94	0.91
American Indian	4	838	22.76	0.60	7.03	0.86
	8	766	23.01	0.58	8.07	0.88
	10	707	26.96	0.55	9.77	0.89
Two or More	4	2484	25.73	0.68	7.28	0.88
	8	1756	26.09	0.65	8.46	0.90
	10	1594	31.26	0.63	10.26	0.91



Table 8-24 Raw Score Descriptive Statistics by Socioeconomic Status

Content	Grade	Economically Disadvantaged					Not Economically Disadvantaged				
		N Count	Mean Raw Score	Test Difficulty	Raw Score SD	Alpha	N Count	Mean Raw Score	Test Difficulty	Raw Score SD	Alpha
English Language Arts	3	26427	23.14	0.51	8.80	0.85	37432	29.57	0.63	8.71	0.86
	4	26030	26.04	0.51	9.63	0.88	38310	33.15	0.65	9.33	0.88
	5	25180	25.35	0.52	8.54	0.85	37766	31.77	0.65	8.36	0.86
	6	23567	26.79	0.53	8.84	0.86	39113	33.38	0.66	8.40	0.85
	7	22842	27.20	0.52	10.19	0.87	40155	34.48	0.66	9.86	0.88
	8	21822	26.68	0.52	9.90	0.87	40188	33.67	0.65	9.79	0.88
Mathematics	3	26533	20.32	0.49	8.51	0.89	37477	26.76	0.64	8.42	0.90
	4	26095	18.31	0.40	8.95	0.90	38376	25.82	0.56	10.01	0.92
	5	25273	16.53	0.36	8.11	0.88	37835	23.40	0.51	9.46	0.91
	6	23630	17.42	0.38	7.91	0.87	39155	24.31	0.53	9.10	0.90
	7	22923	14.68	0.32	7.66	0.87	40196	21.56	0.47	9.56	0.91
	8	21903	14.87	0.33	7.79	0.87	40194	21.66	0.47	9.27	0.91
Science	4	26098	24.75	0.62	7.44	0.87	38363	29.81	0.75	6.49	0.86
	8	21849	24.46	0.61	7.75	0.88	40198	29.75	0.74	6.54	0.86
Social Studies	4	26082	23.16	0.61	7.43	0.87	38367	28.27	0.74	6.48	0.87
	8	21831	23.14	0.58	8.39	0.89	40183	29.25	0.73	7.57	0.89
	10	20058	26.92	0.55	9.94	0.90	43543	34.04	0.68	9.45	0.91

Table 8-25 Raw Score Descriptive Statistics by Disability

Content	Grade	Disabled					Not Disabled				
		N Count	Mean Raw Score	Test Difficulty	Raw Score SD	Alpha	N Count	Mean Raw Score	Test Difficulty	Raw Score SD	Alpha
English Language Arts	3	7109	21.00	0.47	8.74	0.86	56750	27.65	0.60	9.10	0.87
	4	7339	22.77	0.46	9.51	0.88	57001	31.24	0.61	9.73	0.88
	5	7332	21.23	0.45	8.24	0.84	55614	30.26	0.62	8.56	0.86
	6	6919	21.60	0.44	8.32	0.84	55761	32.06	0.63	8.56	0.85
	7	7017	21.10	0.42	8.95	0.84	55980	33.18	0.63	9.98	0.88
	8	6814	20.76	0.42	8.48	0.84	55196	32.50	0.63	9.85	0.88
Mathematics	3	7119	18.34	0.44	9.04	0.91	56891	24.81	0.59	8.77	0.91
	4	7361	16.17	0.35	8.87	0.90	57110	23.63	0.51	10.14	0.92
	5	7341	14.00	0.31	7.68	0.88	55767	21.52	0.47	9.44	0.91
	6	6934	13.92	0.30	7.19	0.85	55851	22.69	0.49	9.06	0.90
	7	7024	11.55	0.25	6.44	0.83	56095	20.00	0.44	9.42	0.91
	8	6814	11.22	0.25	6.39	0.83	55283	20.26	0.44	9.19	0.90
Science	4	7365	23.14	0.58	7.71	0.87	57096	28.36	0.71	7.06	0.87
	8	6803	20.68	0.52	7.78	0.87	55244	28.77	0.72	6.88	0.87
Social Studies	4	7360	21.13	0.56	7.76	0.88	57089	26.86	0.71	7.00	0.88
	8	6798	18.96	0.48	7.87	0.87	55216	28.10	0.70	7.89	0.90
	10	6555	22.57	0.46	9.42	0.88	57046	32.86	0.66	9.69	0.91

Table 8-26 Raw Score Descriptive Statistics by English Language Proficiency

Content	Grade	Limited English Proficient					Fully English Proficient				
		N Count	Mean Raw Score	Test Difficulty	Raw Score SD	Alpha	N Count	Mean Raw Score	Test Difficulty	Raw Score SD	Alpha
English Language Arts	3	5718	21.86	0.47	7.87	0.81	58141	27.41	0.59	9.28	0.87
	4	4849	22.96	0.45	7.82	0.81	59491	30.87	0.61	10.00	0.89
	5	3654	21.25	0.43	6.67	0.75	59292	29.69	0.61	8.90	0.87
	6	2875	21.37	0.42	6.91	0.75	59805	31.36	0.62	8.98	0.86
	7	2637	20.97	0.40	7.36	0.75	60360	32.31	0.62	10.44	0.88
	8	2591	21.21	0.41	7.37	0.76	59419	31.65	0.62	10.27	0.89
Mathematics	3	5865	18.95	0.45	7.97	0.88	58145	24.61	0.59	8.98	0.91
	4	4975	15.34	0.33	7.15	0.84	59496	23.40	0.51	10.26	0.92
	5	3789	13.41	0.29	6.07	0.80	59319	21.11	0.46	9.55	0.91
	6	2970	13.47	0.29	5.51	0.75	59815	22.13	0.48	9.25	0.91
	7	2751	10.74	0.24	4.85	0.71	60368	19.44	0.42	9.50	0.91
	8	2695	11.24	0.25	5.38	0.76	59402	19.63	0.43	9.34	0.91
Science	4	4970	21.96	0.55	6.46	0.81	59491	28.25	0.71	7.18	0.88
	8	2691	19.56	0.49	6.45	0.80	59356	28.26	0.71	7.25	0.88
Social Studies	4	4969	20.53	0.54	6.53	0.82	59480	26.68	0.70	7.19	0.88
	8	2688	18.25	0.46	6.65	0.81	59326	27.50	0.69	8.24	0.90
	10	2094	20.07	0.41	7.16	0.79	61507	32.20	0.65	10.01	0.91

Table 8-27 Scale Score Descriptive Statistics

Content	Grade	N Count	Mean	SD	Skewness	Kurtosis	Min	Max	LOSS	HOSS
<b>English Language Arts</b>	3	63946	559.12	46.93	0.00	0.10	330	755	330	900
	4	64423	585.26	52.44	-0.19	0.57	340	911	340	930
	5	62995	603.24	51.00	-0.26	0.61	350	828	350	940
	6	62754	614.59	49.82	-0.14	0.70	360	910	360	950
	7	63091	626.80	59.14	-0.05	1.22	370	960	370	960
	8	62109	637.69	61.61	-0.01	0.67	380	970	380	970
<b>Mathematics</b>	3	64066	555.03	48.63	-0.39	2.08	360	760	360	760
	4	64533	574.33	54.92	-0.71	1.83	405	800	405	800
	5	63152	599.73	51.00	-0.76	1.67	430	830	430	830
	6	62847	612.93	54.81	-0.55	1.29	440	870	440	870
	7	63200	627.48	58.65	-0.70	1.37	450	880	450	880
	8	62175	641.11	59.36	-0.92	1.51	470	890	470	890
<b>Science</b>	4	64520	399.27	53.16	0.10	1.29	190	600	190	600
	8	62113	594.12	51.25	-0.40	1.73	390	770	390	770
<b>Social Studies</b>	4	64512	397.05	51.71	-0.02	1.46	200	570	200	570
	8	62079	597.60	54.26	0.07	1.51	420	780	420	780
	10	63764	696.92	56.56	-0.48	1.53	490	890	490	890

Table 8-28 Scale Score Descriptive Statistics by Gender

Content	Grade	Male					Female				
		N Count	Mean	SD	Min	Max	N Count	Mean	SD	Min	Max
English Language Arts	3	32569	554.91	46.26	330	755	31377	563.49	47.23	330	755
	4	32975	580.05	52.27	340	792	31448	590.72	52.06	340	911
	5	32305	596.87	50.38	350	800	30690	609.96	50.79	350	828
	6	32028	607.29	49.25	360	893	30726	622.20	49.27	360	910
	7	32352	618.17	58.78	370	960	30739	635.89	58.14	370	960
	8	31914	626.89	60.78	380	926	30195	649.11	60.42	380	970
Mathematics	3	32629	556.46	50.74	360	760	31437	553.54	46.30	360	760
	4	33028	576.76	57.32	405	800	31505	571.79	52.16	405	800
	5	32373	599.99	53.02	430	830	30779	599.46	48.78	430	830
	6	32061	612.27	56.78	440	870	30786	613.61	52.66	440	870
	7	32417	627.21	61.00	450	880	30783	627.78	56.07	450	880
	8	31956	637.89	62.23	470	890	30219	644.52	55.96	470	890
Science	4	33022	399.85	54.85	190	600	31498	398.67	51.31	190	600
	8	31921	592.21	53.98	390	770	30192	596.14	48.10	390	770
Social Studies	4	33016	396.20	52.59	200	570	31496	397.94	50.75	200	570
	8	31900	595.44	56.68	420	780	30179	599.89	51.48	420	780
	10	32530	697.95	60.00	490	890	31234	695.86	52.70	490	890

Table 8-29 Scale Score Descriptive Statistics for English Language Arts by Race/Ethnicity

Race/Ethnicity	Grade	N Count	Mean	SD	Min	Max
White	3	42335	569.50	43.64	330	755
	4	43005	597.57	47.49	356	911
	5	42228	614.71	46.50	350	828
	6	43028	625.03	45.98	360	910
	7	43857	637.73	55.27	370	960
	8	43481	648.90	58.28	380	970
African-American	3	7151	520.89	43.23	330	702
	4	7038	539.31	51.97	340	757
	5	6799	559.58	51.06	350	763
	6	6502	573.00	48.01	360	789
	7	6317	579.09	58.15	370	899
	8	6186	588.82	57.04	380	797
Hispanic	3	8557	541.97	43.16	330	733
	4	8543	564.60	48.80	340	792
	5	8389	585.00	47.05	350	781
	6	7886	594.67	46.31	360	787
	7	7717	605.50	55.41	370	960
	8	7583	615.02	57.14	380	859
Asian	3	2565	558.54	47.57	371	755
	4	2509	584.21	51.90	407	792
	5	2524	601.15	52.75	350	785
	6	2453	617.37	52.20	390	893
	7	2422	634.64	59.23	370	960
	8	2332	645.73	62.71	380	970
American Indian	3	767	538.64	42.74	418	677
	4	838	562.28	47.72	368	696
	5	803	579.65	45.98	409	718
	6	793	591.14	44.44	454	763
	7	793	599.77	55.16	422	905
	8	771	608.86	53.00	453	772
Two or More	3	2571	558.23	47.77	376	743
	4	2490	582.15	51.79	355	885
	5	2252	598.86	50.45	381	785
	6	2092	609.87	49.37	399	823
	7	1985	621.34	58.64	370	960
	8	1756	632.27	62.78	401	970

Table 8-30 Scale Score Descriptive Statistics for Mathematics by Race/Ethnicity

Race/Ethnicity	Grade	N Count	Mean	SD	Min	Max
White	3	42346	566.15	43.63	360	760
	4	43021	587.78	47.43	405	800
	5	42238	611.88	44.58	430	830
	6	43044	625.41	48.57	440	870
	7	43856	640.54	52.14	450	880
	8	43477	653.70	51.96	470	890
African-American	3	7162	515.12	48.26	360	702
	4	7042	525.06	57.34	405	800
	5	6842	553.17	54.31	430	753
	6	6516	561.82	56.36	440	870
	7	6349	572.98	59.26	450	801
	8	6181	585.31	63.92	470	781
Hispanic	3	8618	535.60	46.50	360	760
	4	8599	550.31	54.32	405	800
	5	8456	579.74	48.58	430	760
	6	7923	588.92	51.60	440	870
	7	7770	600.78	57.22	450	880
	8	7629	616.19	58.54	470	801
Asian	3	2601	558.43	52.98	360	760
	4	2544	576.95	59.92	405	800
	5	2562	604.00	53.35	430	830
	6	2476	620.37	57.88	440	870
	7	2446	635.84	57.79	450	880
	8	2355	651.36	62.24	470	890
American Indian	3	768	533.90	46.64	360	679
	4	837	549.37	55.06	405	800
	5	802	575.80	45.57	430	729
	6	795	587.94	50.28	440	718
	7	792	597.05	56.58	450	736
	8	771	610.90	59.49	470	763
Two or More	3	2571	550.93	49.60	360	760
	4	2490	569.95	53.90	405	800
	5	2252	591.94	52.20	430	830
	6	2093	606.81	54.18	440	870
	7	1987	619.74	58.65	450	880
	8	1762	633.79	60.83	470	890

Table 8-31 Scale Score Descriptive Statistics for Science by Race/Ethnicity

Race/Ethnicity	Grade	N Count	Mean	SD	Min	Max
White	4	43028	412.45	48.63	190	600
	8	43469	605.36	46.15	390	770
African-American	4	7032	352.19	49.11	190	600
	8	6139	546.13	52.34	390	770
Hispanic	4	8591	377.05	47.97	190	600
	8	7627	572.03	47.99	390	770
Asian	4	2547	393.10	55.39	190	600
	8	2353	593.17	51.06	390	770
American Indian	4	839	377.40	46.85	190	600
	8	768	573.84	46.09	390	725
Two or More	4	2483	394.87	50.45	190	600
	8	1757	589.66	51.77	390	770

Table 8-32 Scale Score Descriptive Statistics for Social Studies by Race/Ethnicity

Race/Ethnicity	Grade	N Count	Mean	SD	Min	Max
White	4	43022	409.45	46.92	200	570
	8	43457	608.52	51.20	420	780
	10	46533	706.97	51.96	490	890
African-American	4	7029	352.21	49.99	200	570
	8	6129	551.38	49.63	420	780
	10	5459	644.54	59.12	490	841
Hispanic	4	8589	376.38	47.87	200	570
	8	7611	576.02	47.44	420	780
	10	7021	673.46	53.85	490	890
Asian	4	2546	391.13	53.85	200	570
	8	2355	598.68	55.94	420	780
	10	2443	698.92	55.16	490	890
American Indian	4	838	375.49	46.04	200	570
	8	769	572.47	48.20	420	780
	10	709	671.72	55.44	490	890
Two or More	4	2488	394.04	50.92	200	570
	8	1758	591.95	53.90	420	780
	10	1599	694.44	57.37	490	890



Table 8-33 Scale Score Descriptive Statistics by Socioeconomic Status

Content	Grade	Economically Disadvantaged					Not Economically Disadvantaged				
		N Count	Mean	SD	Min	Max	N Count	Mean	SD	Min	Max
English Language Arts	3	26487	540.43	44.15	330	755	37459	572.33	44.28	330	755
	4	26089	563.46	50.51	340	885	38334	600.09	48.39	361	911
	5	25211	581.58	48.97	350	792	37784	617.70	47.06	350	828
	6	23618	592.44	47.33	360	855	39136	627.96	46.41	360	910
	7	22904	601.46	56.48	370	960	40187	641.25	55.65	370	960
	8	21888	611.01	58.20	380	926	40221	652.21	58.49	380	970
Mathematics	3	26574	535.53	47.22	360	760	37492	568.85	44.73	360	760
	4	26141	551.38	55.18	405	800	38392	589.96	48.92	405	800
	5	25301	578.21	51.23	430	756	37851	614.12	45.48	430	830
	6	23675	587.55	54.27	440	870	39172	628.26	49.14	440	870
	7	22977	600.73	58.07	450	880	40223	642.77	53.26	450	880
	8	21961	613.33	61.41	470	890	40214	656.28	52.31	470	890
Science	4	26142	378.12	50.26	190	600	38378	413.68	50.17	190	600
	8	21896	571.52	51.16	390	770	40217	606.42	46.93	390	770
Social Studies	4	26128	376.20	49.34	200	570	38384	411.24	48.36	200	570
	8	21874	573.11	50.46	420	780	40205	610.93	51.54	420	780
	10	20169	670.78	56.49	490	890	43595	709.02	52.34	490	890

Table 8-34 Scale Score Descriptive Statistics by Disability

Content	Grade	Disabled					Not Disabled				
		N Count	Mean	SD	Min	Max	N Count	Mean	SD	Min	Max
English Language Arts	3	7123	529.42	44.39	370	705	56823	562.84	45.91	330	755
	4	7357	546.17	51.83	340	792	57066	590.30	50.36	340	911
	5	7346	556.56	51.04	350	785	55649	609.41	47.70	350	828
	6	6942	564.57	46.88	360	764	55812	620.81	46.56	360	910
	7	7036	567.21	54.12	370	828	56055	634.28	55.38	370	960
	8	6843	576.83	53.42	380	815	55266	645.23	58.29	380	970
Mathematics	3	7128	522.47	55.90	360	760	56938	559.10	46.05	360	760
	4	7374	533.81	62.82	405	800	57159	579.56	51.55	405	800
	5	7351	558.70	57.28	430	830	55801	605.13	47.54	430	830
	6	6956	558.52	59.89	440	764	55891	619.70	50.17	440	870
	7	7042	573.03	61.04	450	880	56158	634.31	54.64	450	880
	8	6837	582.88	63.41	470	890	55338	648.31	54.69	470	890
Science	4	7377	367.67	52.83	190	600	57143	403.35	51.81	190	600
	8	6821	546.90	54.59	390	770	55292	599.95	47.68	390	770
Social Studies	4	7373	363.54	53.25	200	570	57139	401.37	49.89	200	570
	8	6817	548.68	50.56	420	780	55262	603.64	51.58	420	780
	10	6600	646.26	59.26	490	890	57164	702.77	53.22	490	890

Table 8-35 Scale Score Descriptive Statistics by English Language Proficiency

Content	Grade	Limited English Proficient					Fully English Proficient				
		N Count	Mean	SD	Min	Max	N Count	Mean	SD	Min	Max
English Language Arts	3	5740	534.34	39.46	330	700	58206	561.56	46.90	330	755
	4	4862	548.38	41.64	340	696	59561	588.27	52.09	340	911
	5	3656	558.79	40.68	350	705	59339	605.98	50.30	350	828
	6	2888	563.86	39.22	360	762	59866	617.04	48.97	360	910
	7	2643	569.05	45.55	370	706	60448	629.33	58.37	370	960
	8	2600	579.32	46.42	380	761	59509	640.24	60.93	380	970
Mathematics	3	5869	529.26	45.24	360	760	58197	557.62	48.21	360	760
	4	4981	535.77	53.00	405	800	59552	577.56	53.84	405	800
	5	3790	561.18	48.49	430	692	59362	602.19	50.16	430	830
	6	2972	561.34	49.44	440	732	59875	615.49	53.78	440	870
	7	2753	571.95	53.13	450	742	60447	630.01	57.63	450	880
	8	2696	587.95	57.43	470	766	59479	643.52	58.30	470	890
Science	4	4974	359.15	42.78	190	600	59546	402.62	52.56	190	600
	8	2692	541.29	44.32	390	725	59421	596.51	50.24	390	770
Social Studies	4	4972	358.84	42.81	200	570	59540	400.24	51.11	200	570
	8	2689	545.17	42.30	420	698	59390	599.98	53.54	420	780
	10	2098	634.78	48.08	490	890	61666	699.04	55.61	490	890

Table 8-36 Performance Level Cut Scores for All Contents

Content	3			4			5			6			7			8			10		
	B	P	A	B	P	A	B	P	A	B	P	A	B	P	A	B	P	A	B	P	A
English Language Arts	522	570	624	546	592	650	564	610	670	572	622	671	585	638	697	592	652	708			
Mathematics	517	560	611	536	588	633	574	611	658	582	626	688	606	647	712	620	667	718			
Science				348	399	447										552	600	645			
Social Studies				363	396	436										563	599	640	670	703	741

Note: The abbreviation “B” is for the *Basic* performance level, “P” is for the *Proficient* performance level, and “A” is for the *Advanced* performance level.

Table 8-37 Cut Scores and Associated Impact Data, English Language Arts

Grade	Score Range				Impact Data				
	Below Basic	Basic	Proficient	Advanced	Below Basic	Basic	Proficient	Advanced	Proficient + Advanced
3	330–521	522–569	570–623	624–900	21.45	36.72	33.81	8.02	41.83
4	340–545	546–591	592–649	650–930	21.14	32.14	37.00	9.71	46.72
5	350–563	564–609	610–669	670–940	20.36	33.22	37.88	8.54	46.42
6	360–571	572–621	622–670	671–950	18.23	36.52	33.51	11.75	45.26
7	370–584	585–637	638–696	697–960	22.27	34.10	33.52	10.11	43.63
8	380–591	592–651	652–707	708–970	21.66	37.22	29.19	11.93	41.12

Table 8-38 Cut Scores and Associated Impact Data, Mathematics

Grade	Score Range				Impact Data				
	Below Basic	Basic	Proficient	Advanced	Below Basic	Basic	Proficient	Advanced	Proficient + Advanced
3	360–516	517–559	560–610	611–760	18.90	33.06	37.84	10.20	48.03
4	405–535	536–587	588–632	633–800	19.13	37.37	32.67	10.83	43.50
5	430–573	574–610	611–657	658–830	24.97	30.57	34.58	9.88	44.46
6	440–581	582–625	626–687	688–870	24.70	31.68	37.50	6.11	43.61
7	450–605	606–646	647–711	712–880	30.80	29.92	34.53	4.75	39.29
8	470–619	620–666	667–717	718–890	28.43	36.95	28.33	6.29	34.62

Table 8-39 Cut Scores and Associated Impact Data, Science

Grade	Score Range				Impact Data				
	Below Basic	Basic	Proficient	Advanced	Below Basic	Basic	Proficient	Advanced	Proficient + Advanced
4	190–347	348–398	399–446	447–600	15.29	33.63	34.70	16.37	51.07
8	390–551	552–599	600–644	645–770	17.61	34.74	34.11	13.54	47.65

Table 8-40 Cut Scores and Associated Impact Data, Social Studies

Grade	Score Range				Impact Data				
	Below Basic	Basic	Proficient	Advanced	Below Basic	Basic	Proficient	Advanced	Proficient + Advanced
4	200–362	363–395	396–435	436–570	23.02	24.93	31.84	20.20	52.04
8	420–562	563–598	599–639	640–780	23.47	26.50	31.04	18.98	50.03
10	490–669	670–702	703–740	741–890	27.72	24.12	27.83	20.33	48.17

Table 8-41 Percentage of Students in Each Performance Level by Subgroup, English Language Arts

Grade	Performance Level	Examinees		Gender		Race/Ethnicity						ELP		Disability		SES	
		N	%	Female	Male	White	African-American	Hispanic	Asian	American Indian	Two or More	Fully English Proficient	Limited English Proficient	Disabled	Not Disabled	Economically Disadvantaged	Not Economically Disadvantaged
3	BB	13718	21.45	19.22	23.60	13.68	52.96	32.07	21.36	34.81	22.52	19.95	36.72	45.44	18.44	34.08	12.52
	B	23478	36.72	35.14	38.23	35.90	33.46	42.59	39.22	40.55	36.02	35.87	45.30	35.95	36.81	40.74	33.87
	P	21621	33.81	35.88	31.82	40.27	12.29	22.02	30.64	22.29	33.18	35.51	16.60	16.36	36.00	21.97	42.18
	A	5129	8.02	9.75	6.35	10.15	1.29	3.33	8.77	2.35	8.28	8.68	1.38	2.25	8.74	3.20	11.43
<b>Total</b>		63946	100.00	31377	32569	42335	7151	8557	2565	767	2571	58206	5740	7123	56823	26487	37459
4	BB	13619	21.14	18.32	23.83	12.72	54.90	33.90	21.28	34.84	22.53	19.25	44.26	49.76	17.45	34.34	12.16
	B	20708	32.14	30.87	33.36	31.10	29.38	37.42	35.27	38.07	34.74	31.31	42.37	31.81	32.19	36.64	29.08
	P	23838	37.00	39.01	35.08	43.74	14.15	24.93	33.28	24.22	34.74	38.96	12.98	16.16	39.69	25.42	44.88
	A	6258	9.71	11.80	7.73	12.44	1.56	3.75	10.16	2.86	7.99	10.47	0.39	2.27	10.67	3.60	13.88
<b>Total</b>		64423	100.00	31448	32975	43005	7038	8543	2509	838	2490	59561	4862	7357	57066	26089	38334
5	BB	12828	20.36	17.16	23.41	12.68	52.15	30.87	22.54	33.87	22.07	18.41	52.13	55.28	15.75	33.17	11.82
	B	20924	33.22	30.97	35.35	31.71	32.40	39.17	34.55	41.10	37.48	32.84	39.31	30.67	33.55	38.61	29.62
	P	23863	37.88	40.97	34.94	44.73	14.03	26.74	33.40	23.54	33.08	39.70	8.37	12.67	41.21	25.37	46.23
	A	5380	8.54	10.90	6.30	10.88	1.41	3.22	9.51	1.49	7.37	9.05	0.19	1.37	9.49	2.84	12.34
<b>Total</b>		62995	100.00	30690	32305	42228	6799	8389	2524	803	2252	59339	3656	7346	55649	25211	37784

Table 8-41 Percentage of Students in Each Performance Level by Subgroup, English Language Arts (cont.)

Grade	Performance Level	Examinees		Gender		Race/Ethnicity					ELP		Disability		SES		
		N	%	Female	Male	White	African-American	Hispanic	Asian	American Indian	Two or More	Fully English Proficient	Limited English Proficient	Disabled	Not Disabled	Economically Disadvantaged	Not Economically Disadvantaged
6	BB	11437	18.23	14.27	22.02	11.42	47.71	29.51	16.31	32.66	20.89	16.44	55.30	56.55	13.46	31.32	10.33
	B	22916	36.52	34.56	38.39	34.88	37.60	43.28	36.69	43.25	38.53	36.37	39.54	32.99	36.96	42.04	33.18
	P	21026	33.51	36.06	31.05	39.03	12.80	22.46	33.18	21.19	30.98	34.88	5.02	9.13	36.54	22.30	40.27
	A	7375	11.75	15.10	8.54	14.67	1.89	4.76	13.82	2.90	9.61	12.31	0.14	1.33	13.05	4.35	16.22
<b>Total</b>		62754	100.00	30726	32028	43028	6502	7886	2453	793	2092	59866	2888	6942	55812	23618	39136
7	BB	14048	22.27	17.51	26.79	15.62	53.22	33.89	17.13	38.34	25.29	20.55	61.60	64.10	17.02	36.61	14.09
	B	21517	34.10	33.28	34.89	33.45	31.98	38.85	34.85	35.81	35.31	34.10	34.28	26.62	35.04	37.70	32.06
	P	21149	33.52	36.30	30.88	38.51	12.93	23.08	36.04	23.08	30.48	34.81	4.05	8.44	36.67	21.97	40.10
	A	6377	10.11	12.92	7.44	12.42	1.87	4.19	11.97	2.77	8.92	10.55	0.08	0.84	11.27	3.71	13.75
<b>Total</b>		63091	100.00	30739	32352	43857	6317	7717	2422	793	1985	60448	2643	7036	56055	22904	40187
8	BB	13450	21.66	16.00	27.01	15.27	51.58	32.77	18.40	35.93	24.49	20.07	57.92	61.83	16.68	35.63	14.05
	B	23119	37.22	36.13	38.26	36.36	35.43	42.45	38.08	43.06	38.55	37.18	38.19	30.34	38.08	40.97	35.19
	P	18131	29.19	32.27	26.28	33.75	11.06	19.91	28.00	19.07	26.20	30.30	3.73	6.77	31.97	18.99	34.74
	A	7409	11.93	15.60	8.46	14.62	1.92	4.87	15.52	1.95	10.76	12.44	0.15	1.07	13.27	4.41	16.02
<b>Total</b>		62109	100.00	30195	31914	43481	6186	7583	2332	771	1756	59509	2600	6843	55266	21888	40221



Table 8-42 Percentage of Students in Each Performance Level by Subgroup, Mathematics

Grade	Performance Level	Examinees		Gender		Race/Ethnicity						ELP		Disability		SES	
		N	%	Female	Male	White	African-American	Hispanic	Asian	American Indian	Two or More	Fully English Proficient	Limited English Proficient	Disabled	Not Disabled	Economically Disadvantaged	Not Economically Disadvantaged
3	BB	12110	18.90	19.00	18.81	11.16	48.62	30.52	18.65	32.03	21.04	17.26	35.15	43.31	15.85	30.83	10.45
	B	21183	33.06	34.87	31.32	30.94	35.41	40.09	34.45	38.93	34.85	32.20	41.64	32.58	33.13	38.70	29.07
	P	24240	37.84	37.29	38.36	44.91	14.76	25.84	32.72	26.17	34.50	39.57	20.67	19.99	40.07	26.76	45.69
	A	6533	10.20	8.83	11.51	12.99	1.21	3.55	14.19	2.86	9.61	10.97	2.54	4.12	10.96	3.71	14.80
<b>Total</b>		64066	100.00	31437	32629	42346	7162	8618	2601	768	2571	58197	5869	7128	56938	26574	37492
4	BB	12344	19.13	19.66	18.62	10.57	52.43	32.64	19.26	34.05	21.04	17.19	42.32	46.32	15.62	31.82	10.49
	B	24116	37.37	39.93	34.93	35.65	36.44	44.52	39.07	42.89	41.45	36.69	45.53	35.15	37.66	43.10	33.47
	P	21086	32.67	31.66	33.64	39.95	9.84	19.25	27.04	19.35	28.15	34.50	10.88	15.00	34.96	21.29	40.43
	A	6987	10.83	8.74	12.82	13.83	1.29	3.59	14.62	3.70	9.36	11.63	1.26	3.53	11.77	3.79	15.62
<b>Total</b>		64533	100.00	31505	33028	43021	7042	8599	2544	837	2490	59552	4981	7374	57159	26141	38392
5	BB	15768	24.97	24.11	25.78	15.80	61.30	38.85	22.29	44.14	30.73	23.08	54.62	56.99	20.75	39.92	14.97
	B	19306	30.57	32.05	29.16	29.83	26.94	35.45	33.45	35.66	32.15	30.31	34.72	26.94	31.05	33.93	28.33
	P	21838	34.58	35.14	34.05	41.58	10.83	22.79	31.34	18.08	29.35	36.15	9.92	13.60	37.34	22.95	42.35
	A	6240	9.88	8.69	11.01	12.80	0.94	2.91	12.92	2.12	7.77	10.46	0.74	2.48	10.86	3.20	14.35
<b>Total</b>		63152	100.00	30779	32373	42238	6842	8456	2562	802	2252	59362	3790	7351	55801	25301	37851

Table 8-42 Percentage of Students in Each Performance Level by Subgroup, Mathematics (cont.)

Grade	Performance Level	Examinees		Gender		Race/Ethnicity						ELP		Disability		SES	
		N	%	Female	Male	White	African-American	Hispanic	Asian	American Indian	Two or More	Fully English Proficient	Limited English Proficient	Disabled	Not Disabled	Economically Disadvantaged	Not Economically Disadvantaged
6	BB	15526	24.70	23.50	25.86	15.90	61.54	40.57	21.04	42.77	28.57	22.76	63.83	64.13	19.80	41.38	14.63
	B	19913	31.68	32.83	30.59	31.27	27.47	36.68	31.10	34.84	33.97	31.79	29.61	23.62	32.69	34.44	30.02
	P	23566	37.50	38.14	36.88	45.18	10.41	21.13	37.16	21.76	32.11	39.04	6.46	11.21	40.77	22.61	46.50
	A	3842	6.11	5.53	6.67	7.65	0.58	1.63	10.70	0.63	5.35	6.41	0.10	1.04	6.75	1.58	8.86
<b>Total</b>		62847	100.00	30786	32061	43044	6516	7923	2476	795	2093	59875	2972	6956	55891	23675	39172
7	BB	19464	30.80	30.33	31.25	20.99	71.49	49.74	27.60	56.06	36.94	28.77	75.41	72.74	25.54	49.70	20.00
	B	18907	29.92	31.47	28.44	31.10	20.60	30.88	31.07	25.25	30.15	30.35	20.34	17.86	31.43	30.05	29.84
	P	21824	34.53	33.93	35.11	41.99	7.59	18.17	32.58	17.80	28.99	35.92	4.03	8.78	37.76	19.18	43.30
	A	3005	4.75	4.28	5.21	5.91	0.32	1.21	8.75	0.88	3.93	4.96	0.22	0.62	5.27	1.07	6.86
<b>Total</b>		63200	100.00	30783	32417	43856	6349	7770	2446	792	1987	60447	2753	7042	56158	22977	40223
8	BB	17678	28.43	25.55	31.16	19.48	68.11	45.18	23.69	50.19	34.34	26.61	68.58	71.00	23.17	46.68	18.47
	B	22971	36.95	38.57	35.41	38.47	24.59	38.45	36.99	35.15	36.89	37.37	27.63	22.93	38.68	36.77	37.04
	P	17616	28.33	29.96	26.79	34.30	6.81	14.64	27.77	13.23	23.21	29.45	3.64	5.28	31.18	15.07	35.58
	A	3910	6.29	5.92	6.64	7.74	0.49	1.73	11.55	1.43	5.56	6.57	0.15	0.79	6.97	1.48	8.92
<b>Total</b>		62175	100.00	30219	31956	43477	6181	7629	2355	771	1762	59479	2696	6837	55338	21961	40214

Table 8-43 Percentage of Students in Each Performance Level by Subgroup, Science

Grade	Performance Level	Examinees		Gender		Race/Ethnicity						ELP		Disability		SES	
		N	%	Female	Male	White	African-American	Hispanic	Asian	American Indian	Two or More	Fully English Proficient	Limited English Proficient	Disabled	Not Disabled	Economically Disadvantaged	Not Economically Disadvantaged
4	BB	9868	15.29	14.88	15.69	7.90	45.96	25.50	17.08	24.31	16.35	13.64	35.10	34.87	12.77	25.75	8.17
	B	21700	33.63	34.09	33.20	30.17	38.11	43.45	38.79	45.53	37.62	32.30	49.54	38.88	32.96	41.09	28.55
	P	22388	34.70	35.60	33.84	40.85	13.30	24.43	30.43	23.12	32.46	36.43	13.99	19.66	36.64	25.51	40.96
	A	10564	16.37	15.43	17.27	21.07	2.63	6.61	13.70	7.03	13.57	17.63	1.37	6.60	17.63	7.65	22.31
<b>Total</b>		64520	100.00	31498	33022	43028	7032	8591	2547	839	2483	59546	4974	7377	57143	26142	38378
8	BB	10939	17.61	15.35	19.75	10.47	52.01	29.07	17.55	27.60	20.15	15.93	54.75	52.10	13.36	30.94	10.36
	B	21575	34.74	35.82	33.71	32.70	35.23	44.04	36.55	44.92	36.14	34.53	39.30	33.94	34.83	40.60	31.54
	P	21187	34.11	35.60	32.71	39.81	10.95	22.41	32.72	23.44	31.42	35.40	5.72	10.85	36.98	22.94	40.19
	A	8412	13.54	13.24	13.83	17.03	1.81	4.48	13.17	4.04	12.29	14.15	0.22	3.11	14.83	5.52	17.91
<b>Total</b>		62113	100.00	30192	31921	43469	6139	7627	2353	768	1757	59421	2692	6821	55292	21896	40217

Table 8-44 Percentage of Students in Each Performance Level by Subgroup, Social Studies

Grade	Performance Level	Examinees		Gender		Race/Ethnicity						ELP		Disability		SES	
		N	%	Female	Male	White	African-American	Hispanic	Asian	American Indian	Two or More	Fully English Proficient	Limited English Proficient	Disabled	Not Disabled	Economically Disadvantaged	Not Economically Disadvantaged
4	BB	14852	23.02	22.06	23.94	13.83	58.74	36.56	26.71	40.21	24.84	20.80	49.64	48.98	19.67	37.09	13.44
	B	16085	24.93	25.27	24.62	23.67	23.23	30.19	29.03	30.79	27.25	24.22	33.51	25.29	24.89	28.77	22.32
	P	20541	31.84	32.53	31.18	36.83	14.14	24.37	27.10	19.93	30.27	33.25	14.94	18.13	33.61	24.64	36.74
	A	13034	20.20	20.15	20.26	25.68	3.88	8.88	17.16	9.07	17.64	21.73	1.91	7.60	21.83	9.49	27.50
<b>Total</b>		64512	100.00	31496	33016	43022	7029	8589	2546	838	2488	59540	4972	7373	57139	26128	38384
8	BB	14573	23.47	21.13	25.69	15.97	59.15	35.54	22.25	38.36	27.53	21.70	62.70	62.09	18.71	39.24	14.90
	B	16449	26.50	26.84	26.18	25.18	24.93	33.52	28.70	33.81	27.99	26.33	30.16	23.71	26.84	30.58	24.28
	P	19272	31.04	33.29	28.92	35.28	12.64	23.82	29.04	21.72	28.50	32.15	6.73	10.99	33.52	22.59	35.64
	A	11785	18.98	18.74	19.21	23.57	3.28	7.12	20.00	6.11	15.98	19.82	0.41	3.21	20.93	7.58	25.19
<b>Total</b>		62079	100.00	30179	31900	43457	6129	7611	2355	769	1758	59390	2689	6817	55262	21874	40205
10	BB	17675	27.72	27.08	28.33	20.60	65.54	43.38	26.44	46.83	30.52	26.01	78.03	66.52	23.24	45.76	19.37
	B	15377	24.12	25.97	22.34	23.91	20.15	27.97	26.40	24.96	22.83	24.35	17.21	18.83	24.73	25.80	23.34
	P	17747	27.83	29.24	26.48	31.11	11.16	20.41	26.32	20.87	27.45	28.63	4.34	10.02	29.89	20.07	31.42
	A	12965	20.33	17.71	22.86	24.38	3.15	8.23	20.84	7.33	19.20	21.01	0.43	4.64	22.15	8.36	25.87
<b>Total</b>		63764	100.00	31234	32530	46533	5459	7021	2443	709	1599	61666	2098	6600	57164	20169	43595

Table 8-45a Summary Statistics for Content Standards Raw and SPI Scores, English Language Arts

Grade	N	Content Standard	Standard	No. of Items		Total Score Points	Mean	Mean <i>p</i> -Value	SD	SPI	
				MC	CR					Mean	SD
3	63946	A	Reading - Key Ideas and Details	6	2	10	5.69	0.58	2.44	57.04	21.26
	63946	B	Reading - Craft & Structure	3	2	7	3.70	0.51	1.87	52.88	22.51
	63946	C	Reading - Vocabulary Use	3	0	3	1.83	0.61	0.97	n/a*	n/a*
	63946	D	Writing/Language - Text Types and Purposes	1	4	17	6.09	0.55	2.87	36.22	14.32
	63946	E	Writing/Language - Research	3	1	5	3.06	0.64	1.32	60.61	20.42
	63946	F	Writing/Language - Language Conventions	2	1	4	1.82	0.48	1.09	45.94	16.57
	63946	G	Listening	3	2	7	4.69	0.68	1.79	66.38	20.95
4	64423	A	Reading - Key Ideas and Details	4	3	9	5.46	0.59	2.33	60.42	23.06
	64423	B	Reading - Craft & Structure	4	1	6	2.94	0.53	1.57	49.42	20.12
	64423	C	Reading - Vocabulary Use	4	1	5	2.96	0.59	1.35	59.06	21.88
	64423	D	Writing/Language - Text Types and Purposes	3	2	5	3.18	0.64	1.30	63.20	19.62
	64423	E	Writing/Language - Research	3	2	17	6.55	0.53	3.18	39.01	16.09
	64423	F	Writing/Language - Language Conventions	2	2	6	4.30	0.70	1.27	71.21	15.20
	64423	G	Listening	4	2	8	4.86	0.60	1.97	60.58	19.90

\* SPI scores are not computed for content standards with fewer than four score points.

Table 8-45a Summary Statistics for Content Standards Raw and SPI Scores, English Language Arts (cont.)

Grade	N	Content Standard	Standard	No. of Items		Total Score Points	Mean	Mean <i>p</i> -Value	SD	SPI	
				MC	CR					Mean	SD
5	62995	A	Reading - Key Ideas and Details	2	4	10	4.01	0.42	2.14	40.61	16.96
	62995	B	Reading - Craft & Structure	8	0	8	4.84	0.61	1.89	60.62	19.55
	62995	C	Reading - Vocabulary Use	2	0	2	1.40	0.70	0.67	n/a*	n/a*
	62995	D	Writing/Language - Text Types and Purposes	6	0	6	4.32	0.72	1.42	71.45	18.42
	62995	E	Writing/Language - Research	1	2	15	5.60	0.59	2.26	37.59	12.39
	62995	F	Writing/Language - Language Conventions	5	1	7	4.73	0.67	1.68	67.18	19.15
	62995	G	Listening	4	2	8	4.29	0.55	1.95	53.86	19.58
6	62754	A	Reading - Key Ideas and Details	3	3	8	5.18	0.65	1.89	64.51	20.22
	62754	B	Reading - Craft & Structure	7	2	11	7.02	0.63	2.50	63.64	20.08
	62754	C	Reading - Vocabulary Use	1	0	1	0.40	0.40	0.49	n/a*	n/a*
	62754	D	Writing/Language - Text Types and Purposes	4	0	4	3.12	0.78	0.95	77.06	16.42
	62754	E	Writing/Language - Research	1	3	17	6.20	0.40	2.94	37.02	14.46
	62754	F	Writing/Language - Language Conventions	1	4	7	4.16	0.59	1.60	59.31	17.68
	62754	G	Listening	4	2	8	4.81	0.64	1.76	60.09	16.91

\* SPI scores are not computed for content standards with fewer than four score points.

Table 8-45a Summary Statistics for Content Standards Raw and SPI Scores, English Language Arts (cont.)

Grade	N	Content Standard	Standard	No. of Items		Total Score Points	Mean	Mean <i>p</i> -Value	SD	SPI	
				MC	CR					Mean	SD
7	63091	A	Reading - Key Ideas and Details	8	1	10	6.13	0.62	2.72	61.26	24.50
	63091	B	Reading - Craft & Structure	5	2	8	4.87	0.61	2.02	60.69	21.35
	63091	C	Reading - Vocabulary Use	2	0	2	1.44	0.72	0.70	n/a*	n/a*
	63091	D	Writing/Language - Text Types and Purposes	3	2	6	3.50	0.59	1.67	58.12	20.67
	63091	E	Writing/Language - Research	2	2	16	7.38	0.54	3.14	46.68	17.15
	63091	F	Writing/Language - Language Conventions	2	2	6	3.51	0.60	1.22	58.40	14.56
	63091	G	Listening	2	3	8	5.00	0.62	2.21	62.20	23.74
8	62109	A	Reading - Key Ideas and Details	5	3	10	5.87	0.60	2.37	58.77	20.15
	62109	B	Reading - Craft & Structure	4	2	8	5.12	0.65	1.95	63.63	21.45
	62109	C	Reading - Vocabulary Use	1	1	2	1.19	0.60	0.75	n/a*	n/a*
	62109	D	Writing/Language - Text Types and Purposes	5	1	7	4.40	0.61	1.65	62.79	18.69
	62109	E	Writing/Language - Research	2	3	17	7.17	0.51	3.41	42.76	17.33
	62109	F	Writing/Language - Language Conventions	2	1	4	2.49	0.66	1.09	61.94	20.40
	62109	G	Listening	4	2	8	4.96	0.64	2.16	61.71	21.82

\* SPI scores are not computed for content standards with fewer than four score points.

Table 8-45b Summary Statistics for Domain Raw and SPI Scores, English Language Arts

Grade	N	Domain	No. of Items		Total Score Points	Mean	Mean <i>p</i> -Value	SD	SPI	
			MC	CR					Mean	SD
3	63946	Listening	3	2	7	4.69	0.68	1.79	66.38	20.95
	63946	Reading	12	4	20	11.23	0.56	4.46	56.16	21.28
	63946	Writing	6	6	26	10.97	0.56	4.23	42.37	15.14
4	64423	Listening	4	2	8	4.86	0.60	1.97	60.58	19.90
	64423	Reading	12	5	20	11.36	0.58	4.49	56.78	21.46
	64423	Writing	8	6	28	14.03	0.62	4.77	50.26	15.94
5	62995	Listening	4	2	8	4.29	0.55	1.95	53.86	19.58
	62995	Reading	12	4	20	10.25	0.55	3.94	51.39	18.05
	62995	Writing	12	3	28	14.65	0.67	4.36	52.28	14.70
6	62754	Listening	4	2	8	4.81	0.64	1.76	60.09	16.91
	62754	Reading	11	5	20	12.60	0.62	4.18	62.88	19.86
	62754	Writing	6	7	28	13.48	0.59	4.46	48.31	14.76
7	63091	Listening	2	3	8	5.00	0.62	2.21	62.20	23.74
	63091	Reading	15	3	20	12.44	0.63	4.70	62.12	22.50
	63091	Writing	7	6	28	14.38	0.58	4.89	51.58	16.41
8	62109	Listening	4	2	8	4.96	0.64	2.16	61.71	21.82
	62109	Reading	10	6	20	12.18	0.62	4.34	60.80	20.53
	62109	Writing	9	5	28	14.05	0.58	5.17	50.46	17.41



Table 8-46 Summary Statistics for Content Standards Raw and SPI Scores, Mathematics

Grade	N	Content Standard	Standard	No. of Items		Total Score Points	Mean	Mean <i>p</i> -Value	SD	SPI	
				MC	CR					Mean	SD
3	64066	A	Operations and Algebraic Thinking	8	1	9	5.38	0.60	2.37	59.86	23.46
	64066	B	Number and Operations in Base Ten	6	2	8	5.11	0.64	2.19	63.69	24.36
	64066	C	Number and Operations - Fractions	5	3	8	5.02	0.63	1.96	62.71	20.48
	64066	D	Measurement and Data	7	3	10	4.72	0.47	2.37	47.55	21.33
	64066	E	Geometry	4	3	7	3.85	0.55	1.81	55.18	20.79
4	64533	A	Operations and Algebraic Thinking	9	1	10	5.40	0.54	2.17	54.05	18.12
	64533	B	Number and Operations in Base Ten	4	5	9	4.57	0.51	2.35	50.63	23.52
	64533	C	Number and Operations - Fractions	8	2	10	4.33	0.43	2.96	43.72	27.14
	64533	D	Measurement and Data	8	2	10	4.64	0.47	2.67	46.69	24.12
	64533	E	Geometry	7	0	7	3.82	0.55	1.94	54.57	22.10
5	63152	A	Operations and Algebraic Thinking	5	4	9	4.30	0.48	2.29	47.78	22.56
	63152	B	Number and Operations in Base Ten	6	3	9	5.11	0.57	2.28	56.31	22.44
	63152	C	Number and Operations - Fractions	7	2	9	4.07	0.45	2.34	45.48	22.65
	63152	D	Measurement and Data	7	3	10	4.02	0.40	2.29	40.41	19.40
	63152	E	Geometry	4	5	9	3.13	0.35	2.15	35.13	20.40

Table 8-46 Summary Statistics for Content Standards Raw and SPI Scores, Mathematics (cont.)

Grade	N	Content Standard	Standard	No. of Items		Total Score Points	Mean	Mean <i>p</i> -Value	SD	SPI	
				MC	CR					Mean	SD
6	62847	E	Geometry	4	3	7	2.55	0.37	1.81	37.01	21.21
	62847	F	Ratios and Proportional Relationships	4	3	7	3.67	0.53	1.70	52.34	19.82
	62847	G	The Number System	7	4	11	5.59	0.51	2.86	50.87	23.66
	62847	H	Expressions and Equations	8	3	11	5.13	0.47	2.63	46.72	21.46
	62847	I	Statistics and Probability	9	1	10	4.76	0.48	2.06	47.62	16.52
7	63200	E	Geometry	6	4	10	3.95	0.40	2.15	39.64	17.77
	63200	F	Ratios and Proportional Relationships	6	2	8	4.43	0.56	2.14	54.81	23.75
	63200	G	The Number System	4	3	7	2.75	0.39	1.93	39.51	23.89
	63200	H	Expressions and Equations	7	3	10	3.38	0.34	2.34	34.20	20.61
	63200	I	Statistics and Probability	7	4	11	4.54	0.42	2.62	41.36	21.11
8	62175	E	Geometry	5	5	10	3.67	0.37	2.29	36.82	19.56
	62175	G	The Number System	5	3	8	2.78	0.35	1.91	35.24	18.87
	62175	H	Expressions and Equations	6	4	10	4.01	0.40	2.39	40.29	21.28
	62175	I	Statistics and Probability	7	1	8	4.29	0.54	2.07	53.55	22.09
	62175	J	Functions	6	4	10	4.51	0.45	2.45	44.97	21.97

Table 8-47 Summary Statistics for Content Standards Raw and SPI Scores, Science

Grade	N	Content Standard	Standard	No. of Items		Total Score Points	Mean	Mean <i>p</i> -Value	SD	SPI	
				MC	CR					Mean	SD
4	64520	A/B	Science Connections & Nature of Science	7	0	7	5.12	0.73	1.68	73.07	20.29
	64520	C	Science Inquiry	9	0	9	5.84	0.65	2.16	65.23	21.06
	64520	D	Physical Science	5	0	5	3.33	0.67	1.10	66.82	13.73
	64520	E	Earth and Space Science	4	1	5	2.90	0.58	1.29	58.52	17.97
	64520	F	Life & Environmental Science	6	0	6	4.48	0.75	1.30	74.55	16.82
	64520	G/H	Science Applications & Personal Social Perspectives	8	0	8	6.09	0.76	1.84	76.02	20.18
8	62113	A/B	Science Connections & Nature of Science	7	0	7	5.26	0.75	1.65	75.12	20.16
	62113	C	Science Inquiry	9	0	9	6.63	0.74	2.21	73.69	22.21
	62113	D	Physical Science	5	0	5	3.86	0.77	1.21	76.59	18.80
	62113	E	Earth and Space Science	5	0	5	3.06	0.61	1.24	61.57	16.36
	62113	F	Life & Environmental Science	6	0	6	3.75	0.63	1.51	62.99	19.40
	62113	G/H	Science Applications & Personal Social Perspectives	8	0	8	5.31	0.67	1.62	66.73	16.98

Table 8-48 Summary Statistics for Content Standards Raw and SPI Scores, Social Studies

Grade	N	Content Standard	Standard	No. of Items		Total Score Points	Mean	Mean <i>p</i> -Value	SD	SPI	
				MC	CR					Mean	SD
4	64512	A	Geography	9	0	9	6.65	0.74	1.99	73.74	19.51
	64512	B	History	8	0	8	5.71	0.72	1.87	71.37	20.32
	64512	C	Political Science and Citizenship	7	0	7	4.79	0.69	1.55	68.59	17.57
	64512	D	Economics	6	0	6	3.30	0.55	1.69	55.82	23.00
	64512	E	The Behavioral Sciences	8	0	8	5.74	0.72	1.84	71.80	19.48
8	62079	A	Geography	10	0	10	7.10	0.71	2.33	70.88	21.00
	62079	B	History	12	0	12	7.68	0.64	2.78	64.33	21.11
	62079	C	Political Science and Citizenship	6	0	6	4.05	0.68	1.61	67.53	22.21
	62079	D	Economics	7	0	7	4.78	0.68	1.78	68.32	21.88
	62079	E	The Behavioral Sciences	5	0	5	3.47	0.70	1.37	69.46	21.75
10	63764	A	Geography	10	0	10	6.64	0.67	2.41	66.43	20.93
	63764	B	History	12	0	12	8.47	0.71	2.62	70.39	19.78
	63764	C	Political Science and Citizenship	12	0	12	7.35	0.62	2.82	61.51	21.16
	63764	D	Economics	8	0	8	4.34	0.55	1.99	54.75	20.88
	63764	E	The Behavioral Sciences	8	0	8	4.96	0.63	2.00	62.26	20.97

Table 8-49 SPI Cut Scores, English Language Arts

Content Standard/Domain	Performance Level	Grade 3		Grade 4		Grade 5	
		Score Lower Bound	Score Upper Bound	Score Lower Bound	Score Upper Bound	Score Lower Bound	Score Upper Bound
Reading - Key Ideas and Details	1	0	37	0	38	0	25
	2	38	61	39	65	26	40
	3	62	87	66	88	41	64
	4	88	100	89	100	65	100
Reading - Craft & Structure	1	0	32	0	30	0	41
	2	33	59	31	49	42	64
	3	60	82	50	76	65	85
	4	83	100	77	100	86	100
Reading - Vocabulary Use*	1	*	*	0	39	*	*
	2	*	*	40	62	*	*
	3	*	*	63	87	*	*
	4	*	*	88	100	*	*
Writing/Language - Text Types and Purposes	1	0	24	0	47	0	56
	2	25	38	48	68	57	77
	3	39	53	69	85	78	91
	4	54	100	86	100	92	100
Writing/Language - Research	1	0	44	0	25	0	28
	2	45	67	26	39	29	38
	3	68	84	40	58	39	51
	4	85	100	59	100	52	100
Writing/Language - Language Conventions	1	0	32	0	60	0	51
	2	33	48	61	74	52	72
	3	49	66	75	87	73	88
	4	67	100	88	100	89	100
Listening	1	0	47	0	43	0	36
	2	48	74	44	63	37	54
	3	75	91	64	85	55	81
	4	92	100	86	100	82	100
Reading	1	0	36	0	36	0	34
	2	37	61	37	60	35	53
	3	62	85	61	84	54	75
	4	86	100	85	100	76	100
Writing	1	0	29	0	36	0	40
	2	30	45	37	52	41	54
	3	46	61	53	69	55	69
	4	62	100	70	100	70	100

\* SPI scores are not computed for content standards with fewer than four score points.

Table 8-49 SPI Cut Scores, English Language Arts (cont.)

Content Standard/Domain	Performance Level	Grade 6		Grade 7		Grade 8	
		Score Lower Bound	Score Upper Bound	Score Lower Bound	Score Upper Bound	Score Lower Bound	Score Upper Bound
Reading - Key Ideas and Details	1	0	44	0	38	0	40
	2	45	69	39	69	41	64
	3	70	87	70	90	65	82
	4	88	100	91	100	83	100
Reading - Craft & Structure	1	0	43	0	43	0	45
	2	44	68	44	66	46	72
	3	69	86	67	86	73	86
	4	87	100	87	100	87	100
Reading - Vocabulary Use*	1	*	*	*	*	*	*
	2	*	*	*	*	*	*
	3	*	*	*	*	*	*
	4	*	*	*	*	*	*
Writing/Language - Text Types and Purposes	1	0	64	0	41	0	46
	2	65	82	42	61	47	67
	3	83	91	62	83	68	85
	4	92	100	84	100	86	100
Writing/Language - Research	1	0	24	0	33	0	29
	2	25	37	34	48	30	44
	3	38	51	49	65	45	61
	4	52	100	66	100	62	100
Writing/Language - Language Conventions	1	0	43	0	47	0	45
	2	44	63	48	60	46	68
	3	64	78	61	74	69	83
	4	79	100	75	100	84	100
Listening	1	0	44	0	42	0	42
	2	45	63	43	70	43	69
	3	64	77	71	89	70	86
	4	78	100	90	100	87	100
Reading	1	0	42	0	41	0	42
	2	43	67	42	69	43	68
	3	68	86	70	89	69	84
	4	87	100	90	100	85	100
Writing	1	0	34	0	38	0	35
	2	35	50	39	53	36	53
	3	51	64	54	71	54	70
	4	65	100	72	100	71	100

\* SPI scores are not computed for content standards with fewer than four score points.

Table 8-50 SPI Cut Scores, Mathematics

Content Standard/Domain	Performance Level	Grade 3		Grade 4		Grade 5	
		Score Lower Bound	Score Upper Bound	Score Lower Bound	Score Upper Bound	Score Lower Bound	Score Upper Bound
Operations and Algebraic Thinking	1	0	35	0	36	0	29
	2	36	62	37	56	30	49
	3	63	89	57	76	50	79
	4	90	100	77	100	80	100
Number and Operations in Base Ten	1	0	37	0	27	0	39
	2	38	68	28	54	40	61
	3	69	93	55	80	62	84
	4	94	100	81	100	85	100
Number and Operations - Fractions	1	0	42	0	16	0	25
	2	43	64	17	44	26	45
	3	65	88	45	83	46	78
	4	89	100	84	100	79	100
Measurement and Data	1	0	25	0	21	0	24
	2	26	48	22	49	25	38
	3	49	74	50	79	39	68
	4	75	100	80	100	69	100
Geometry	1	0	34	0	33	0	17
	2	35	56	34	55	18	34
	3	57	81	56	85	35	63
	4	82	100	86	100	64	100

Table 8-50 SPI Cut Scores, Mathematics (cont.)

Content Standard/Domain	Performance Level	Grade 6		Grade 7		Grade 8	
		Score Lower Bound	Score Upper Bound	Score Lower Bound	Score Upper Bound	Score Lower Bound	Score Upper Bound
Geometry	1	0	19	0	27	0	23
	2	20	32	28	40	24	41
	3	33	77	41	74	42	70
	4	78	100	75	100	71	100
Ratios and Proportional Relationships*	1	0	37	0	40		
	2	38	56	41	64		
	3	57	80	65	89		
	4	81	100	90	100		
The Number System	1	0	30	0	21	0	20
	2	31	54	22	44	21	39
	3	55	88	45	82	40	67
	4	89	100	83	100	68	100
Expressions and Equations	1	0	28	0	18	0	23
	2	29	48	19	33	24	46
	3	49	83	34	76	47	77
	4	84	100	77	100	78	100
Statistics and Probability	1	0	35	0	25	0	37
	2	36	48	26	43	38	64
	3	49	74	44	82	65	86
	4	75	100	83	100	87	100
Functions**	1					0	28
	2					29	54
	3					55	79
	4					80	100

\* Content standard in grades 6 and 7 only.

\*\* Content standard in grade 8 only.



Table 8-51 SPI Cut Scores, Science

Content Standard/Domain	Performance Level	Grade 4		Grade 8	
		Score Lower Bound	Score Upper Bound	Score Lower Bound	Score Upper Bound
Science Connections & Nature of Science	1	0	49	0	55
	2	50	77	56	81
	3	78	92	82	94
	4	93	100	95	100
Science Inquiry	1	0	40	0	50
	2	41	67	51	81
	3	68	86	82	95
	4	87	100	96	100
Physical Science	1	0	52	0	60
	2	53	67	61	82
	3	68	78	83	93
	4	79	100	94	100
Earth and Space Science	1	0	38	0	46
	2	39	58	47	63
	3	59	75	64	77
	4	76	100	78	100
Life & Environmental Science	1	0	55	0	42
	2	56	76	43	65
	3	77	90	66	84
	4	91	100	85	100
Science Applications & Social and Personal Perspectives	1	0	52	0	50
	2	53	80	51	69
	3	81	94	70	83
	4	95	100	84	100

Table 8-52 SPI Cut Scores, Social Studies

Content Standard/Domain	Performance Level	Grade 4		Grade 8		Grade 10	
		Score Lower Bound	Score Upper Bound	Score Lower Bound	Score Upper Bound	Score Lower Bound	Score Upper Bound
<b>Geography</b>	1	0	60	0	56	0	52
	2	61	77	57	75	53	69
	3	78	90	76	89	70	86
	4	91	100	90	100	87	100
<b>History</b>	1	0	56	0	47	0	60
	2	57	74	48	66	61	75
	3	75	89	67	84	76	87
	4	90	100	85	100	88	100
<b>Political Science and Citizenship</b>	1	0	56	0	47	0	45
	2	57	70	48	71	46	64
	3	71	82	72	88	65	81
	4	83	100	89	100	82	100
<b>Economics</b>	1	0	34	0	51	0	40
	2	35	52	52	72	41	55
	3	53	77	73	88	56	72
	4	78	100	89	100	73	100
<b>The Behavioral Sciences</b>	1	0	57	0	51	0	47
	2	58	75	52	72	48	64
	3	76	88	73	90	65	81
	4	89	100	91	100	82	100

Table 8-53 Longitudinal Comparison of State-Level Scale Score Means: ELA

Grade	Year	N	Mean	Stand. Dev
3	2016	64107	560.57	47.31
	2017	63946	559.12	46.93
4	2016	62609	582.71	49.41
	2017	64423	585.26	52.44
5	2016	62300	599.62	51.11
	2017	62995	603.24	51.00
6	2016	62728	610.36	52.16
	2017	62754	614.59	49.82
7	2016	62084	623.84	54.85
	2017	63091	626.80	59.14
8	2016	61486	637.23	57.27
	2017	62109	637.69	61.61

Table 8-54 Longitudinal Comparison of State-Level Scale Score Means: Mathematics

Grade	Year	N	Mean	Stand. Dev
3	2016	64194	554.28	46.47
	2017	64066	555.03	48.63
4	2016	62674	573.45	56.15
	2017	64533	574.33	54.92
5	2016	62368	599.57	50.19
	2017	63152	599.73	51.00
6	2016	62772	612.67	53.00
	2017	62847	612.93	54.81
7	2016	62144	627.49	57.40
	2017	63200	627.48	58.65
8	2016	61551	640.79	57.54
	2017	62175	641.11	59.36

Table 8-55 Longitudinal Comparison of State-Level Scale Score Means: Science

Grade	Year	N	Mean	Stand. Dev
4	2016	62636	398.83	51.65
	2017	64520	399.27	53.16
8	2016	61471	597.92	52.54
	2017	62113	594.12	51.25

Table 8-56 Longitudinal Comparison of State-Level Scale Score Means: Social Studies

Grade	Year	N	Mean	Stand. Dev
4	2016	62630	398.02	51.49
	2017	64512	397.05	51.71
8	2016	61496	598.06	51.68
	2017	62079	597.60	54.26
10	2016	63991	698.51	53.74
	2017	63764	696.92	56.56

Table 8-57 Longitudinal Comparison of State-Level Impact Data: ELA

Grade	Year	N	Below Basic	Basic	Proficient	Advanced	Prof. & Adv.
3	2016	64107	21.99	34.88	34.29	8.84	43.13
	2017	63946	21.45	36.72	33.81	8.02	41.83
4	2016	62609	22.81	33.88	34.77	8.54	43.30
	2017	64423	21.14	32.14	37.00	9.71	46.72
5	2016	62300	23.17	34.37	34.55	7.91	42.47
	2017	62995	20.36	33.22	37.88	8.54	46.42
6	2016	62728	21.12	36.30	31.67	10.91	42.58
	2017	62754	18.23	36.52	33.51	11.75	45.26
7	2016	62084	23.11	34.91	34.09	7.89	41.98
	2017	63091	22.27	34.10	33.52	10.11	43.63
8	2016	61486	21.24	37.21	31.26	10.30	41.56
	2017	62109	21.66	37.22	29.19	11.93	41.12

Table 8-58 Longitudinal Comparison of State-Level Impact Data: Mathematics

Grade	Year	N	Below Basic	Basic	Proficient	Advanced	Prof. & Adv.
3	2016	64194	18.59	33.41	38.90	9.10	48.00
	2017	64066	18.90	33.06	37.84	10.20	48.03
4	2016	62674	19.59	36.22	33.33	10.86	44.20
	2017	64533	19.13	37.37	32.67	10.83	43.50
5	2016	62368	25.94	29.98	34.14	9.94	44.08
	2017	63152	24.97	30.57	34.58	9.88	44.46
6	2016	62772	25.51	31.66	36.78	6.05	42.84
	2017	62847	24.70	31.68	37.50	6.11	43.61
7	2016	62144	30.45	30.28	34.81	4.45	39.26
	2017	63200	30.80	29.92	34.53	4.75	39.29
8	2016	61551	28.66	37.48	28.12	5.74	33.86
	2017	62175	28.43	36.95	28.33	6.29	34.62

Table 8-59 Longitudinal Comparison of State-Level Impact Data: Science

Grade	Year	N	Below Basic	Basic	Proficient	Advanced	Prof. & Adv.
4	2016	62636	14.85	33.73	35.70	15.73	51.42
	2017	64520	15.29	33.63	34.70	16.37	51.07
8	2016	61471	16.31	34.07	34.36	15.27	49.63
	2017	62113	17.61	34.74	34.11	13.54	47.65

Table 8-60 Longitudinal Comparison of State-Level Impact Data: Social Studies

Grade	Year	N	Below Basic	Basic	Proficient	Advanced	Prof. & Adv.
4	2016	62630	22.55	24.52	32.26	20.66	52.93
	2017	64512	23.02	24.93	31.84	20.20	52.04
8	2016	61496	22.74	27.47	30.82	18.96	49.78
	2017	62079	23.47	26.50	31.04	18.98	50.03
10	2016	63991	26.32	25.18	28.80	19.70	48.50
	2017	63764	27.72	24.12	27.83	20.33	48.17

## Part 9: Reliability

---

Part 9 of the Technical Report builds upon existing analyses of the summary results by providing additional estimates of the reliability of those results. Reliability can be defined as the consistency of an assessment when the testing procedure is repeated with the same testing target group. A reliable assessment is one that would produce stable scores if the same group of students were to take the same test repeatedly, without any fatigue or memory of the test. As detailed below, the reliability of the Spring 2017 Wisconsin Forward Exam was estimated in four ways:

1. Internal consistency was assessed for all items using Cronbach's alpha (1951).
2. Standard error of measurement (SEM) was calculated for raw score and scale score.
3. Classification consistency and classification accuracy were estimated for the performance level classifications.
4. Inter-rater reliability was estimated for the TDA items.

The present chapter addresses AERA, APA, & NCME (2014) Standards 2.0, 2.3, 2.7, 2.11, 2.13, 2.14, and 2.16, which are cited below.

**Standard 2.0** Appropriate evidence of reliability/precision should be provided for the interpretation for each intended score use. (42)

**Standard 2.3** For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant indices of reliability/precision should be reported. (43)

**Standard 2.7** When subjective judgment enters into test scoring, evidence should be provided on both interrater consistency in scoring and within-examinee consistency over repeated measurements. A clear distinction should be made among reliability data based on (a) independent panels of raters scoring the same performance or products, (b) a single panel scoring successive performances or new products, and (c) independent panels scoring successive performances or new products. (44)

**Standard 2.11** Test publishers should provide estimates of reliability/precision as soon as feasible for each relevant subgroup for which the test is recommended. (45)

**Standard 2.13** The standard error of measurement, both overall and conditional (if reported), should be provided in units of each reported score. (45)

**Standard 2.14** When possible and appropriate, conditional standard errors of measurement should be reported at several score levels unless there is evidence that the standard error is constant across score levels. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score. (46)

**Standard 2.16** When a test or combination of measures is used to make classification decisions, estimates should be provided of the percentage of test takers who would be classified in the same way on two replications of the procedure. (46)

Standard 2.3 advises providing reliability estimates and the SEM for all total scores and subscores reported; Standard 2.13 advises reporting SEM in both raw score and scale score units; and Standard 2.11 advises assessing reliability and SEM for all population subgroups. This chapter of the report presents raw score reliability coefficients and SEMs for the four Wisconsin Forward Exam content areas, for each reported content standard for the total group of examinees, and for the subgroups identified by gender, race/ethnicity, socioeconomic status, disability status, and English language proficiency. The scale score conditional SEMs are provided in Section 6.3.1.

Standard 2.16 advises that when testing measures are used to make categorical decisions, the reliability of those decisions should be estimated. In the present context, Standard 2.16 applies specifically to performance level determinations, such as *Proficient* or *Advanced*. As described below, the Spring 2017 Wisconsin Forward Exam adhered to this standard by applying a detailed analysis of classification consistency and classification accuracy—two related measures used to evaluate the reliability of the performance level classifications used in the test program. This analysis also addresses Standard 2.14 by providing a conditional SEM for the cut scores that separate the performance levels.

Standard 2.7 advises reporting measures of inter-rater consistency in which subjective judgment is involved in scoring. As discussed in Part 5, ELA TDA items were scored by the AI engine with second reads performed by human scorers. As this section will show, a detailed assessment of inter-rater consistency was applied to the Wisconsin Forward Exam. The assessment conducted is termed inter-rater reliability; it measures the reliability of the AI engine versus human scorers in terms of the scores given to TDA items.

Combined, Cronbach's alpha, SEM, classification consistency, classification accuracy, and inter-rater reliability provide several forms of evidence related to the reliability of the Wisconsin Forward Exam. Cronbach's alpha and the SEM operate at the content level: they provide estimates of reliability for student scores in ELA or Mathematics, for example. Classification consistency and classification accuracy operate on the associated performance level classifications. These are of particular interest in the context of the *Elementary and Secondary Education Act* and the associated accountability requirements. Inter-rater reliability probes further, looking at individual items and evaluating the reliability of the AI engine versus human scorers as the scores are assigned to TDA items. In addition, statistics on Cronbach's alpha and the SEM and the procedure for setting the standard performance index (SPI) cut scores at the reported content standard level present reliability and precision evidence in support of the diagnostic use of the Wisconsin Forward Exam subscores. Altogether, the provided evidence in this part of the Technical Report targeted at each intended use of the Wisconsin Forward Exam scores addresses Standard 2.0.

## 9.1 Measures of Internal Consistency and Standard Error of Measurement

Cronbach's alpha is a frequently used measure of internal consistency for tests consisting of MC and CR items. Cronbach's alpha ( $\alpha$ ) is computed as

$$\hat{\alpha} = \frac{k}{k-1} \left( 1 - \frac{\sum \sigma_i^2}{\sigma_x^2} \right),$$

where  $k$  = number of items,  $\sigma_x^2$  = the total score variance, and  $\sigma_i^2$  = the variance of item  $i$  (Crocker & Algina, 1986). SEM is defined as

$$SEM = SD \sqrt{1 - reliability},$$

where  $SD$  represents the standard deviation (SD) of the raw score distribution and *reliability* represents Cronbach's alpha.

Cronbach's alpha and the SEM are shown in Tables 9-1 and 9-2, respectively. These tables include information for all students and for the subgroup categories of gender, race/ethnicity, socioeconomic status, disability status, and English language proficiency.

As indicated in Table 9-1, reliability was highest in Mathematics and Social Studies. Looking at all examinees together in the "Total" column, reliability ranges from 0.87 to 0.89 across grades for ELA, from 0.91 to 0.92 for Mathematics, 0.88 for both grades in Science, and from 0.88 to 0.91 for Social Studies. Ideally, we would like all reliability coefficients to be 0.90 or above. However, for relatively short tests that are designed to measure a fairly broad range of content, this is not always a realistic expectation. If 0.90 is considered a conservative criterion for an acceptable level of reliability, as measured by Cronbach's alpha, then none of the ELA assessments, Science assessments, or Social Studies grade 4 assessments would meet this criterion. The reliability coefficients for these tests are consistent with the small number of items (and score points) and the diversity of the content being assessed. Applying the Spearman-Brown prophecy formula to these results indicates that to achieve the 0.90 reliability threshold, the current ELA assessments for grades 3 through 8 would need to be increased from 53, 56, 56, 56, 56, and 56 points to 69, 62, 75, 76, 64, and 64 score points, respectively. For the current Science assessments in grades 4 and 8, the increase would need to be from 40 points for both grades to 49 and 47 score points, respectively. For the current Social Studies assessment in grade 4, the increase would need to be from 38 to 45 score points.

Table 9-1 shows that many of the subgroup reliability coefficients were similar to, albeit slightly lower than, the total reliability coefficients. Reliability coefficients are particularly sensitive to the score distribution and variance, so this result is consistent with the generally larger SDs (as previously discussed in Part 8 of this report and summarized in Tables 8-19 through 8-26) among many of these subgroups.



The differences in reliability among most subgroups on most tests were generally small. Differences between male and female students were within 0.02 of one another for all grades and content areas.

Most differences among the five racial/ethnic groups were also quite small, within 0.04 of one another for all grades in ELA, Science, and Social Studies. In Mathematics, higher test reliabilities were observed for White or Asian students and the lowest reliability was observed for African-American students.

The differences between disabled and not disabled and economically disadvantaged and not disadvantaged students were within 0.07 of one another for all grades and content areas. The greatest differences were between fully English proficient and limited English proficient students, with consistently lower reliability among limited English proficient students. In fact, the test reliability coefficients for limited English proficiency students were lower than for other subgroups for most grades and content areas. The reliability coefficient is affected, among other factors, by the variability of the students' scores. The higher the variability of scores, the higher the reliability coefficient will tend to be. Based on the evaluation of the distribution of the limited English proficiency student test scores, it was observed that the variance of these scores was often lower than the variance of the scores for other groups. The limited English proficiency student groups appear to be more homogeneous on the ability being measured by the test, leading to lower test reliability for this group of students in different grades and content areas.

Table 9-2 presents the raw score SEM for the total population and for the subgroups described above. These values provide important information for raw score interpretation since we can expect an individual's obtained score to fall within two standard errors of his or her true score approximately 95% of the time. Although there were some observable differences in SEM for the different subgroups, all differences were within one-half of a score point. The SEMs for ELA were slightly larger than those for the other content areas. Because these SEMs are on the raw score scale, this result is consistent with the fact that ELA tests have more raw score points and relatively larger raw score SDs than other content areas. For every grade and content area, the conditional SEM for individual scale scores is provided in the scoring tables previously discussed in Part 6 (Tables 6-9 through 6-25).

Reliability, as measured by Cronbach's alpha, and SEM were also computed for content standards within each content area as well as for each language domain in ELA. The exceptions were content standards with fewer than 4 score points for which the reliability coefficients and SEM were not computed.

Table 9-3 shows these reliability coefficients by content standard/domain. The last column presents the reliability for the total content area (with all content standards or domains) for all examinees. It is clear that the reliability per content standard/domain is lower than the reliability for the total test per content area. The number of items (or score points) has a close relationship with reliability, and a smaller number of items (or score points) is generally associated with lower reliability. The number of score points ranged from 7 to 28 per domain and from 1 to 17 per standard for ELA, from 7 to 11 per standard for Mathematics, from 5 to 9 per standard for Science, and from 5 to 12 per standard for Social Studies. A lower level of

reliability statistics per content standard or domain is therefore expected. The lower level of reliability per standard or domain is one of the reasons why the information based on the content standards or domains should be used for low-stakes purposes only (this issue was previously discussed in the context of SPI).

By content standard/domain, the reliability ranges were as follows (Table 9-3):

- For ELA, reliability indices by content standard or domain ranged from 0.30 (for standard F in grade 3) to 0.81 (for the Reading domain in grades 4 and 7).
- For Mathematics, reliability indices by content standard ranged from 0.59 (for standard F in grade 6) to 0.82 (for standard C in grade 4).
- For Science, reliability indices by content standard ranged from 0.30 (for standard D in grade 4) to 0.74 (for standard C in grade 8).
- For Social Studies, reliability indices by content standard ranged from 0.51 (for standard C in grade 4) to 0.74 (for standard C in grade 10).

The SEM associated with each content standard is presented in Table 9-4 by content area and grade level. Some differences in SEM by content standard can be observed. As indicated by the discussion above, these SEMs were smaller than those for the total test and were generally consistent with the number of items within each content standard.

In summary, the reliability indices, as measured by Cronbach’s alpha at the test level, are in a reasonable range given the number of items in each test. As described above, readers should also note that, because the reliability is influenced by the number of items, lower reliability for the content standards with fewer items is to be expected.

### 9.1.1 Conditional Standard Error of Measurement

In contrast to SEM, the conditional standard error of measurement (CSEM) expresses the degree of measurement error in scale score units and are conditioned on the ability of the student. The CSEMs are defined as the reciprocal of the square root of the test information function and can be estimated across all points of the ability continuum (Hambleton & Swaminathan, 1985):

$$\text{CSEM}(\theta_i) = \frac{1}{\sqrt{I(\theta_i)}},$$

where  $I(\theta_i)$  is the test information function, as a sum of item information function 2, obtained as

$$I(\theta_i) = \sum_j \frac{p'_{ij}(\theta_i)^2}{p_{ij}(\theta_i)q_{ij}(\theta_i)},$$

where  $p'_{ij}(\theta_i)$  is the derivative of  $p_{ij}(\theta_i)$ , and  $q_{ij}(\theta_i) = 1 - p_{ij}(\theta_i)$ .

Note that the CSEMs vary in magnitude across the entire range of student ability estimates (i.e., scale scores) and are smaller in the middle of the score distribution and higher at

the tails. This pattern is seen for all Wisconsin Forward Exam CSEMs and is to be expected when IRT methods are used. In compliance with Standard 2.14, the CSEM of each cut score was presented in the raw score-to-scale score tables (Tables 6-9 to 6-25) for all grades and content areas in Part 6 of this report. In addition, graphical representation of the CSEM with the cut scores is presented in Figures I-1 through I-17 of Appendix I for all grades and content areas. As shown in Appendix I, the estimates of measurement error tend to be higher at the low and high ends of the scale score range. The measurement error increases when there are few observations at a particular ability level. Generally, there are few students with extreme scores, and these score levels cannot be estimated as accurately as levels toward the middle of the ability range. Figures I-1 through I-17 demonstrate that the measurement error is minimized at the cut scores and in the middle of the scale range where most students are located.

## 9.2 Classification Consistency and Accuracy

One of the primary goals of education policy is to improve the performance of all students, with a specific goal of having all students become *Proficient*. Because of this heavy emphasis on moving all students to levels of academic performance at or above each state’s self-defined *Proficient* category, the consistency and accuracy of the classification of students into these performance levels are of particular interest. The following section describes how the consistency and accuracy of these classifications were evaluated and provides evidence that supports the validity of these classifications.

Conceptually, classification consistency is defined as the extent to which two classifications of a single student agree, based either on two independent administrations of the same test or on one administration of two parallel test forms. However, it is difficult to obtain data from repeated administrations of the same form because of the cost, time, and student memory from prior administrations. It is also difficult to construct two psychometrically parallel forms. For these reasons, the common practice is to estimate classification consistency from a single administration.

A contingency table representing the probability of particular classification outcomes under specific scenarios is a convenient way to measure classification consistency. The table below is a contingency table of  $(H + 1) \times (H + 1)$ , where H is the number of cut scores. Three cut scores yield a  $4 \times 4$  contingency table, as can be seen below in Table 9-A.

It is common to report two indices of classification consistency: the classification agreement “P” and the coefficient kappa. Hambleton and Novick (1973) proposed P as a measure of classification consistency, where P is defined as the sum of diagonal values of the contingency table:

$$P = P_{11} + P_{22} + P_{33} + P_{44}.$$

Table 9-A Example Contingency Table with Three Cut Scores

	Level 1	Level 2	Level 3	Level 4	Sum
Level 1	P <sub>11</sub>	P <sub>21</sub>	P <sub>31</sub>	P <sub>41</sub>	P. <sub>1</sub>
Level 2	P <sub>12</sub>	P <sub>22</sub>	P <sub>32</sub>	P <sub>42</sub>	P. <sub>2</sub>
Level 3	P <sub>13</sub>	P <sub>23</sub>	P <sub>33</sub>	P <sub>43</sub>	P. <sub>3</sub>
Level 4	P <sub>14</sub>	P <sub>24</sub>	P <sub>34</sub>	P <sub>44</sub>	P. <sub>4</sub>
Sum	P <sub>1.</sub>	P <sub>2.</sub>	P <sub>3.</sub>	P <sub>4.</sub>	1.0

To reflect statistical chance agreement, Swaminathan, Hambleton, and Algina (1974) suggest using Cohen’s kappa (1960) as

$$\text{kappa} = \frac{P - P_c}{1 - P_c},$$

where  $P_c$  is the chance probability of a consistent classification under two completely random assignments. Probability  $P_c$  is the sum of the probabilities obtained by multiplying the marginal probability of the first administration and the corresponding marginal probability of the second administration as

$$P_c = (P_{1.} \times P_{.1}) + (P_{2.} \times P_{.2}) + (P_{3.} \times P_{.3}) + (P_{4.} \times P_{.4}).$$

Landis and Koch (1977) suggest that values of kappa greater than 0.75 indicate “excellent agreement,” values between 0.40 and 0.74 represent “good agreement” beyond chance, and values below 0.40 denote “poor agreement.”

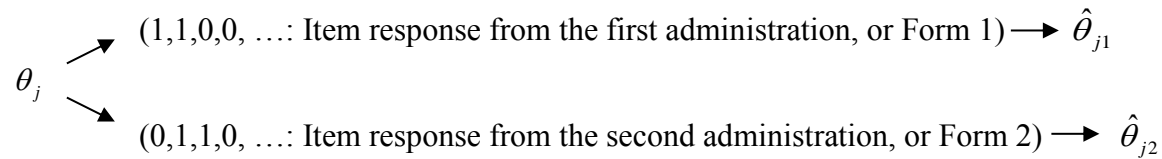
While classification *consistency* refers to the agreement between two observed scores, classification *accuracy* refers to the agreement between the observed score and the true score. Classification accuracy is defined as the extent to which the actual classifications of test takers agree with the classifications that would be made on the basis of their true scores (Livingston & Lewis, 1995). It is common to estimate classification accuracy by assuming the psychometric model to find true scores that correspond to observed scores. For the Wisconsin Forward Exam, the method used to estimate classification accuracy and consistency is the Kolen and Kim method (2004), which is described in the next section of this report (see also Kim, Choi, Um, & Kim, 2006; Kim, Barton, & Kim, 2007).

### 9.2.1 Kolen and Kim’s Method for Pattern Scoring

As stated in Part 6, when item response theory (IRT) is applied to score examinees’ responses, two types of scoring are available: number-correct scoring and item-pattern scoring. The Wisconsin Forward Exam uses item-pattern scoring. Many methods of estimating the consistency and accuracy of classification based on number-correct scoring have been suggested in psychometric literature. However, there have been relatively few studies dealing with item-pattern scoring based on IRT. Kolen and Kim (2004) suggest a simple procedure for pattern scoring (KKM) based on IRT and simulated item responses. The procedure is described below and was implemented with KKCLASS software (Kim, 2005):

Step 1: Obtain item parameters (I) and the ability distribution weight ( $\hat{g}(\theta)$ ) at each quadrature point.

Step 2: Compute two ability estimates at each quadrature point. At a given quadrature point,  $\theta_j$ , generate two sets of item responses using the item parameters from a test form, assuming that the same test form was administered twice to an examinee with the true ability  $\theta_j$ .



If two parallel (or alternative) forms (e.g., Form 1 and Form 2) are available, the two response patterns can be generated based on the item parameters from the two forms.

Step 3: Construct a classification matrix at each quadrature point. Determine the joint event for the cells in Table 9-B by using the two ability estimates obtained from Step 2. Note that this table is constructed for each quadrature point and replication. One, and only one, cell will have a value of one and zeros elsewhere.

Table 9-B Example Classification Table for One Cut Point ( $C_1$ )

	First Administration; or Form 1		
	$\hat{\theta}_{j1} \geq C_1$	$\hat{\theta}_{j1} < C_1$	
$\hat{\theta}_{j2} \geq C_1$			Second Administration; or Form 2
$\hat{\theta}_{j2} < C_1$			

Step 4: Repeat Steps 2 and 3  $R$  times and get average values over  $R$  replications.  $R$  should be a large number (e.g., 500) to obtain stable results.

Step 5: Multiply distribution weight ( $\hat{g}(\theta)$ ) by the average values in Step 4 for each quadrature point and sum across all quadrature points. From this, a final contingency table and classification consistency indices, such as kappa, can be computed.

Because the examinees' abilities are estimated at each quadrature point, this quadrature point can be considered the true score. Therefore, classification accuracy is computed using the examinees' estimated abilities (observed scores) and quadrature points (true scores). Just as 0.90 is generally considered the criterion for acceptable test score reliability, the criterion value of 0.90 is considered to be an acceptably high level of classification accuracy.

In Tables 9-5 through 9-21, there are two tables for each grade and content area. The first table is a contingency table with all three cut scores, which was prepared based on the KKM

procedure. The rows represent the first administration of an assessment, and the columns represent the second administration of the same assessment to the same students. As mentioned above, in the KKM procedure, the score distributions for the first administration and the second administration are estimated using a simulation. So, the value in each cell represents the probability of belonging to a particular pair of performance levels in the first administration and the second administration. For example, when considering the first column of data in the ELA grade 3 table, 0.18 represents the probability of belonging to *Below Basic* in both the first and second administrations. The 0.05 value represents the probability of belonging to *Basic* in the first administration and *Below Basic* in the second administration. The probability of belonging to *Proficient* or *Advanced* in the first administration and *Below Basic* in the second administration is 0.00. “Sum” is obtained simply by adding the four row values or the four column values. This sum is not always identical to the sum of the values shown in the table because the values displayed have been rounded to two decimal places.

The second table shows indices for classification consistency and classification accuracy. Because there are four performance levels for the Wisconsin Forward Exam, there are three cut scores. The values in “All Cuts” were obtained by applying all three cuts together. In Table 9-5 for ELA grade 3, when all three cuts were used for the computation, classification consistency (P) is 0.73, probability of chance is 0.29, kappa ( $k$ ) is 0.61, and classification accuracy is 0.81. The values for “Cut 1” were obtained by applying only the first cut score. There are two levels whenever only one cut is applied (i.e., performance levels above and below the cut). It is clear that the values for P,  $k$ , and classification accuracy with all three cuts are smaller than those for any single cut point. The probability of assigning students to the incorrect performance level will increase with the number of cut scores.

Because the *Proficient* cut score is a criterion for accountability reports, the reliability values for this second cut need to be considered carefully. In Table 9-5, for example, the P for the second cut, which establishes the *Proficient* performance level, was 0.89, kappa was 0.77, and classification accuracy was 0.92. The interpretation of the values illustrated for Table 9-5 is the same for Tables 9-6 through 9-21.

As shown in Tables 9-5 through 9-21, when only the *Proficient* cut score was applied, the classification consistency (P) was greater than or equal to 0.86, and the classification accuracy was greater than or equal to 0.90 for all tests. The kappa value was greater than or equal to 0.72 for all tests. According to Landis and Koch’s criteria for  $k$  (presented previously in this report in the discussion of classification consistency), all tests showed good or excellent agreement based on the cut for the *Proficient* performance level.

In addition, the indices for classification consistency and classification accuracy were computed for the subgroups of students. These data are presented in Appendix J. As seen in Tables J-1 through J-17, when the *Proficient* cut is considered, classification consistency and accuracy coefficients, and the kappa values were good or very good for all subgroups, grades, and content areas. Specifically, the classification consistency was greater than or equal to 0.86 and the classification accuracy was greater than or equal to 0.90 for all ELA subgroups across all grades. The classification consistency was greater than or equal to 0.88 and the classification accuracy was greater than or equal to 0.92 for all Mathematics subgroups across all grades. For

Science, the classification consistency and accuracy was greater than or equal to 0.86 for all subgroups across both grades. For Social Studies, the classification consistency was greater than or equal to 0.87 and the classification accuracy was greater than or equal to 0.90 for all subgroups across all grades. The kappa values were greater than or equal to 0.54 for all subgroups in ELA, greater than or equal to 0.64 for all subgroups in Mathematics, greater than or equal to 0.57 in Science, and greater than or equal to 0.60 for all subgroups in Social Studies. The lowest kappa values were observed for the limited English proficiency subgroups in each content area. This is consistent with the trend of the test reliability coefficients, which were found to be lower for the limited English proficiency students compared to other subgroups.

### 9.3 Inter-Rater Reliability for TDA Items

The reliability of scoring of TDA items was measured in two ways: (1) tabulations of exact and adjacent agreement of two scorers and (2) reliability coefficients. Reliability for TDA items was examined by calculating indices of inter-rater agreement, which is the degree of reliability with which the AI engine and a human scorer assign scores to a given student response. Two indices for inter-rater reliability, intraclass correlation and weighted kappa, are presented here.

**Notation:** To assess reliability, it is necessary to replicate the scoring process for a subset of papers. This is usually done with “blind double-reads.” Suppose that we have  $N$  responses, each of which is scored twice. We denote the two scores of response  $n$  by  $X_{n1}$  and  $X_{n2}$ , where  $n = 1, 2, \dots, N$ . The resulting data may be presented in two ways: enumeration by response and cross-tabulation.

**Data Structure 1: Enumeration by Response.** Each row represents a single student response:

Response #	Score 1	Score 2	Mean Score
1	$X_{11}$	$X_{12}$	$\bar{X}_{1.}$
2	$X_{21}$	$X_{22}$	$\bar{X}_{2.}$
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮
$N$	$X_{N1}$	$X_{N2}$	$\bar{X}_{N.}$
Column Mean	$\bar{X}_{.1}$	$\bar{X}_{.2}$	$\bar{X}$

where

$$\bar{X}_{1.} = (X_{11} + X_{12}) / 2$$

is the mean score for Response 1 (similarly for responses 2, 3, ... $N$ ),

$$\bar{X}_{.1} = \frac{1}{N} \sum_{n=1}^N X_{n1} = (X_{11} + X_{21} + \dots + X_{N1}) / N$$

is the mean of Score 1 over all responses (similarly for Score 2), and

$$\bar{X}_{..} = \frac{1}{N} \sum_{n=1}^N 1(X_{n1} + X_{n2})/2$$

is the overall mean score across both scores of all responses.

**Data Structure 2: Cross-Tabulation of Score 1 and Score 2.** As an alternative, we may create a square table of counts for each Score 1 by Score 2 (i.e.,  $X_{n1} \times X_{n2}$ ) combination:

		Score 2				Row Total
		0	1	...	$m$	
Score 1	0	$n_{00}$	$n_{01}$	...	$n_{0m}$	$n_{0+}$
	1	$n_{10}$	$n_{11}$	...	$n_{1m}$	$n_{1+}$
	.	.	.	...	.	.
	.	.	.	...	.	.
	$m$	$n_{m0}$	$n_{m1}$	...	$n_{mm}$	$n_{m+}$
Column Total		$n_{+0}$	$n_{+1}$	...	$n_{+m}$	$n_{++}$

where  $m$  is the maximum score (for a rubric including zero) obtainable for the item;  $n_{ij}$  is the number of responses for which Score 1 =  $i$  and Score 2 =  $j$ ;  $n_{i+}$  is the number of responses for which Score 1 =  $i$ ; and  $n_{+j}$  is the number of responses for which Score 2 =  $j$ .

Formulas for the two reliability coefficients of interest are then given:

1. Intraclass Correlation,  $\rho_{IC}$ , describes the percentage of overall score variance accounted for by the variance of mean response scores:

$$\rho_{IC} = \frac{Var_n(\bar{X}_n)}{Var_n(X_{n1}, X_{n2})} = \frac{\frac{1}{N-1} \sum_{n=1}^N (\bar{X}_n - \bar{X}_{..})^2}{\frac{1}{2(N-1)} \sum_{n=1}^N [(X_{n1} - \bar{X}_{..})^2 + (X_{n2} - \bar{X}_{..})^2]}$$

If agreement is perfect,  $\rho_{IC} = 1$ . The following is always true:  $0 \leq \rho_{IC} \leq 1$ .

2. Weighted Kappa,  $k$ , is used in many contexts as a measure of association in square contingency tables:

$$k = \frac{\sum_{i=0}^m \sum_{j=0}^m w_{ij} \frac{n_{ij}}{n_{++}} - \sum_{i=0}^m \sum_{j=0}^m w_{ij} \frac{n_{i+} n_{+j}}{n_{++}^2}}{1 - \sum_{i=0}^m \sum_{j=0}^m w_{ij} \frac{n_{i+} n_{+j}}{n_{++}^2}}, \text{ where } w_{ij} = 1 - \frac{(i-j)^2}{M^2}.$$



If agreement is perfect,  $k = 1$ . If agreement is what would be expected by chance,  $k = 0$ . The following is always true:  $0 \leq k \leq 1$ .

Ordinal rating scales (e.g., 0, 1, 2) used in scoring TDA items contain a certain level of chance agreement that is expected. Although the intraclass correlation is reported in this report, it does not take into account the possibility of chance agreement between the two raters. Cohen's kappa ( $k$ ) does take this into consideration. In general,  $k$  will have values equal to or smaller than the intraclass correlation. If agreement is perfect, the value of  $k$  is 1.0. If agreement is at chance levels, the value of  $k$  is 0. As noted in Section 9.2, Landis and Koch (1977) suggest that values of  $k$  greater than 0.75 indicate "excellent agreement," values between 0.40 and 0.74 represent "good agreement" beyond chance, and values below 0.40 denote "poor agreement." Specific criteria for intraclass correlation or weighted  $k$  are not established.

Table 9-22 presents the rater agreement statistics for TDA items. The evidence supporting inter-rater reliability is presented in terms of the percentage of agreement between raters (the AI engine and a human rater), two indices of inter-rater reliability, and the distributions of scores across score levels. In the table, "Exact" agreement is defined as scores that are exactly the same. "Adjacent" agreement is defined as scores differing by 1 point. "Discrepant" cases are those cases where the scores of the two raters differed by more than one raw score point. For example, as shown in Table 9-22, for a grade 3 TDA item, the exact agreement, adjacent agreement, and discrepant agreement rates are 85.68%, 13.82%, and 0.50%, respectively. "Mean" reflects the item mean score from the second reads (by human scorers). "Number of Second Reads" is the number of student responses selected for the purpose of the second read and computing inter-rater reliability. The "Score Frequency" columns represent the scoring outcomes for the student responses based on the raw scores given by the second (human) scorers. The column for "Codes" reflects the number of students who received the condition codes B, C, N, R, or T (described in detail in Part 5, Table 5-2 of this report).

Overall, the rater agreement was very high. Exact scores ranged from 77.57% in grade 7 to 85.68% in grade 3. Adjacent scores ranged from 13.82% in grade 3 to 22.05% in grade 7. Non-discrepant scores (exact plus adjacent agreement) were over 99% in each grade. The intraclass correlation coefficients ranged from 0.79 in grade 3 to 0.89 in grade 8. The weighted kappa ranged from 0.59 in grade 3 to 0.78 in grade 8.

## 9.4 Summary

Overall, the analyses discussed in this section of the report indicated acceptable levels of reliability for the Wisconsin Forward Exam. The internal consistency reliability estimates, as measured by Cronbach's alpha coefficient, were reasonable given the number of items in each test. The analyses of classification consistency and accuracy indicated acceptable levels of consistency and accuracy of student proficiency level classifications, and the SEM around the *Proficient* cut score was low in every grade and content area. The levels of rater agreement were high, and the discrepancy rates were low, with acceptably high values for the weighted kappa and intraclass correlations. The results of the inter-rater reliability analyses indicated a high degree of reliability for scores on the ELA TDA items in the Wisconsin Forward Exam.

Table 9-1 Reliability for Total Group and Subgroups Using Cronbach's Alpha

Content	Grade	Total	Gender		Race/Ethnicity						ELP		Disability		SES	
			Female	Male	White	African-American	Hispanic	Asian	American Indian	Two or More	Fully English Proficient	Limited English Proficient	Disabled	Not Disabled	Economically Disadvantaged	Not Economically Disadvantaged
English Language Arts	3	0.87	0.87	0.87	0.86	0.83	0.85	0.87	0.85	0.88	0.87	0.81	0.86	0.87	0.85	0.86
	4	0.89	0.89	0.89	0.87	0.87	0.87	0.89	0.86	0.89	0.89	0.81	0.88	0.88	0.88	0.88
	5	0.87	0.87	0.87	0.85	0.84	0.84	0.87	0.84	0.86	0.87	0.75	0.84	0.86	0.85	0.86
	6	0.87	0.87	0.87	0.85	0.84	0.85	0.87	0.83	0.87	0.86	0.75	0.84	0.85	0.86	0.85
	7	0.89	0.88	0.89	0.88	0.86	0.87	0.88	0.87	0.88	0.88	0.75	0.84	0.88	0.87	0.88
	8	0.89	0.88	0.89	0.88	0.86	0.87	0.89	0.86	0.89	0.89	0.76	0.84	0.88	0.87	0.88
Mathematics	3	0.91	0.91	0.92	0.90	0.87	0.89	0.92	0.89	0.91	0.91	0.88	0.91	0.91	0.89	0.90
	4	0.92	0.92	0.93	0.91	0.86	0.89	0.94	0.90	0.92	0.92	0.84	0.90	0.92	0.90	0.92
	5	0.91	0.91	0.92	0.91	0.84	0.88	0.92	0.86	0.91	0.91	0.80	0.88	0.91	0.88	0.91
	6	0.91	0.90	0.91	0.90	0.83	0.87	0.92	0.86	0.90	0.91	0.75	0.85	0.90	0.87	0.90
	7	0.91	0.91	0.92	0.91	0.80	0.87	0.92	0.87	0.91	0.91	0.71	0.83	0.91	0.87	0.91
	8	0.91	0.90	0.91	0.90	0.83	0.87	0.92	0.87	0.90	0.91	0.76	0.83	0.90	0.87	0.91
Science	4	0.88	0.88	0.89	0.85	0.85	0.86	0.88	0.85	0.87	0.88	0.81	0.87	0.87	0.87	0.86
	8	0.88	0.87	0.89	0.86	0.86	0.86	0.88	0.86	0.88	0.88	0.80	0.87	0.87	0.88	0.86
Social Studies	4	0.88	0.88	0.89	0.86	0.86	0.86	0.88	0.86	0.88	0.88	0.82	0.88	0.88	0.87	0.87
	8	0.91	0.90	0.91	0.89	0.87	0.88	0.91	0.88	0.90	0.90	0.81	0.87	0.90	0.89	0.89
	10	0.91	0.90	0.92	0.90	0.88	0.89	0.91	0.89	0.91	0.91	0.79	0.88	0.91	0.90	0.91

Table 9-2 Standard Error of Measurement for Total Group and Subgroups

Content	Grade	Total	Gender		Race/Ethnicity					ELP		Disability		SES		
			Female	Male	White	African-American	Hispanic	Asian	American Indian	Two or More	Fully English Proficient	Limited English Proficient	Disabled	Not Disabled	Economically Disadvantaged	Not Economically Disadvantaged
English Language Arts	3	3.30	3.29	3.33	3.24	3.30	3.39	3.37	3.31	3.32	3.28	3.49	3.41	3.28	3.36	3.24
	4	3.06	3.08	3.06	3.03	3.06	3.14	3.15	3.20	3.09	3.05	3.31	3.14	3.07	3.13	3.03
	5	3.69	3.73	3.68	3.75	3.51	3.71	3.70	3.68	3.78	3.71	3.78	3.55	3.76	3.67	3.75
	6	3.72	3.82	3.66	3.81	3.55	3.72	3.72	3.75	3.74	3.76	3.66	3.49	3.82	3.66	3.82
	7	4.01	4.14	3.91	4.08	3.78	4.02	4.17	3.97	4.07	4.04	4.05	3.65	4.14	3.94	4.12
	8	3.34	3.39	3.35	3.30	3.38	3.45	3.38	3.52	3.35	3.34	3.67	3.47	3.38	3.44	3.31
Mathematics	3	2.76	2.85	2.69	2.74	2.92	2.87	2.67	2.85	2.79	2.76	2.82	2.76	2.76	2.86	2.71
	4	2.58	2.68	2.50	2.49	3.17	2.86	2.41	2.77	2.62	2.56	3.02	2.84	2.55	2.86	2.46
	5	3.00	3.15	2.86	2.96	3.53	3.30	2.84	3.40	3.04	2.99	3.33	3.08	2.98	3.23	2.93
	6	3.12	3.24	3.00	3.08	3.71	3.48	2.96	3.47	3.19	3.11	3.68	3.31	3.11	3.41	3.05
	7	3.55	3.73	3.38	3.54	4.48	4.07	3.34	3.97	3.65	3.55	4.36	3.76	3.55	3.98	3.49
	8	3.02	3.10	2.94	2.96	3.75	3.44	2.76	3.41	3.10	3.00	3.72	3.54	2.99	3.43	2.91
Science	4	3.21	3.28	3.14	3.25	3.28	3.32	3.34	3.34	3.31	3.21	3.33	3.18	3.21	3.23	3.22
	8	3.40	3.53	3.27	3.53	3.39	3.50	3.51	3.48	3.43	3.42	3.41	3.15	3.48	3.37	3.49
Social Studies	4	3.15	3.21	3.09	3.19	3.12	3.23	3.24	3.30	3.19	3.15	3.22	3.11	3.17	3.18	3.16
	8	3.06	3.16	2.97	3.07	3.21	3.24	3.08	3.21	3.13	3.07	3.27	3.12	3.10	3.16	3.08
	10	3.00	3.03	2.97	2.92	3.28	3.18	2.99	3.21	3.05	2.99	3.31	3.24	2.97	3.18	2.91

Table 9-3 Cronbach's Alpha Reliability Coefficients for Content Standard and Domain

English Language Arts

Grade	Alpha per Content Standard and Domain									
	A	B	C	D	E	F	G/Listening	Reading	Writing	Total
3	0.64	0.55	*	0.50	0.50	0.30	0.53	0.79	0.69	0.87
4	0.67	0.49	0.57	0.49	0.58	0.43	0.55	0.81	0.75	0.89
5	0.54	0.58	*	0.49	0.46	0.55	0.54	0.74	0.75	0.87
6	0.60	0.67	*	0.36	0.47	0.49	0.51	0.79	0.69	0.87
7	0.72	0.56	*	0.42	0.47	0.39	0.64	0.81	0.69	0.89
8	0.61	0.64	*	0.55	0.53	0.48	0.57	0.79	0.76	0.89

\* Results are not reported for the content standards with fewer than four score points.

Mathematics

Grade	Alpha per Content Standard										
	A	B	C	D	E	F	G	H	I	J	Total
3	0.73	0.73	0.67	0.69	0.61						0.91
4	0.61	0.73	0.82	0.74	0.67						0.92
5	0.72	0.71	0.70	0.67	0.71						0.91
6					0.64	0.59	0.77	0.72	0.60		0.91
7					0.61	0.69	0.69	0.69	0.73		0.91
8					0.67		0.62	0.70	0.68	0.71	0.91

Science

Grade	Alpha per Content Standard						Total
	A/B	C	D	E	F	G/H	
4	0.63	0.67	0.30	0.41	0.50	0.67	0.88
8	0.63	0.74	0.52	0.33	0.50	0.55	0.88

Social Studies

Grade	Alpha per Content Standard					
	A	B	C	D	E	Total
4	0.67	0.64	0.51	0.59	0.64	0.88
8	0.72	0.72	0.60	0.65	0.56	0.91
10	0.69	0.71	0.74	0.61	0.63	0.91

Table 9-4 Standard Error of Measurement per Content Standard and Domain

English Language Arts

Grade	SEM per Content Standard and Domain									
	A	B	C	D	E	F	G/Listening	Reading	Writing	Total
3	1.45	1.25	*	2.02	0.93	0.91	1.22	2.07	2.34	3.30
4	1.35	1.12	0.88	0.93	2.06	0.96	1.31	1.97	2.40	3.06
5	1.45	1.23	*	1.01	1.65	1.13	1.32	1.99	2.18	3.69
6	1.19	1.45	*	0.76	2.15	1.15	1.23	1.93	2.48	3.72
7	1.44	1.34	*	1.27	2.28	0.95	1.32	2.05	2.71	4.01
8	1.47	1.16	*	1.11	2.35	0.79	1.41	1.98	2.54	3.34

\* Results are not reported for the content standards with fewer than four score points.

Mathematics

Grade	SEM per Content Standard										
	A	B	C	D	E	F	G	H	I	J	Total
3	1.24	1.14	1.13	1.32	1.13						2.76
4	1.36	1.22	1.25	1.35	1.11						2.58
5	1.22	1.23	1.28	1.31	1.16						3.00
6					1.09	1.09	1.36	1.39	1.30		3.12
7					1.35	1.20	1.07	1.30	1.37		3.55
8					1.32		1.17	1.30	1.17	1.33	3.02

Science

Grade	SEM per Content Standard						
	A/B	C	D	E	F	G/H	Total
4	1.03	1.24	0.92	0.99	0.92	1.06	3.21
8	1.01	1.13	0.84	1.01	1.07	1.09	3.40

Social Studies

Grade	SEM per Content Standard					
	A	B	C	D	E	Total
4	1.15	1.12	1.08	1.08	1.11	3.15
8	1.23	1.46	1.02	1.05	0.91	3.06
10	1.35	1.42	1.45	1.24	1.22	3.00

Table 9-5 Classification Consistency and Classification Accuracy for English Language Arts Grade 3

Contingency Table with All Cut Scores

Performance Level	Below Basic	Basic	Proficient	Advanced	Sum
<b>Below Basic</b>	0.18	0.04	0.00	0.00	0.22
<b>Basic</b>	0.05	0.25	0.05	0.00	0.36
<b>Proficient</b>	0.00	0.06	0.24	0.03	0.33
<b>Advanced</b>	0.00	0.00	0.03	0.06	0.09
<b>Sum</b>	0.23	0.36	0.33	0.09	

Indexes for Classification Consistency and Classification Accuracy

Indexes	Cut 1	Cut 2	Cut 3	All Cuts
<b>Classification Consistency (P)</b>	0.90	0.89	0.93	0.73
<b>Probability of Chance</b>	0.65	0.51	0.83	0.29
<b>Kappa (k)</b>	0.72	0.77	0.60	0.61
<b>Classification Accuracy</b>	0.93	0.92	0.95	0.81

Table 9-6 Classification Consistency and Classification Accuracy for English Language Arts Grade 4

Contingency Table with All Cut Scores

Performance Level	Below Basic	Basic	Proficient	Advanced	Sum
<b>Below Basic</b>	0.18	0.04	0.00	0.00	0.22
<b>Basic</b>	0.04	0.21	0.06	0.00	0.31
<b>Proficient</b>	0.00	0.06	0.27	0.03	0.36
<b>Advanced</b>	0.00	0.00	0.03	0.08	0.11
<b>Sum</b>	0.22	0.31	0.36	0.11	

Indexes for Classification Consistency and Classification Accuracy

Indexes	Cut 1	Cut 2	Cut 3	All Cuts
<b>Classification Consistency (P)</b>	0.92	0.89	0.93	0.73
<b>Probability of Chance</b>	0.66	0.50	0.80	0.29
<b>Kappa (k)</b>	0.75	0.77	0.66	0.63
<b>Classification Accuracy</b>	0.94	0.92	0.95	0.81

Table 9-7 Classification Consistency and Classification Accuracy for English Language Arts Grade 5

Contingency Table with All Cut Scores

Performance Level	Below Basic	Basic	Proficient	Advanced	Sum
<b>Below Basic</b>	0.16	0.05	0.00	0.00	0.21
<b>Basic</b>	0.05	0.22	0.06	0.00	0.32
<b>Proficient</b>	0.00	0.06	0.27	0.03	0.37
<b>Advanced</b>	0.00	0.00	0.03	0.07	0.10
<b>Sum</b>	0.21	0.33	0.36	0.10	

Indexes for Classification Consistency and Classification Accuracy

Indexes	Cut 1	Cut 2	Cut 3	All Cuts
<b>Classification Consistency (P)</b>	0.91	0.88	0.94	0.72
<b>Probability of Chance</b>	0.67	0.50	0.82	0.29
<b>Kappa (k)</b>	0.72	0.76	0.64	0.61
<b>Classification Accuracy</b>	0.93	0.91	0.96	0.80

Table 9-8 Classification Consistency and Classification Accuracy for English Language Arts Grade 6

Contingency Table with All Cut Scores

Performance Level	Below Basic	Basic	Proficient	Advanced	Sum
<b>Below Basic</b>	0.15	0.04	0.00	0.00	0.19
<b>Basic</b>	0.04	0.25	0.06	0.00	0.35
<b>Proficient</b>	0.00	0.07	0.22	0.04	0.33
<b>Advanced</b>	0.00	0.00	0.04	0.09	0.13
<b>Sum</b>	0.19	0.36	0.32	0.13	

Indexes for Classification Consistency and Classification Accuracy

Indexes	Cut 1	Cut 2	Cut 3	All Cuts
<b>Classification Consistency (P)</b>	0.92	0.87	0.92	0.71
<b>Probability of Chance</b>	0.70	0.50	0.77	0.28
<b>Kappa (k)</b>	0.73	0.74	0.64	0.59
<b>Classification Accuracy</b>	0.94	0.91	0.94	0.78

Table 9-9 Classification Consistency and Classification Accuracy for English Language Arts Grade 7

Contingency Table with All Cut Scores

Performance Level	Below Basic	Basic	Proficient	Advanced	Sum
Below Basic	0.18	0.04	0.00	0.00	0.23
Basic	0.04	0.23	0.06	0.00	0.33
Proficient	0.00	0.06	0.23	0.04	0.33
Advanced	0.00	0.00	0.03	0.08	0.11
Sum	0.23	0.33	0.32	0.12	

Indexes for Classification Consistency and Classification Accuracy

Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Classification Consistency (P)	0.91	0.89	0.93	0.73
Probability of Chance	0.65	0.51	0.79	0.28
Kappa (k)	0.75	0.77	0.65	0.62
Classification Accuracy	0.93	0.92	0.95	0.81

Table 9-10 Classification Consistency and Classification Accuracy for English Language Arts Grade 8

Contingency Table with All Cut Scores

Performance Level	Below Basic	Basic	Proficient	Advanced	Sum
Below Basic	0.18	0.04	0.00	0.00	0.22
Basic	0.04	0.26	0.06	0.00	0.36
Proficient	0.00	0.06	0.18	0.04	0.28
Advanced	0.00	0.00	0.04	0.10	0.14
Sum	0.22	0.36	0.28	0.14	

Indexes for Classification Consistency and Classification Accuracy

Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Classification Consistency (P)	0.91	0.88	0.92	0.72
Probability of Chance	0.65	0.51	0.77	0.28
Kappa (k)	0.75	0.76	0.66	0.61
Classification Accuracy	0.94	0.92	0.94	0.80



Table 9-11 Classification Consistency and Classification Accuracy for Mathematics Grade 3

Contingency Table with All Cut Scores

Performance Level	Below Basic	Basic	Proficient	Advanced	Sum
Below Basic	0.04	0.01	0.00	0.00	0.05
Basic	0.01	0.08	0.03	0.00	0.11
Proficient	0.00	0.02	0.24	0.06	0.33
Advanced	0.00	0.00	0.06	0.46	0.52
Sum	0.04	0.11	0.33	0.52	

Indexes for Classification Consistency and Classification Accuracy

Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Classification Consistency (P)	0.98	0.95	0.87	0.80
Probability of Chance	0.91	0.74	0.50	0.39
Kappa (k)	0.77	0.81	0.75	0.68
Classification Accuracy	0.99	0.97	0.92	0.87

Table 9-12 Classification Consistency and Classification Accuracy for Mathematics Grade 4

Contingency Table with All Cut Scores

Performance Level	Below Basic	Basic	Proficient	Advanced	Sum
Below Basic	0.08	0.02	0.00	0.00	0.10
Basic	0.02	0.14	0.03	0.00	0.20
Proficient	0.00	0.04	0.26	0.06	0.35
Advanced	0.00	0.00	0.04	0.31	0.35
Sum	0.10	0.21	0.33	0.37	

Indexes for Classification Consistency and Classification Accuracy

Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Classification Consistency (P)	0.95	0.93	0.91	0.79
Probability of Chance	0.82	0.58	0.54	0.30
Kappa (k)	0.74	0.83	0.79	0.70
Classification Accuracy	0.97	0.95	0.94	0.86

Table 9-13 Classification Consistency and Classification Accuracy for Mathematics Grade 5

Contingency Table with All Cut Scores

Performance Level	Below Basic	Basic	Proficient	Advanced	Sum
Below Basic	0.19	0.04	0.00	0.00	0.23
Basic	0.04	0.16	0.04	0.00	0.24
Proficient	0.00	0.04	0.27	0.03	0.34
Advanced	0.00	0.00	0.03	0.15	0.18
Sum	0.23	0.24	0.34	0.18	

Indexes for Classification Consistency and Classification Accuracy

Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Classification Consistency (P)	0.92	0.91	0.94	0.77
Probability of Chance	0.64	0.50	0.70	0.26
Kappa (k)	0.77	0.83	0.79	0.68
Classification Accuracy	0.94	0.94	0.96	0.83

Table 9-14 Classification Consistency and Classification Accuracy for Mathematics Grade 6

Contingency Table with All Cut Scores

Performance Level	Below Basic	Basic	Proficient	Advanced	Sum
Below Basic	0.22	0.05	0.00	0.00	0.27
Basic	0.05	0.21	0.05	0.00	0.31
Proficient	0.00	0.06	0.29	0.01	0.36
Advanced	0.00	0.00	0.02	0.04	0.06
Sum	0.27	0.32	0.35	0.06	

Indexes for Classification Consistency and Classification Accuracy

Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Classification Consistency (P)	0.90	0.89	0.97	0.76
Probability of Chance	0.61	0.52	0.89	0.30
Kappa (k)	0.75	0.78	0.73	0.66
Classification Accuracy	0.92	0.92	0.98	0.82

Table 9-15 Classification Consistency and Classification Accuracy for Mathematics Grade 7

Contingency Table with All Cut Scores

Performance Level	Below Basic	Basic	Proficient	Advanced	Sum
Below Basic	0.39	0.06	0.00	0.00	0.45
Basic	0.07	0.20	0.04	0.00	0.31
Proficient	0.00	0.04	0.18	0.01	0.23
Advanced	0.00	0.00	0.01	0.01	0.02
Sum	0.45	0.30	0.23	0.02	

Indexes for Classification Consistency and Classification Accuracy

Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Classification Consistency (P)	0.87	0.92	0.99	0.78
Probability of Chance	0.50	0.63	0.96	0.35
Kappa (k)	0.73	0.79	0.68	0.66
Classification Accuracy	0.91	0.94	0.99	0.85

Table 9-16 Classification Consistency and Classification Accuracy for Mathematics Grade 8

Contingency Table with All Cut Scores

Performance Level	Below Basic	Basic	Proficient	Advanced	Sum
Below Basic	0.50	0.07	0.00	0.00	0.57
Basic	0.08	0.21	0.02	0.00	0.31
Proficient	0.00	0.03	0.08	0.00	0.11
Advanced	0.00	0.00	0.00	0.01	0.01
Sum	0.57	0.31	0.11	0.01	

Indexes for Classification Consistency and Classification Accuracy

Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Classification Consistency (P)	0.85	0.95	0.99	0.80
Probability of Chance	0.51	0.79	0.98	0.43
Kappa (k)	0.70	0.76	0.69	0.64
Classification Accuracy	0.90	0.96	0.99	0.86

Table 9-17 Classification Consistency and Classification Accuracy for Science Grade 4

Contingency Table with All Cut Scores

Performance Level	Below Basic	Basic	Proficient	Advanced	Sum
Below Basic	0.11	0.03	0.00	0.00	0.14
Basic	0.04	0.23	0.06	0.00	0.33
Proficient	0.00	0.06	0.21	0.06	0.34
Advanced	0.00	0.00	0.06	0.13	0.19
Sum	0.14	0.33	0.34	0.19	

Indexes for Classification Consistency and Classification Accuracy

Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Classification Consistency (P)	0.93	0.87	0.88	0.69
Probability of Chance	0.75	0.50	0.69	0.28
Kappa (k)	0.71	0.75	0.61	0.56
Classification Accuracy	0.95	0.91	0.92	0.78

Table 9-18 Classification Consistency and Classification Accuracy for Science Grade 8

Contingency Table with All Cut Scores

Performance Level	Below Basic	Basic	Proficient	Advanced	Sum
Below Basic	0.12	0.03	0.00	0.00	0.15
Basic	0.03	0.22	0.06	0.00	0.32
Proficient	0.00	0.07	0.23	0.06	0.36
Advanced	0.00	0.00	0.06	0.12	0.18
Sum	0.15	0.33	0.35	0.18	

Indexes for Classification Consistency and Classification Accuracy

Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Classification Consistency (P)	0.94	0.86	0.88	0.68
Probability of Chance	0.75	0.50	0.71	0.28
Kappa (k)	0.76	0.72	0.59	0.56
Classification Accuracy	0.96	0.90	0.91	0.77

Table 9-19 Classification Consistency and Classification Accuracy for Social Studies Grade 4

Contingency Table with All Cut Scores

Performance Level	Below Basic	Basic	Proficient	Advanced	Sum
Below Basic	0.14	0.04	0.00	0.00	0.18
Basic	0.04	0.14	0.06	0.00	0.24
Proficient	0.00	0.06	0.19	0.07	0.32
Advanced	0.00	0.00	0.06	0.20	0.26
Sum	0.18	0.24	0.32	0.27	

Indexes for Classification Consistency and Classification Accuracy

Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Classification Consistency (P)	0.92	0.87	0.87	0.67
Probability of Chance	0.71	0.51	0.61	0.26
Kappa (k)	0.73	0.73	0.66	0.55
Classification Accuracy	0.94	0.91	0.90	0.75

Table 9-20 Classification Consistency and Classification Accuracy for Social Studies Grade 8

Contingency Table with All Cut Scores

Performance Level	Below Basic	Basic	Proficient	Advanced	Sum
Below Basic	0.17	0.04	0.00	0.00	0.21
Basic	0.03	0.18	0.05	0.00	0.27
Proficient	0.00	0.06	0.20	0.05	0.31
Advanced	0.00	0.00	0.06	0.15	0.21
Sum	0.21	0.27	0.31	0.21	

Indexes for Classification Consistency and Classification Accuracy

Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Classification Consistency (P)	0.93	0.88	0.88	0.70
Probability of Chance	0.67	0.50	0.67	0.26
Kappa (k)	0.78	0.77	0.65	0.60
Classification Accuracy	0.94	0.92	0.92	0.78

Table 9-21 Classification Consistency and Classification Accuracy for Social Studies Grade 10

Contingency Table with All Cut Scores

<b>Performance Level</b>	<b>Below Basic</b>	<b>Basic</b>	<b>Proficient</b>	<b>Advanced</b>	<b>Sum</b>
<b>Below Basic</b>	0.21	0.04	0.00	0.00	0.25
<b>Basic</b>	0.05	0.15	0.06	0.00	0.25
<b>Proficient</b>	0.00	0.06	0.18	0.05	0.29
<b>Advanced</b>	0.00	0.00	0.05	0.16	0.21
<b>Sum</b>	0.26	0.25	0.28	0.22	

Indexes for Classification Consistency and Classification Accuracy

<b>Indexes</b>	<b>Cut 1</b>	<b>Cut 2</b>	<b>Cut 3</b>	<b>All Cuts</b>
<b>Classification Consistency (P)</b>	0.91	0.88	0.90	0.70
<b>Probability of Chance</b>	0.62	0.50	0.67	0.25
<b>Kappa (k)</b>	0.77	0.76	0.70	0.60
<b>Classification Accuracy</b>	0.94	0.91	0.93	0.78

Table 9-22 Inter-Rater Reliability, English Language Arts

Grade	Item No.	Max	Percentage of Agreement			Intra. Corr.	Weighted Kappa	Mean	Score Frequency					
			Exact	Adjacent	Discrepant				No. of Second Reads	1	2	3	4	Codes
3	4	4	85.68	13.82	0.50	0.79	0.59	1.21	25478	8489	1348	151	11	15479
4	6	4	80.88	18.46	0.66	0.84	0.69	1.36	18760	6739	1984	359	46	9632
5	4	4	80.99	18.70	0.31	0.80	0.61	1.20	15506	7161	2512	292	46	5495
6	4	4	77.92	21.57	0.51	0.85	0.70	1.50	11588	4445	2127	280	23	4713
7	4	4	77.57	22.05	0.39	0.88	0.77	1.72	10422	3835	2994	611	44	2938
8	5	4	80.87	18.64	0.49	0.89	0.78	1.49	15070	5881	2370	540	60	6219

Note: The sum of the modes of agreement and codes may not equal exactly 100% due to rounding.

## Part 10: Validity

---

Validity is the overarching component of the Wisconsin Forward Exam program. The following excerpt is from the *Standards for Educational and Psychological Testing* (hereafter the *Standards*; American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014):

Ultimately, the validity of an intended interpretation of test scores relies on all the available evidence relevant to the technical quality of a testing system. Different components of validity evidence . . . include evidence of careful test construction; adequate score reliability; appropriate test administration and scoring; accurate score scaling, equating, and standard setting; and careful attention to fairness for all test takers, as appropriate to the test interpretation in question. (22)

As stated by the *Standards*, the validity of a testing program hinges on the use of the test scores. Validity evidence that supports the interpretation and uses of the Wisconsin Forward Exam scores is provided in this Technical Report. The purpose of test score validation is not to validate the test itself, but to validate interpretations of the test scores for particular purposes or actions. Test score validation is not a quantifiable property but an ongoing process, beginning at initial conceptualization and continuing throughout the entire assessment process. Every aspect of an assessment provides evidence in support of (or a challenge to) the validity of an intended interpretation of test scores, including design, content specifications, item development, psychometric quality, and inferences made from the results.

As the Technical Report has progressed part by part, it has moved through the phases of the testing cycle. Each part of the Technical Report details the procedures and processes applied in the Wisconsin Forward Exam program, as well as the test results. Each part also highlights the meaning and significance of the procedures, processes, and results in terms of validity or a relationship to the *Standards*. Part 10 addresses four final issues related to the evidence of the validity of an intended interpretation of test scores: the issue of test fairness, evidence of validity based on the test internal structure, evidence of validity based on relationship with other variables, and test integrity. The analyses presented here add to the perspectives provided in Parts 2 through 9. Below is a brief review.

Part 2 of the Technical Report describes the the test blueprint and the involvement of Wisconsin educators, DPI, and DRC in the test development process. As indicated in Part 2, the test development process and the involvement of Wisconsin educators in that process forms an important part of the validity of the entire Wisconsin Forward Exam program. The knowledge, expertise, and professional judgment offered by Wisconsin educators ultimately ensures that the content of the Wisconsin Forward Exam forms an adequate and representative sample of appropriate content and that the content formed a legitimate basis upon which to derive valid conclusions about student achievement.

Part 3 of this report presents the test design and describes the key development tasks related to creating the Spring 2017 Wisconsin Forward Exam operational test forms. The test blueprint and item development activities described in Part 2 explain how specific development



processes provide evidence in support of the validity of an intended interpretation of test scores, primarily based on the test content and through the use of expert professional judgment from Wisconsin educators and from DRC test development specialists. The foundational documents—test blueprints and test designs—developed and approved during the initial phases of the test development served as critical guides throughout development and field testing of items. These documents contributed to ensuring that each form of the test accurately measured the content in consistent and stable ways, thus providing evidence supporting the test’s score use as an indicator of student achievement of state standards.

Parts 2 and 3 together provide evidence to support the validity of an intended interpretation of test scores based on test content of the Wisconsin Forward Exam and address AERA, APA, & NCME (2014) Standards 3.1, 3.2, 4.0, 4.1, 4.7, and 4.12.

Part 4 of the Technical Report describes the process, procedures, and policies that guided the administration of the Wisconsin Forward Exam, including accommodations, security, and the written procedures provided to test administrators and school personnel. The following AERA, APA, & NCME (2014) Standards are addressed: 4.15, 4.16, 6.1, 6.2, 6.3, 6.4, 6.6, and 6.7. The process, procedures, and policies detailed in this section contribute to the validity of an intended interpretation of test scores of the Wisconsin Forward Exam by reducing the impact of construct-irrelevant variables (e.g., nonstandardized administration methods, limitations associated with student disabilities, security breaches) on test performance.

Part 5 of the Technical Report demonstrates adherence to AERA, APA, & NCME (2014) Standards 4.18, 4.20, 6.8, and 6.9. It describes how MC, MS, ESR, SA, and TE auto-scored items, and TDA writing items were scored, including the handscoring process, the training and selection of scorers, the scoring rubrics used for scoring TDA items, and the resulting score distributions. The procedures described in this section contribute to the evidence of the validity of an intended interpretation of test scores of the Wisconsin Forward Exam by preventing hardware- or software-related errors in machine scoring and reducing construct-irrelevant score variance associated with variations in raters’ interpretation and application of scoring rubrics.

Part 6 describes the sample data used for the item calibration, test equating, and test scaling. The calibration, equating, and scaling methods as well as processes and procedures for deriving scale scores from response patterns are also described in this part of the Technical Report. Some references to introductory and advanced discussions of IRT are provided. Several axes upon which to evaluate the calibration, equating, and scaling procedures, such as the models and data used, the software applied, the vertical relationship across grades, the successful estimation of parameters, the fit, the SEM, and the IRT scoring method, are discussed. Part 6 of this report addresses AERA, APA, & NCME (2014) Standards 1.8, 2.13, 5.2, and 7.2. These processes and procedures contribute to the validity of an intended interpretation of test scores of the Wisconsin Forward Exam by providing the opportunity to evaluate items contributing to the accurate and reliable measurement of the intended constructs and by ensuring stability of the Wisconsin Forward Exam in its second administration year.

Part 7 of the Technical Report provides a brief summary of the Wisconsin Forward Exam standard setting, conducted in June 2016, during which the cut scores were set for all content

areas. The process of the standard setting adhered to AERA, APA, & NCME (2014) Standards 5.21 and 5.22, providing evidence of the procedural validity of the standard setting process, methodology, and outcomes.

Part 8 presents classical item analysis data, raw score results, scale score results, performance level information, and SPI scores. Scale score results provided a basic quantitative reference to student performance as derived through the IRT models applied. The performance level information reflected the performance level requirements of the DPI policy environment, as well as interests of parents, students, and educators. The SPI scores then probed further, assessing specific skills and abilities. Combined, scale scores, performance levels, and SPI scores provided a comprehensive set of tools to assess Wisconsin student performance by content and grade level and by gender, race/ethnicity, socioeconomic status, disability status, and English language proficiency. In addition, longitudinal evaluation of student performance on the tests is included in this part of the Technical Report. Part 8 thus addresses AERA, APA, & NCME (2014) Standards 1.8, 4.14, 5.1, 5.2, 5.21, 7.0, and 7.1. The analyses addressed in Part 8 contribute to the validity of an intended interpretation of test scores of the Wisconsin Forward Exam by providing further evidence of the tests being accurate and reliable measurements of the intended constructs.

Part 9 demonstrates adherence to AERA, APA, & NCME (2014) standards through several analyses of the reliability of the Spring 2017 Wisconsin Forward Exam. It presents a reliability analysis using Cronbach's alpha, SEM results, a detailed analysis of classification consistency and classification accuracy, and a full analysis of inter-rater reliability for TDA items. The Spring 2017 Wisconsin Forward Exam Technical Report thereby addresses AERA, APA, & NCME (2014) Standards 2.0, 2.3, 2.7, 2.11, 2.13, 2.14, and 2.16. Reliability is a prerequisite to score validity, and the analyses in that section contribute to the evidence of the validity of an intended interpretation of test scores by establishing the reliability of the Wisconsin Forward Exam scores and proficiency classifications.

In the subsequent pages, Part 10 will, as stated, present additional metrics with which to evaluate the validity of an intended interpretation of test scores of the Wisconsin Forward Exam program. As described below, the Wisconsin Forward Exam program formally assessed the issue of test fairness through an analysis of differential item functioning (DIF). It is possible for items to function differently across different population groups, and it is also possible that results for an item do not reflect student ability but instead reflect irrelevant information influenced by demographic factors. The DIF analysis provided below serves to determine whether that possibility occurred and, if so, to what degree, item by item, for each of the categories of gender, race/ethnicity, socioeconomic status, disability status, and English language proficiency.

This part is particularly relevant to AERA, APA, & NCME (2014) Standards 3.1 through 3.6. These standards are from Chapter 3 of the AERA, APA, & NCME (2014) *Standards* "Fairness in Testing." Each of these standards will be presented, as will be the way the standard is addressed in this part.

**Standard 3.6** Where credible evidence indicates that test scores may differ in meaning for relevant subgroups in the intended examinee population, test developers and/or users

are responsible for examining the evidence for validity of score interpretations for intended uses for individuals from those subgroups. What constitutes a significant difference in subgroup scores and what actions are taken in response to such differences may be defined by applicable laws. (65)

There is no particular research on the Wisconsin Forward Exam showing that the test scores of examinee subgroups differ in meaning; however, this is an ongoing concern in any large-scale testing program. To lessen the possibility of differences in test score meaning, DRC has several steps that are followed in item development and selection as is explained in Part 3. These practices adhere to Standard 3.3.

**Standard 3.3** Those responsible for test development should include relevant subgroups in validity, reliability/precision, and other preliminary studies used when constructing the test. (64)

DRC conducted DIF studies following the operational administration of the Wisconsin Forward Exam. Often items are evaluated for possible DIF in the field test phase of the test development, and items flagged for DIF are typically further examined for possible bias. In the case of the Wisconsin Forward Exam, the DIF analyses were conducted after the first operational test administration in Spring 2016. Items flagged for DIF were reviewed again by DRC content experts for potential bias and were avoided during the selection of the Spring 2017 forms. Only items deemed to be free of bias were included in the selection of the Spring 2017 forms. An additional DIF analysis was performed on the Spring 2017 operational test items. Items flagged for DIF were, again, evaluated by DRC content experts for potential bias. Section 10.1 of this part of the Technical Report explains the steps taken to evaluate the Wisconsin Forward Exam items through the use of DIF.

Section 3.2.3 of Part 3 discusses the form quality review conducted for the Wisconsin Forward Exam and the steps taken by DRC to minimize words, phrases, and content that may be regarded as offensive by members of particular demographic subgroups. This review is also critical in fulfilling AERA, APA, & NCME (2014) Standards 3.1 and 3.2.

**Standard 3.1** Those responsible for test development, revision, and administration should design all steps of the testing process to promote valid score interpretations for intended score uses for the widest possible range of individuals and relevant subgroups in the intended population. (63)

**Standard 3.2** Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests' being affected by construct-irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics. (64)

The present part of the report also provides the evidence of the validity of an intended interpretation of test scores related to test construct. Two measures are provided: correlations between content area objectives and principal components analysis. Both of these measures are provided to demonstrate the existence of a single, underlying trait or ability for each content

area, such as ELA ability or Mathematics ability. The presence of a single, underlying trait is a fundamental issue when scaling and analyzing results through IRT models. As such, these analyses are essential elements in assessing the validity of the Wisconsin Forward Exam. Next, the relationship between the Wisconsin Forward Exam scores and other variables is explored in order to support the evidence of the validity of an intended interpretation of test scores. These measures include evaluation of the correlations of the content area scores with other content area scores for the total population and by subgroups, as well as comparison of the student performance on the Wisconsin Forward Exam with the performance on the National Assessment of Educational Progress (NAEP). In addition, this chapter outlines the forensic analysis procedures that were employed to ensure the integrity of test scores by identifying schools and individual students who might have engaged in inappropriate behaviors during testing. Last but not least, a summary of standardized test administration procedures is provided as additional evidence supporting the validity of an intended interpretation of test scores.

## **10.1 Differential Item Functioning**

An empirical DIF approach was used to examine potential item bias and to determine whether item performance differences between identifiable subgroups were due to extraneous or construct-irrelevant information, making the items unfairly difficult for a particular subgroup in the student population. An item was flagged for DIF when there was a significant difference in the scores between a focal group of students and a reference group of students, with both groups at the same overall ability level. Thus, an item flagged for DIF is more difficult for a particular group of students than would be expected based on their total test scores (Camilli & Shepard, 1994; Green, 1975).

DIF analyses were conducted based on gender, race/ethnicity, socioeconomic status, disability status, and English language proficiency (ELP) groups. For the DIF analysis by gender, the reference group is male, meaning that the results for female students are considered with reference to male student performance. In the DIF analysis for race/ethnicity, the reference group is White. This means that the performance of students of each race/ethnicity is considered with reference to the performance of White students. The DIF analysis on socioeconomic status defines students identified as not economically disadvantaged as the reference group and students identified as economically disadvantaged as the focal group. The DIF analysis for disability status uses students identified as not disabled as a reference group to assess DIF within the student population identified as disabled. The DIF analysis for ELP compares item functioning among students identified as fully English proficient to those identified as limited English proficient. Students identified as fully English proficient comprise the reference group, and those identified as limited English proficient comprise the focal group.

Two DIF statistics that are commonly used for this purpose are the Mantel-Haenszel (MH) statistic (1959) and the Standardized Mean Difference (SMD) between the reference and focal groups, proposed by Dorans and Schmitt (1991).

The MH statistic is computed as follows (Zwick, Donoghue, & Grima, 1993):

$$\text{Mantel } \chi^2 = \frac{\left(\sum_k F_k - \sum_k E(F_k)\right)^2}{\sum_k \text{Var}(F_k)},$$

where  $F_k$  is the sum of scores for the focal group at the  $k$  level of the matching variable. Note that the MH statistic is sensitive to  $N$  such that larger sample sizes increase the value of chi square.

In addition to the MH chi-square statistic, the delta statistic (MH-D DIF) was computed for all items. Educational Testing Service (ETS) first developed the MH-D DIF statistic (Holland & Thayer, 1985, 1986). To compute delta, alpha (the odds ratio) is first computed:

$$\alpha_{MH} = \frac{\sum_{k=1}^K N_{r1k}N_{f0k} / N_k}{\sum_{k=1}^K N_{f1k}N_{r0k} / N_k},$$

where  $N_{r1k}$  is the number of correct responses in the reference group at ability level  $k$ ,  $N_{f0k}$  is the number of incorrect responses in the focal group at ability level  $k$ ,  $N_k$  is the total number of responses,  $N_{f1k}$  is the number of correct responses in the focal group at ability level  $k$ , and  $N_{r0k}$  is the number of incorrect responses in the reference group at ability level  $k$ . MH-D DIF is then computed:

$$\text{MH-D DIF} = -2.35 \ln(\alpha_{MH}).$$

For selected-response items, the MH ( $\chi_{MH}^2$ ) statistic was used to evaluate potential DIF items. In the MH procedure, subgroups are matched by their raw total test score using a contingency table with  $k$  ability levels. When applying the MH procedure, the log-odds ratio  $\alpha$  is assumed to be constant across the  $k$  matched levels. The  $\chi_{MH}^2$ , then, estimates a pooled common-odds ratio. Taking the natural logarithm of the common-odds ratio and its confidence limits and multiplying these with the constant  $-2.35$ , the resulting values may then be placed on the MH delta metric ( $\Delta_{MH}$ ) for interpretive purposes. Items were flagged for DIF using the following criteria:

- Moderate DIF: Significant MH chi-square statistic ( $p < 0.05$ ) and  $1.0 \leq |\text{MH D-DIF}| < 1.5$
- Large DIF: Significant MH chi-square statistic ( $p < 0.05$ ) and  $|\text{MH D-DIF}| \geq 1.5$

For constructed-response items, an effect size (ES) statistic based on the MH chi-square was used. The ES is obtained by dividing the SMD statistics by the standard deviation of the item. The SMD is an effect size index of DIF, which is relatively easy to interpret (Zwick et al., 1993). The SMD compares the mean of the reference and focal group, adjusting for the

distribution of the reference and focal group members on the conditioning variable (Zwick et al., 1993), which for these analyses is the Wisconsin Forward Exam raw score. SMD is computed as follows (Zwick et al., 1993):

$$SMD = p_{Fk} \left( \sum_k m_{Fk} - \sum_k m_{Rk} \right),$$

where  $p_{Fk}$  = proportion of the focal group members at the  $k$ th level of the matching variable,  $m_{Fk} = 1/N_{F1k}$ , and  $m_{Rk} = 1/N_{R1k}$ . Items are flagged using the same rules that are used in NAEP:

- Moderate DIF: If the MH statistic is significant ( $p < .05$ ) and  $|ES|$  is between 0.17 and 0.25
- Large DIF: If the MH statistic is significant ( $p < .05$ ) and  $|ES| \geq 0.25$

A positive DIF value indicates that the item favors the focal group, while a negative value indicates that the item disadvantages the focal group. Tables 10-1 through 10-7 show the DIF results for the following subgroups:

- **Gender:** Focal group is females; reference group is males.
- **Race/Ethnicity:** Focal groups are students whose race/ethnicity is reported as African-American, Hispanic, Asian, American Indian, or Two or More Ethnicities; reference group is students whose race/ethnicity is reported as White.
- **English Language Proficiency:** Focal group is students who are classified as not fully English language proficient; reference group is all others.
- **Disability Status:** Focal group is students with one or more disabilities; reference group is all others.
- **SES Status:** Focal group is students who are socioeconomically disadvantaged; reference group is all others.

A negative SMD value implies that the focal group has a lower mean item score than the reference group, whereas a positive value implies that the focal group has a higher mean item score than the reference group, conditioned on the matching test score.

The minimum case count for the focal group was set at 200, and the minimum case count for the reference group was set at 400. The DIF analyses were not performed for subgroups of fewer than 200 students. In these cases, the statistical procedures do not have sufficient power to detect differences should they exist.

Tables 10-1 through 10-7 show items that were flagged based on the criteria described above. The B flag represents a lower threshold for DIF. Only items that were flagged with a B or C flag were included in the tables described below.

The DIF results for gender are presented in Table 10-1; results for race/ethnicity are presented in Tables 10-2 through 10-5; English language proficiency (ELP) results are presented in Table 10-6; and results based on disability status are presented in Table 10-7. No operational test items were flagged for DIF based on SES.

Each DIF table references the grade and content area of the items flagged for DIF, as well as the item number on the test and the item type. The tables present the SMD statistics and the Mantel-Haenszel statistic ( $\Delta_{MH}$ ). After specifying these statistics for each item, the final column provides a flag status. The flag is based on SMD statistics for constructed-response items and on MH ( $\Delta_{MH}$ ) statistics.

In Table 10-1, looking at all items and all grades and content areas, 9 items were flagged for moderate (B flag) gender DIF in ELA; 8 items were flagged for moderate DIF and 1 item was flagged for large DIF (C flag) in Mathematics; 2 items were flagged for moderate DIF in Science; and 8 items were flagged for moderate DIF in Social Studies. Overall, 7 items were flagged in favor of the focal group (Females) and 21 items were flagged against the focal group. Of all items flagged for gender DIF, only one displayed large DIF (in grade 4 Mathematics) and 27 items displayed moderate DIF.

The other DIF results in Tables 10-2 through 10-7 can be understood in the same fashion. Note that a single item can be flagged for multiple subgroup categories, such as for ethnicity and language proficiency.

When looking at DIF results by item type, it was observed that most of the flagged items were MC items across all content areas and subgroups. The exceptions were DIF results for ELA conducted for subgroups of students with and without disabilities. As can be seen in Table 10-7, the majority of flagged items were either TDA or TE items. The items were flagged against students with disabilities.

The Spring 2017 Wisconsin Forward Exam was developed to minimize item and test bias. As stated earlier in this part of the Technical Report, all operational and field test items flagged for DIF in the Spring 2016 were reviewed by DRC's content experts for potential content-related bias. Only items deemed to be free of bias were included in the selection of the Spring 2017 forms. Items flagged for DIF after the Spring 2017 test administration were, again, evaluated by DRC content experts for potential bias.

Combined, the DIF statistical analyses discussed above and the expert reviews provide an appropriate set of tools with which to minimize the extraneous or construct-irrelevant information associated with item bias, or DIF, in the Wisconsin Forward Exam. It should be noted that in large-scale assessments, such as the Wisconsin Forward Exam, it is expected that some items will show DIF. All of the items in the Spring 2017 Wisconsin Forward Exam flagged for DIF were notated as such in the classical item analyses and in the item pool so that content experts would be able to reevaluate these items in future item selection activities. Items with DIF (particularly items flagged for strong DIF) are to be avoided in future selections.

## 10.2 Validity Evidence Based on Internal Test Structure

Construct-related evidence of the validity of an intended interpretation of test scores can be defined as the extent to which tests measure the skills or constructs they intend to measure, and is the central concept underlying the Spring 2017 Wisconsin Forward Exam validation process. Evidence for construct-related validity is comprehensive and integrates evidence from both content- and criterion-related validity. The Wisconsin Forward Exam test development process included specifications, item writing, review, and test construction.

Threats to construct-related validity include the unintended measurement of variables unrelated to the desired constructs and multidimensionality of the tests. To ensure that the test items are focused on the desired constructs, standardized procedures are employed to select items with sound statistical properties to align the items to content standards and to ensure that each test form meets the Wisconsin Forward Exam blueprint. A test can be said to be unidimensional when all of the items in the test measure the same underlying ability or trait.

### 10.2.1 Correlations between Content Standards

Analyses of the internal structure of a test can indicate the extent to which the relationships between test items and components conform to the construct the test purports to measure. For educational assessments that are designed to measure a single construct or content domain, the correlations between content standards within a test can be expected to be relatively high. Table 10-8 shows the correlations between main test domains for ELA, and Tables 10-9 through 10-12 show the correlations between content standards for each Wisconsin Forward Exam content area. The correlation coefficients here reflect the degree of linear relationship and direction between any two given content standards. The correlation can range from +1 to -1. A correlation of +1 indicates a perfect positive linear relationship, and a correlation of -1 indicates a perfect negative linear relationship between two content standards. A correlation of zero means there is no linear relationship. In general, the size of the correlation coefficient is influenced by the number of items or score points and by the score variance. Readers are cautioned not to confuse correlation with causation. The presence of a high correlation between two content standards should not be taken as an indication that there is a causal relationship between them.

As may be observed in Table 10-8, the correlations between the ELA main test domains of Reading, Writing, and Listening are moderate to high and range from 0.54 to 0.74 across all grades. With a few exceptions, the correlations between ELA content standards (see Table 10-9) are typically moderate for all grades and all standard pairs and range from 0.19 to 0.65. It should be noted, however, that the number of items in several content standards, particularly the standard C, measuring Reading - Vocabulary Use, was small, which was very likely a contributing factor to the lower correlations at the standard level compared to the correlations at the ELA domain level.

As indicated in Table 10-10, correlations between Mathematics content standards are also moderate to high and range from 0.52 to 0.74. The correlations between Science content standards range from 0.35 to 0.68 (see Table 10-11), and the correlations between Social Studies



content standards range from 0.52 to 0.71 (as shown in Table 10-12). Overall, the correlations for all content areas are within the moderate to high range.

Although it may be tempting to try to interpret the differences in magnitude within and across content areas, it is important to note that these correlations are highly dependent upon the numbers of items and the score variance for the different standards. The important finding is that within each content area the correlations between content standards are low enough to indicate that the standards are, as intended, somewhat distinct from one another but high enough to indicate that the individual standards are measuring related components of a single content area.

### **10.2.2 Principal Component Analysis**

Wisconsin Forward Exam items are calibrated using unidimensional IRT models, which posit that the test items are measuring an essentially unidimensional construct. To assess the dimensionality of the Wisconsin Forward Exam, a principal components analysis was conducted for each content area and grade. Principal components analysis is a statistical technique commonly used to evaluate dimensionality by detecting patterns of relationships among items. This method is useful in determining whether the observed scores on a test can be explained largely or entirely in terms of a much smaller number of components. For example, if answering the Mathematics items in a Mathematics test required a lot of reading ability, the Mathematics test would not be only a measure of mathematics ability, it would be a measure of reading ability as well. Such a test would be said to be multidimensional rather than essentially unidimensional. One way of evaluating the dimensions detected in the analysis is by examining the eigenvectors and eigenvalues. In principal components analysis, the eigenvectors correspond to factors, and the eigenvalues correspond to the variance explained by these factors. The sum of the eigenvalues is equal to the number of items in the test. The eigenvalues can be ordered from first to last in terms of the amount of the common variance that each explains. Data are generally considered to be unidimensional if the second eigenvalue is less than or equal to 1.0. Previous research shows that the examination of the ratio of the first two (i.e., the two largest) eigenvalues can be useful in determining the existence of dominant factors. Specifically, where large ratios exist between the first and second eigenvalues, a single dominant factor can be said to exist. Although the definition of “large” in the present context is subjective, the results in Table 10-13 show that the eigenvalue of the first factor, in most cases, is at least five times as large as the eigenvalue of the second factor.

As may be seen in Table 10-13, the ratios of the first two eigenvalues range from 4.91 to 7.42. The eigenvalues are proportional to the amount of common variance explained by each component, so these ratios indicate that the variance explained by the first component alone is approximately 5 to 7 times greater than the variance explained by the second component. The eigenvalue ratios range from 5.30 to 7.42 in ELA, from 4.91 to 6.91 in Mathematics, from 5.67 to 5.74 in Science, and from 5.95 to 7.22 in Social Studies. These ratios suggest that the unidimensionality of each of the Wisconsin Forward Exam content assessments is sufficient to meet the requirements of a unidimensional IRT calibration model.

Overall, these results provide support for the construct validity of the Wisconsin Forward Exam assessments. The correlations between content standards and the presence of a single

dominant factor for each test confirm that the content standards are sufficiently unidimensional to be combined into a single score.

### **10.3 Validity Evidence Based on Relationship with Other Variables**

The Wisconsin Forward Exam test score relationship with other variables was examined to further support the validity of the intended score interpretation. This was done using two measures: evaluation of correlations between the Wisconsin Forward Exam content area scores, and comparisons of the percentages of students classified in different proficiency levels (impact data) on the State assessment and on the NAEP assessment.

#### **10.3.1 Correlations between Content Area Test Scores**

The test score relationship with other variables can be assessed by the extent to which measures of constructs that theoretically should not be related to each other are, in fact, observed as not related to each other. Typically, correlation coefficients among measures of unrelated or distantly related constructs are examined in support of the validity evidence based on relationship of the test scores with other variables.

To assess the relationship between the Wisconsin Forward Exam content area scores, the correlations between the ELA, Mathematics, Science, and Social Studies scale scores for students who took more than one subject area test in 2017 were computed and examined for the total student population and by subgroups. Table 10-14 shows the correlations between the content area scores for the total population of Wisconsin students. These correlations ranged from 0.72 (between Mathematics and Science in grade 4, and between Mathematics and Social Studies in grades 4 and 8) to 0.81 (between ELA and Social Studies in grade 4, and between Social Studies and Science, also in grade 4).

Tables 10-15 through 10-19 show correlation coefficients between the content area scores by gender, ethnicity, English language proficiency status, SES status, and disability status, respectively. As seen in Table 10-15, the correlations between the content area scores for male or female groups ranged from 0.71 to 0.81 and were comparable for the two gender groups for each pair of correlated scores. The correlations between the content area scores for different ethnic groups ranged from 0.60 to 0.82 (see Table 10-16). The highest correlations by ethnic group were observed for White and Asian students. Correlations between the content area scores for the African-American student subgroup were lower than the correlations for other subgroups. As shown in Table 10-17, the correlations between the content area scores by English proficiency status ranged from 0.46 to 0.80. Lower correlations were observed for the group of students not fully English proficient compared to the fully English proficient group of students in all grade levels and for all pairs of correlated scores. The correlations between the content area scores by student socioeconomic status are presented in Table 10-18. These correlations ranged from 0.67 to 0.81 across all grades and pairs of correlated scores. The correlations between each pair of scores were comparable for the groups of students considered economically disadvantaged and non-economically disadvantaged in all grade levels. The correlations between the content area scores by student disability status are shown in Table 10-19. These correlations ranged from 0.55

to 0.80 across all grades and pairs of correlated scores. The correlations between each pair of scores were fairly comparable for the groups of students with and without disabilities in grades 3 and 4. In higher grades, between each pair of scores, correlations were lower for the group of students with disabilities compared to the group of students without disabilities.

Overall, the correlations between the content area scores for the total population of students were found to be highly related. The correlations between the content area scores for the subgroups of students were found to be moderately to highly related. Despite high correlations, the tests are not perfectly related to each other, suggesting that different constructs are being tapped; however, if the test scores are highly related to one another, they may be tapping into a similar knowledge base or general underlying ability.

### **Partial Correlations**

In addition to the simple correlations between the content area scores, partial correlations, which are measures of the strength of the relationship between the content area scores while controlling for the student demographic characteristics (gender, ethnicity, English proficiency status, disability status, and SES status), were also computed. Partial correlations allow for evaluation of the relationship two content area scores with the effect of the student demographic characteristics removed (or held constant). The partial correlations between the ELA, Mathematics, Science, and Social Studies test scores for the total population of students and at each grade level are presented in Table 10-20. These correlations ranged from 0.62 (between Mathematics and Science in grade 4, and between Mathematics and Social Studies in grades 4 and 8) to 0.75 (between ELA and Social Studies in grade 4, and between Social Studies and Science, also in grade 4). Although, the magnitude of these correlations is considered to be strong, as expected, the partial correlations between the content area scores were lower than the corresponding simple correlations, indicating that the student demographic characteristics did contribute to strength of the relationship between the content area test scores. The differences between the simple correlation and corresponding partial correlation coefficients were, however, relatively small, indicating that the effect of the student demographic characteristics on the relationship between the ELA, Mathematics, Science, and Social Studies test scores was small.

### **10.3.2 Comparison of the Wisconsin Forward Exam and Wisconsin NAEP Impact Data**

The NAEP is the largest nationally representative and continuing assessment of what America's students know and can do in various subject areas. Assessments in several content areas, including Reading, Mathematics, and Science, are administered to students in grades 4, 8, and 12 and conducted periodically. Representative samples of students from different states, including Wisconsin, participated in the latest NAEP assessment, which occurred in Spring 2017.

The main NAEP assessments are constructed using detailed frameworks that result from a comprehensive national process in which teachers, curriculum experts, policymakers, and members of the general public work to create a unified vision of how a particular subject ought to be assessed. This vision is based on current educational research on achievement and its

measurement, as well as good educational practices. These frameworks are updated about every decade in order to keep them current (for details, refer to <https://nces.ed.gov>).

The NAEP results are reported for all assessed content areas and for all participating grades at the national level. At the state level, the results for Reading, Mathematics, Science, and Writing are reported for grades 4 and 8. The results may also be reported at the district level (within a state) for these four content areas. No results are reported at the student level.

Wisconsin students participated in the last two NAEP assessments in Spring 2017 and Spring 2015. As of the time of this Technical Report development, the Spring 2017 state-level NAEP results are not yet available. Consequently, the Wisconsin Forward Exam state assessment results are compared to the Spring 2015 state-level NAEP results in Reading, Mathematics, and Science in grades 4 and 8. The percentages of Wisconsin students classified in different proficiency levels on the Wisconsin Forward Exam and the corresponding NAEP assessments are presented in Table 10-21. With two exceptions, the percentages of students classified in different performance levels on the NAEP assessments and on the Wisconsin Forward Exam were comparable within 8% or less for any performance level in both grades and all three content areas. The exceptions were percentages of students classified in the *Advanced* level for Science, where the differences were over 15% in grade 4 and over 11% in grade 8, with a larger percentage of students classified as *Advanced* on the Wisconsin Forward Exam compared to the NAEP Science assessment.

Looking at the percentages of students classified as *Proficient* or above, higher proportions of students were classified in these two combined categories on the Wisconsin Forward Exam in ELA and Science compared to the corresponding NAEP Reading and Science assessment. The opposite was true for Mathematics, where higher proportions of students were classified in the *Proficient* or above category on the NAEP Mathematics assessment compared to the Wisconsin Forward Exam in Mathematics for both grade levels. All differences were 10% or less.

When considering the percentages of students classified as *Basic* or above, higher proportions of students were classified in these three combined categories on the Wisconsin Forward Exam in grade 4 ELA and both Science grades compared to the corresponding NAEP Reading and Science assessments. More students were classified in the *Basic* or above category on the NAEP grade 8 Reading assessment and both grades of the Mathematics assessment compared to the Wisconsin Forward Exam in ELA and Mathematics for the corresponding grade levels. All differences were less than 8%.

It should be noted that the Spring 2015 Reading and Mathematics Wisconsin NAEP impact data were used as benchmarks during the Wisconsin Forward Exam standard setting after the Spring 2016 test administration. While the standard setting participants were free to deviate from the benchmarks while placing their bookmarks in the ordered item booklets in consideration of the Wisconsin performance level descriptors, the final Wisconsin impact data achieved after the standard setting were generally aligned with the Wisconsin state-level NAEP data. When considering the Wisconsin content standards and impact data articulation across grades, the Wisconsin Forward Exam cut scores for ELA, Mathematics, and Science remained in

alignment with the benchmarks, further supporting the evidence of the relationship between the state and the national assessments in these content areas.

#### **10.4 Test Integrity: Data Forensic Analyses**

With the high-stakes nature of large-scale statewide assessment programs there can be situations in which student responses, and hence their scores, may not be a true representation of students' own abilities. Various activities may take place, such as a student copying from another student's paper, students receiving inappropriate assistance before or during testing, or students' responses being altered during or after testing. To maintain the integrity of the Wisconsin Forward Exam and the validity of the results, it is important that any such instances be discovered.

Three studies were conducted to evaluate the Wisconsin Forward Exam student data for any indicators of possible inappropriate testing behavior. The first study examines incorrect student responses to multiple-choice items on the Spring 2017 Wisconsin Forward Exam ELA, Mathematics, Science, and Social Studies tests that were changed to correct responses. We refer to these answer changes as wrong-to-right (WTR) answer changes. Inordinate numbers of WTR answer changes in a specifically identifiable testing administration group may indicate inappropriate intervention on students' answer documents by an educator.

The second study evaluates students' time spent on the test and individual test items. These analyses serve to inform of any events in which students (e.g. within one schools) spent very short or very long time on the test or specific items. Inordinate numbers of unusual test or item response times may indicate inappropriate pre-knowledge of the items or other interventions during the testing session.

The results of the two studies are provided to DPI for evaluation. We emphasize that the results from these studies may be used in conjunction with other information to investigate whether inappropriate interventions may have taken place. The statistical results, by themselves, may simply be coincidental and do not necessarily indicate inappropriate behavior.

#### **10.5 Standardized Test Administration**

Unstandardized testing conditions can pose a serious threat to test validity by adding construct-irrelevant variance to the test scores. McCallin (2006) described a number of such threats to validity, including alterations in test administration requirements (e.g., changing time limits, modifying test instructions, giving hints to examinees), variability across test sites (e.g., differences in facilities/equipment, inadvertent posting of instructional aids in classrooms), interruptions during test sessions (e.g., power outages, relocation of students during testing, disturbances, other distractions), test administrator practices that may exacerbate test anxiety in particular students, practices that elicit test-wiseness, and security breaches that may result in the exposure of test forms or items. Construct-irrelevant variance may exert a systematic effect on

the scores of individual students or groups of students, resulting in an overestimation or underestimation of their true ability.

Standardized test administration, extensive training of the test scorers and AI engine, and rigorous scoring rules for auto-scored items for the Wisconsin Forward Exam comply with AERA, APA, & NCME (2014) Standards 3.4 and 3.5.

**Standard 3.4** Test takers should receive comparable treatment during the test administration and scoring process. (65)

**Standard 3.5** Test developers should specify and document provisions that have been made to test administration and scoring procedures to remove construct-irrelevant barriers for all relevant subgroups in the test-taker population. (65)

Taken together, the standardized Wisconsin Forward Exam test administration procedures described in Part 4 of this report were designed to address these potential threats to validity through the use of comprehensive security measures and the provision of detailed Test Administration Manuals and other training materials for District Assessment Coordinators, School Assessment Coordinators, and test administrators.

## 10.6 Summary

In summary, the overall purpose of Part 10 was to provide additional evidence of the validity of an intended interpretation of test scores related to test construct. Through the measures of correlations between content area objectives and principal components analysis, the existence of a single, underlying trait or ability for each content area was demonstrated. Next, the relationship between the Wisconsin Forward Exam scores and other variables was explored and validated through the measures of correlations of the content area scores with other content area scores for the total population and by subgroups, as well as comparisons of the student performance on the Wisconsin Forward Exam with the performance on the National Assessment of Educational Progress (NAEP). The forensic analysis procedures that were employed to ensure the integrity of test scores by identifying schools and individual students who might have engaged in inappropriate behaviors during testing were also described in this part of the report. In addition, a summary of standardized test administration procedures was provided as additional evidence supporting the validity of an intended interpretation of test scores.

Table 10-1 Items Flagged for DIF by Gender, Focal Group: Female

Content	Grade	Item Number	Item Type	MH SMD Statistic	MH Delta Statistic	DIF Flag
ELA	5	7	MC	-0.05	-1.25	B-
	5	28	MC	-0.10	-1.41	B-
	6	4	TDA	0.14		B
	6	6	MC	-0.04	-1.25	B-
	6	7	MC	-0.07	-1.21	B-
	7	4	TDA	0.18		B
	7	14	MC	-0.10	-1.25	B-
	7	23	MC	-0.08	-1.03	B-
	8	5	TDA	0.18		B
Math	3	4	MC	-0.07	-1.18	B-
	3	19	MC	-0.06	-1.04	B-
	4	1	MC	-0.09	-1.53	C-
	4	7	MC	-0.09	-1.20	B-
	4	39	MC	-0.07	-1.05	B-
	5	5	MC	-0.09	-1.20	B-
	6	7	MC	-0.02	-1.03	B-
	6	25	MC	0.03	1.12	B
	6	32	MC	-0.08	-1.02	B-
Science	4	22	MC	-0.04	-1.23	B-
	8	20	MC	0.07	1.00	B
Social Studies	8	10	MC	-0.06	-1.03	B-
	8	22	MC	-0.09	-1.05	B-
	8	37	MC	0.07	1.09	B
	10	8	MC	-0.08	-1.01	B-
	10	11	MC	-0.10	-1.20	B-
	10	34	MC	0.06	1.25	B
	10	35	MC	-0.05	-1.14	B-
	10	39	MC	-0.09	-1.26	B-

Table 10-2 Items Flagged for DIF by Race/Ethnicity, Focal Group: African-American

Content	Grade	Item Number	Item Type	MH SMD Statistic	MH Delta Statistic	DIF Flag
ELA	5	7	MC	-0.11	-1.59	C-
	5	28	MC	-0.10	-1.26	B-
	8	2	ESR	0.20		B
	8	7	MC	-0.09	-1.09	B-
Math	3	5	MC	-0.08	-1.15	B-
	5	7	SA	-0.10	-1.20	C-
	6	7	MC	-0.07	-1.31	B-
	8	16	TE	-0.12	-2.03	B-
Science	8	6	MC	-0.07	-1.03	B-
	8	8	MC	-0.07	-1.04	B-
Social Studies	4	4	MC	0.09	1.07	B
	4	8	MC	0.07	1.09	B
	4	13	MC	0.10	1.22	B
	4	18	MC	-0.09	-1.42	B-
	8	13	MC	-0.08	-1.06	B-
	8	22	MC	-0.10	-1.45	B-
	8	36	MC	-0.12	-1.68	C-
	10	50	MC	-0.08	-1.01	B-



Table 10-3 Items Flagged for DIF by Race/Ethnicity, Focal Group: Hispanic

<b>Content</b>	<b>Grade</b>	<b>Item Number</b>	<b>Item Type</b>	<b>MH SMD Statistic</b>	<b>MH Delta Statistic</b>	<b>DIF Flag</b>
ELA	5	7	MC	-0.09	-1.57	C-
	6	6	MC	-0.06	-1.11	B-
Science	8	8	MC	-0.07	-1.25	B-
Social Studies	4	7	MC	-0.06	-1.01	B-
	8	18	MC	0.12	1.41	B
	10	50	MC	-0.09	-1.16	B-

Table 10-4 Items Flagged for DIF by Race/Ethnicity, Focal Group: Asian

Content	Grade	Item Number	Item Type	MH SMD Statistic	MH Delta Statistic	DIF Flag
ELA	5	4	TDA	0.11		B
	5	7	MC	-0.12	-2.37	C-
	5	16	MC	0.10	1.18	B
	5	20	MC	-0.09	-1.08	B-
	5	28	MC	-0.09	-1.19	B-
	6	6	MC	-0.04	-1.06	B-
	6	7	MC	-0.06	-1.02	B-
	6	12	TE	0.13		B
	7	19	MC	-0.09	-1.10	B-
8	25	MC	-0.10	-1.73	C-	
Math	3	4	MC	-0.07	-1.11	B-
	3	7	MC	0.06	1.39	B
	3	37	MC	0.07	1.11	B
	4	2	MC	-0.11	-1.30	B-
	4	29	TE	0.09	1.42	B
	5	5	MC	-0.08	-1.04	B-
	5	20	TE	0.10	1.37	B
	6	7	MC	-0.03	-1.48	B-
	6	34	MC	-0.09	-1.18	B-
	7	17	MC	0.08	1.09	B
	7	31	MC	-0.06	-1.03	B-
8	17	MC	-0.08	-1.05	B-	
Science	4	34	MC	0.06	1.10	B
	8	6	MC	-0.05	-1.14	B-
	8	8	MC	-0.06	-1.32	B-
	8	30	MC	0.07	1.13	B
Social Studies	4	7	MC	-0.09	-1.77	C-
	4	30	MC	-0.08	-1.07	B-
	8	10	MC	-0.07	-1.12	B-
	8	12	MC	0.07	1.13	B
	8	24	MC	-0.11	-1.80	C-
	8	35	MC	0.07	1.05	B
	8	36	MC	-0.08	-1.29	B-
	8	40	MC	0.07	1.56	C
	10	1	MC	-0.07	-1.33	B-
10	50	MC	-0.07	-1.11	B-	

Table 10-5 Items Flagged for DIF by Race/Ethnicity, Focal Group: American Indian

Content	Grade	Item Number	Item Type	MH SMD Statistic	MH Delta Statistic	DIF Flag
ELA	5	7	MC	-0.08	-1.43	B-
Math	4	41	MC	-0.04	-1.00	B-
Social Studies	8	20	MC	0.08	1.06	B

Table 10-6 Items Flagged for DIF by English Language Proficiency, Focal Group: Students Not English Language Proficient

Content	Grade	Item Number	Item Type	MH SMD Statistic	MH Delta Statistic	DIF Flag
ELA	3	25	MC	-0.09	-1.07	B-
	5	7	MC	-0.14	-1.74	C-
	5	18	TE	-0.12		B-
	6	6	MC	-0.08	-1.07	B-
	8	7	MC	-0.09	-1.05	B-
	8	11	TE	-0.13		B-
	8	25	MC	-0.11	-1.33	B-
Science	8	8	MC	-0.12	-1.42	B-
Social Studies	4	7	MC	-0.13	-1.66	C-
	8	10	MC	-0.11	-1.25	B-
	8	18	MC	0.11	1.18	B
	10	1	MC	-0.09	-1.05	B-
	10	50	MC	-0.12	-1.28	B-

Table 10-7 Items Flagged for DIF by Disability Status, Focal Group: Students with One or More Disabilities

Content	Grade	Item Number	Item Type	MH SMD Statistic	MH Delta Statistic	DIF Flag
ELA	4	19	TE	-0.09		B-
	5	4	TDA	-0.14		B-
	5	28	MC	0.09	1.14	B
	6	12	TE	-0.17		B-
	7	4	TDA	-0.15		B-
	8	11	TE	-0.17		C-
Math	3	33	SA	-0.06	-0.96	B-
	6	7	MC	-0.10	-1.68	C-
	7	1	MC	0.06	1.24	B
Science	4	17	MC	-0.10	-1.33	B-
	4	21	MC	-0.07	-1.09	B-

Table 10-8 Correlations among English Language Arts Test Domains

Grade	ELA Domain	Listening	Reading
3	Reading	0.64	
	Writing	0.59	0.70
4	Reading	0.66	
	Writing	0.61	0.74
5	Reading	0.61	
	Writing	0.58	0.68
6	Reading	0.59	
	Writing	0.54	0.70
7	Reading	0.69	
	Writing	0.63	0.73
8	Reading	0.63	
	Writing	0.60	0.73

Table 10-9 Correlations among English Language Arts Standards

Grade	Standard Code	A	B	C	D	E	F
3	B	0.62					
	C	0.47	0.46				
	D	0.55	0.53	0.40			
	E	0.52	0.49	0.39	0.48		
	F	0.38	0.37	0.29	0.35	0.31	
	G	0.57	0.57	0.43	0.52	0.50	0.35
4	B	0.57					
	C	0.63	0.53				
	D	0.55	0.45	0.49			
	E	0.61	0.51	0.54	0.50		
	F	0.50	0.41	0.45	0.41	0.47	
	G	0.61	0.51	0.55	0.49	0.53	0.43
5	B	0.56					
	C	0.41	0.45				
	D	0.44	0.50	0.40			
	E	0.45	0.50	0.37	0.46		
	F	0.46	0.52	0.40	0.49	0.50	
	G	0.51	0.54	0.44	0.48	0.46	0.49
6	B	0.65					
	C	0.27	0.31				
	D	0.43	0.45	0.19			
	E	0.53	0.53	0.24	0.35		
	F	0.52	0.54	0.23	0.38	0.49	
	G	0.53	0.55	0.23	0.38	0.45	0.45
7	B	0.64					
	C	0.46	0.46				
	D	0.52	0.48	0.37			
	E	0.60	0.53	0.39	0.47		
	F	0.46	0.44	0.34	0.40	0.42	
	G	0.62	0.60	0.47	0.50	0.53	0.46
8	B	0.65					
	C	0.42	0.45				
	D	0.55	0.57	0.37			
	E	0.57	0.56	0.37	0.51		
	F	0.50	0.51	0.34	0.51	0.49	
	G	0.56	0.58	0.37	0.53	0.50	0.46

Note: Standard Codes are as follows: A = Reading - Key Ideas and Details; B = Reading - Craft & Structure/Integration of Knowledge & Ideas; C = Reading - Vocabulary Use; D = Writing/Language - Text Types and Purpose; E = Writing/Language - Research; F = Writing/Language - Language Conventions; G = Listening

Table 10-10 Correlations among Mathematics Standards

Grade	Standard Code	A	B	C	D	E	F	G	H	I
3	B	0.72								
	C	0.60	0.61							
	D	0.70	0.70	0.64						
	E	0.59	0.59	0.58	0.62					
4	B	0.67								
	C	0.63	0.69							
	D	0.65	0.71	0.74						
	E	0.52	0.55	0.62	0.64					
5	B	0.67								
	C	0.67	0.64							
	D	0.63	0.59	0.65						
	E	0.67	0.61	0.62	0.61					
6	F					0.53				
	G					0.64	0.68			
	H					0.64	0.63	0.73		
	I					0.53	0.54	0.62	0.61	
7	F					0.59				
	G					0.62	0.66			
	H					0.64	0.64	0.68		
	I					0.64	0.67	0.68	0.70	
8	G					0.57				
	H					0.64		0.62		
	I					0.62		0.54	0.66	
	J					0.65		0.58	0.71	0.71

Note: Standard Codes are as follows: A = Operations and Algebraic Thinking; B = Number and Operations in Base Ten; C = Number and Operations - Fractions; D = Measurement and Data; E = Geometry; F = Ratios and Proportional Relationships; G = The Number System; H = Expressions and Equations; I = Statistics and Probability; J = Functions

Table 10-11 Correlations among Science Standards

Grade	Standard Code	A/B	C	D	E	F
4	C	0.65				
	D	0.40	0.43			
	E	0.46	0.50	0.35		
	F	0.54	0.56	0.40	0.44	
	G/H	0.65	0.66	0.44	0.49	0.59
8	C	0.68				
	D	0.57	0.57			
	E	0.47	0.47	0.42		
	F	0.54	0.57	0.47	0.40	
	G/H	0.60	0.64	0.52	0.43	0.52

Note: Standard Codes are as follows: A/B = Science Connections & Nature of Science; C = Science Inquiry; D = Physical Science; E = Earth and Space Science; F = Life & Environmental Science; G/H = Science Applications & Social and Personal Perspectives

Table 10-12 Correlations among Social Studies Standards

Grade	Standard Code	A	B	C	D
4	B	0.66			
	C	0.58	0.58		
	D	0.57	0.57	0.52	
	E	0.63	0.61	0.56	0.56
8	B	0.70			
	C	0.63	0.66		
	D	0.66	0.68	0.62	
	E	0.61	0.63	0.59	0.58
10	B	0.70			
	C	0.70	0.71		
	D	0.64	0.64	0.67	
	E	0.65	0.67	0.68	0.62

Note: Standard Codes are as follows: A = Geography; B = History; C = Political Science and Citizenship; D = Economics; E = The Behavioral Sciences

Table 10-13 Principal Components Analysis

Content Area	Grade	First Eigenvalue	Second Eigenvalue	Ratio of First Two Eigenvalues
ELA	3	6.792	1.167	5.822
	4	7.962	1.229	6.478
	5	6.982	1.318	5.298
	6	6.830	1.179	5.795
	7	7.711	1.148	6.718
Mathematics	8	7.876	1.061	7.421
	3	9.417	1.559	6.040
	4	10.691	1.547	6.910
	5	10.010	1.612	6.211
	6	9.738	1.982	4.912
Science	7	10.040	1.615	6.217
	8	9.666	1.615	5.986
Social Studies	4	7.570	1.335	5.672
	8	7.846	1.367	5.739
	10	9.927	1.374	7.224

Table 10-14 Correlations between Content Area Scale Scores

Grade	ELA & Mathematics	ELA & Science	ELA & Social Studies	Mathematics & Science	Mathematics & Social Studies	Science & Social Studies
3	0.74					
4	0.73	0.80	0.81	0.72	0.72	0.81
5	0.73					
6	0.77					
7	0.73					
8	0.73	0.78	0.79	0.73	0.72	0.80



Table 10-15 Correlations between Content Area Scale Scores by Gender

Grade	Demographic Group	ELA & Math	ELA & Science	ELA & Social Studies	Math & Science	Math & Social Studies	Science & Social Studies
3	Female	0.75					
	Male	0.74					
4	Female	0.74	0.81	0.81	0.72	0.72	0.81
	Male	0.74	0.80	0.81	0.71	0.72	0.81
5	Female	0.74					
	Male	0.74					
6	Female	0.77					
	Male	0.78					
7	Female	0.74					
	Male	0.73					
8	Female	0.73	0.78	0.80	0.74	0.72	0.80
	Male	0.74	0.78	0.80	0.73	0.72	0.80

Table 10-16 Correlations between Content Area Scale Scores by Ethnicity/Race

Grade	Demographic Group	ELA & Math	ELA & Science	ELA & Social Studies	Math & Science	Math & Social Studies	Science & Social Studies
3	White	0.70					
	African-American	0.65					
	Hispanic	0.68					
	Asian	0.75					
	American Indian	0.70					
	Two or More	0.73					
4	White	0.70	0.75	0.78	0.67	0.67	0.77
	African-American	0.63	0.75	0.74	0.61	0.62	0.76
	Hispanic	0.68	0.78	0.79	0.68	0.67	0.81
	Asian	0.74	0.80	0.81	0.74	0.74	0.81
	American Indian	0.66	0.77	0.78	0.67	0.65	0.81
	Two or More	0.71	0.80	0.80	0.71	0.70	0.81
5	White	0.70					
	African-American	0.62					
	Hispanic	0.66					
	Asian	0.75					
	American Indian	0.65					
	Two or More	0.70					
6	White	0.73					
	African-American	0.70					
	Hispanic	0.73					
	Asian	0.78					
	American Indian	0.71					
	Two or More	0.76					
7	White	0.71					
	African-American	0.60					
	Hispanic	0.66					
	Asian	0.73					
	American Indian	0.66					
	Two or More	0.70					
8	White	0.72	0.75	0.77	0.71	0.69	0.77
	African-American	0.61	0.73	0.74	0.61	0.60	0.73
	Hispanic	0.67	0.75	0.77	0.67	0.66	0.79
	Asian	0.75	0.77	0.78	0.77	0.75	0.82
	American Indian	0.65	0.73	0.77	0.65	0.66	0.77
	Two or More	0.72	0.77	0.79	0.72	0.71	0.78

Table 10-17 Correlations between Content Area Scale Scores by English Proficiency Status

<b>Grade</b>	<b>Demographic Group</b>	<b>ELA &amp; Math</b>	<b>ELA &amp; Science</b>	<b>ELA &amp; Social Studies</b>	<b>Math &amp; Science</b>	<b>Math &amp; Social Studies</b>	<b>Science &amp; Social Studies</b>
3	Fully English Proficient	0.74					
	Limited English Proficiency	0.65					
4	Fully English Proficient	0.73	0.79	0.80	0.71	0.71	0.80
	Limited English Proficiency	0.60	0.71	0.71	0.61	0.61	0.75
5	Fully English Proficient	0.73					
	Limited English Proficiency	0.54					
6	Fully English Proficient	0.76					
	Limited English Proficiency	0.60					
7	Fully English Proficient	0.72					
	Limited English Proficiency	0.46					
8	Fully English Proficient	0.73	0.77	0.79	0.73	0.72	0.79
	Limited English Proficiency	0.49	0.61	0.61	0.53	0.50	0.68

Table 10-18 Correlations between Content Area Scale Scores by SES Status

Grade	Demographic Group	ELA & Math	ELA & Science	ELA & Social Studies	Math & Science	Math & Social Studies	Science & Social Studies
3	Economically Disadvantaged	0.70					
	Not Economically Disadvantaged	0.71					
4	Economically Disadvantaged	0.69	0.79	0.79	0.68	0.69	0.81
	Not Economically Disadvantaged	0.70	0.76	0.78	0.68	0.68	0.77
5	Economically Disadvantaged	0.68					
	Not Economically Disadvantaged	0.71					
6	Economically Disadvantaged	0.73					
	Not Economically Disadvantaged	0.74					
7	Economically Disadvantaged	0.67					
	Not Economically Disadvantaged	0.71					
8	Economically Disadvantaged	0.68	0.76	0.78	0.68	0.67	0.79
	Not Economically Disadvantaged	0.72	0.74	0.76	0.71	0.69	0.77

Table 10-19 Correlations between Content Area Scale Scores by Disability Status

Grade	Demographic Group	ELA & Math	ELA & Science	ELA & Social Studies	Math & Science	Math & Social Studies	Science & Social Studies
3	Disabled	0.70					
	Not Disabled	0.73					
4	Disabled	0.68	0.77	0.77	0.67	0.67	0.80
	Not Disabled	0.72	0.79	0.80	0.71	0.71	0.80
5	Disabled	0.63					
	Not Disabled	0.71					
6	Disabled	0.68					
	Not Disabled	0.74					
7	Disabled	0.55					
	Not Disabled	0.71					
8	Disabled	0.59	0.70	0.70	0.60	0.58	0.74
	Not Disabled	0.71	0.76	0.77	0.72	0.70	0.78

Table 10-20 Partial Correlations between Content Area Scale Scores

Grade	ELA & Mathematics	ELA & Science	ELA & Social Studies	Mathematics & Science	Mathematics & Social Studies	Science & Social Studies
3	0.67					
4	0.65	0.74	0.75	0.62	0.62	0.75
5	0.64					
6	0.68					
7	0.64					
8	0.64	0.71	0.73	0.63	0.62	0.73

Table 10-21 Comparison of Spring 2015 Wisconsin NAEP and Spring 2017 Wisconsin Forward Exam Impact Data

Content	Grade	Wisconsin NAEP Spring 2015 Percentages of Students						Wisconsin Forward Exam Spring 2017 Percentages of Students					
		Below Basic	Basic	Proficient	Advanced	At or Above Proficient	At or Above Basic	Below Basic	Basic	Proficient	Advanced	At or Above Proficient	At or Above Basic
<b>Reading/ ELA</b>	4	29	34	29	8	37	71	21.14	32.14	37.00	9.71	46.72	78.86
<b>Reading/ ELA</b>	8	21	40	35	4	39	79	21.66	37.22	29.19	11.93	41.12	78.34
<b>Math</b>	4	17	37	36	9	45	82	19.13	37.37	32.67	10.83	43.50	80.87
<b>Math</b>	8	22	37	30	11	41	78	28.43	36.95	28.33	6.29	34.62	71.57
<b>Science</b>	4	21	38	40	1	41	79	15.29	33.63	34.70	16.37	51.07	84.71
<b>Science</b>	8	25	35	38	2	40	75	17.61	34.74	34.11	13.54	47.65	82.39

\* NEAP assessed student knowledge and skills in Reading while Wisconsin Forward Exam assessed student knowledge and skills in ELA, which included Reading, Listening, and Writing

## Part 11: Summary Recommendations

---

Results and key findings of the Spring 2017 Wisconsin Forward Exam test administration are presented throughout the body of this report. This last section of the report presents some recommendations for DPI consideration.

The 2017 Wisconsin Forward Exam administration was the second administration of the assessment. The assessment results were reported on the same scales and students were classified into the proficiency levels using the same cut scores for two consecutive years, allowing for longitudinal tracking of student performance. We recommend continuing to use the same scales and the same cut scores for Wisconsin assessments and monitoring student growth in the upcoming administration years.

Following the Spring 2016 and 2017 field test of new test items in Wisconsin, we recommend that, in the future, all items be field tested in Wisconsin prior to their operational test administration to provide accurate information on how students may perform on these items once they are administered operationally. We recommend continuing to develop and embed field test items in each operational test administration for all content areas in order to build a high-quality Wisconsin item bank for future form development.

DRC also recommends continuing to use an artificial intelligence (AI) engine in the scoring of text-dependent analysis items for its efficiency and accuracy. As indicated in Part 5 and Part 9 of this report, the AI scores were in very high agreement with scores by trained human scorers.

From the psychometric perspective, it was noticed that the ELA grade 5 test continues to be relatively difficult for grade 5 students. The properties of the ELA vertical scale described in Part 6 of this report indicate that the ELA grade 5 and grade 6 tests were of similar difficulty, as indicated by the test characteristic curves. In order to achieve better ordinality of the ELA assessments' overall difficulty across grade levels, easier items could be added to the grade 5 test. However, it should be noted that because equating requires tests to maintain a similar level of difficulty from year to year, increasing or decreasing the test rigor would likely require a cut score review and an examination regarding whether a new test scale should be set.

Several items, particularly in higher grades of Mathematics assessments, were found to be very difficult for Wisconsin students. While use of some difficult items may be necessary to fulfill the test content specifications, both DPI and DRC recommend careful review of these items and determination whether they should be included in the Wisconsin Forward Exam item bank for future use or be removed and replaced with other items measuring the same content standards.

In addition, DRC recommends continuing to compare the Wisconsin Forward Exam results for grades 4 and 8 with the most recent and available state-level NAEP data in order to monitor the alignment of impact data between the state and national assessments.

## References

---

- Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- 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. (2002). PARDUX [Computer program]. Unpublished.
- Camilli, G., & Shepard, A. L. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage Publications.
- Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing bias in item response theory. *Applied Psychological Measurement*, 12(3), 253–260.
- Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage Publications.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth Group/Thomson Learning.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334.
- CTB/McGraw-Hill. (1997). *TerraNova* (1st ed.). Monterey, CA: Author.
- CTB/McGraw-Hill. (2000). *TerraNova* (2nd ed.). Monterey, CA: Author.
- CTB/McGraw-Hill. (2009). *TerraNova 3rd Edition Technical Addendum: Forms E and F*. Monterey, CA: Author.
- Dorans, N. J., & Schmitt, A. P. (1991). *Constructed response and differential item functioning: A pragmatic approach*. Princeton, NJ: Educational Testing Service.
- Fitzpatrick, A. R. (1991). *Status report on the results of preliminary analyses of dichotomous and multi-level items using the PARMATE program*. Monterey, CA: CTB/McGraw-Hill.



- Fitzpatrick, A. R., & Julian, M. W. (1996). *Two studies comparing the parameter estimates produced by PARDUX and PARSCALE*. Unpublished manuscript.
- Green, D. R. (1975). *Procedures for assessing bias in achievement tests*. Paper presented at the National Institute of Education Conference on Test Bias, Annapolis, MD.
- Hambleton, R. K., & Novick, M. R. (1973). Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement, 10*(3), 159–170.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Hingham, MA: Kluwer-Nijhoff Publishing.
- Holland, P. W., & Thayer, D. T. (1985). *An alternate definition of the ETS delta scale of item difficulty* (Research Report RR-85-43). Princeton, NJ: Educational Testing Service.
- Holland, P. W., & Thayer, D. T. (1986). *Differential item performance and the Mantel-Haenszel procedure*. Paper presented at the annual meeting of the American Educational Research Association Annual Meeting, San Francisco, CA.
- Karantonis, A., & Sireci, S. G. (2006). The Bookmark standard-setting method: A literature review. *Educational Measurement: Issues and Practice, 25*(1), 4–12.
- Kim, D. (2005). KKCLASS [Computer program]. Unpublished.
- Kim, D., Barton, K., & Kim, J. (2007). *Estimating classification consistency and classification accuracy with pattern scoring*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.
- Kim, D., Choi, S., Um, K., & Kim, J. (2006). *A comparison of methods for estimating classification consistency*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.
- Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices*. New York, NY: Springer-Verlag.
- Kolen, M., & Kim, D. (2004). [Personal correspondence].
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.
- Linn, R. L. (Ed.) (1989). *Educational measurement* (3rd ed.). New York, NY: Macmillan.
- Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32*(2), 179–197.

- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22(4), 719–748.
- McCallin, R. C. (2006). Test Administration. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 625–652). Mahwah, NJ: Lawrence Erlbaum Associates.
- Michaelides, M. P., & Haertel, E. H. (2004). *Sampling of common items: An unrecognized source of error in test equating*. Los Angeles, CA: Center for the Study of Evaluation.
- Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement*, 14(1), 59–71.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16(2), 159–176.
- Muraki, E., & Bock, R. D. (1991). PARSCALE: Parameter scaling of rating rata [Computer program]. Chicago, IL: Scientific Software, Inc.
- Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7(2), 201–210.
- Swaminathan, H., Hambleton, R. K., & Algina, J. (1974). Reliability of criterion-referenced tests: A decision theoretic formulation. *Journal of Educational Measurement*, 11(4), 263–267.
- Swineford, F. (1956). *Technical manual for users of test analysis* (Statistical Report 56–42). Princeton, NJ: Educational Testing Service.
- Thissen, D. (1982). Marginal maximum-likelihood estimation for the one-parameter logistic model. *Psychometrika*, 47(2), 175–186.
- Thissen, D. (1990). MULTILOG: Multiple categorical item analysis and test scoring (Version 6) [Computer program]. Chicago, IL: Scientific Software, Inc.
- Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large-scale assessment* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- van der Linden, W. J., & Hambleton, R. K. (Eds.) (1997). *Handbook of modern item response theory*. New York: Springer.

- Wright, B. D., & Linacre, J. M. (1992). BIGSTEPS Rasch Analysis [Computer program]. Chicago, IL: MESA Press.
- Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, 5(2), 245–262.
- Yen, W. M. (1984). Obtaining maximum likelihood trait estimates from number-correct scores for the three-parameter logistic model. *Journal of Educational Measurement*, 21(2), 93–111.
- Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30(3), 187–213.
- Yen, W. M., & Burket, G. R. (1997). Comparison of item response theory and Thurstone methods of vertical scaling. *Journal of Educational Measurement*, 34(4), 293–313.
- Yen, W. M., & Candell, G. L. (1991). Increasing score reliability with item-pattern scoring: An empirical study in five score metrics. *Applied Measurement in Education*, 4(3), 209–228.
- Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, 30(3), 233–251.

## **Appendix A**

### **Spring 2016 Field Test Data Review Training Slides**

# Wisconsin Forward Data Review

Sept. 26-27, 2016

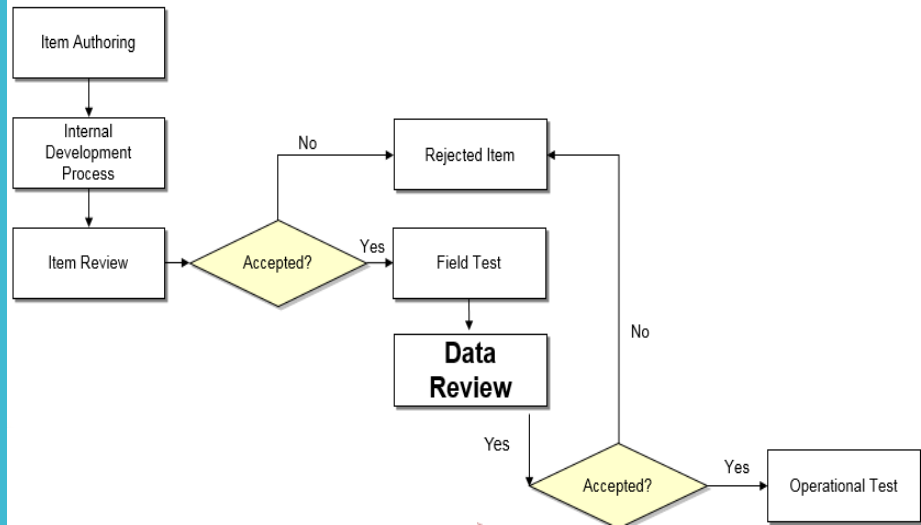


## Meeting Goals/Key Objectives

- Review Data of Flagged Items from Spring 2016 WI Forward Exam
  - Item data card layout
  - Understand and interpret statistics
- Review Data of Field-Test Items from Spring 2016 WI Forward Exam
- Receive DPI Approval for placement on Spring 2017 WI Forward Exam



# Life of an Item



# Test Development Summary



# Item Data Review Process

Step 1: Item	Step 2: Difficulty	Step 3: Discrimination	Step 4: Distractors / Steps	Step 5: DIF	Step 6: Decision
•Carefully read the item stem and/or all response options	•Determine the item difficulty	•Determine the item's ability to differentiate between low and high achievers	•Determine the proportion of each distractor/score point and its item total corr	•Determine if the item is potentially biased to the groups of interest	•Decide to ACCEPT or REJECT the item
•Item key, standard, and depth of knowledge	•P-Value	•Item Total Correlation	•Proportion •Correlation	•Bias Code	•ACCEPT •REJECT
•MC: Pay attention to the distractors	•0.30 – 0.90	•Above 0.25	•MC: Corr should be $\leq 0.0$ • Proportion should be $\geq 0.05$	•Potentially reject C- or C+ items if can posit a reason	Better to approve items with <b>marginal statistics</b> if no clear flaw is evident.



# Item Data Card

Standard: Determine if a relation is a function given a set of points or a graph.

1. The sets of ordered pairs (x, y) form a relation. Determine if the relation is a function.

Which ordered pair could be included in the set so that the relation remains a function of y?

A. (4, 7)  
B. (1, 6)  
C. (3, 3)  
D. (0, 4)

**Item Stem**

**Item Options**

**Content Standard**

**Depth of Knowledge**

**Key**

PK3 - Data Card continued

Item Name	The Function	Item ID	Year	Session	Calc	Model/Est	Grade
11	F1	77	2	Spring	2013	1	No

N	P-Val	Mean	Item Total Corr
4181	0.08		0.32

Label	Final	Final S.E.	Preliminary	Preliminary S.E.
Location	-1.08	0.03		

Label	Proportion	Corr	Avg Mean	Threshold
A	0.05	-0.14		
B	0.10	0.03		
C	0.17	-0.25		
D	0.68	0.30		
METS	0.05			
OMTS	0.05			

Category	Bias Code	Item Mean	N - Raw	N - Final
NONFUNCTIONAL	C-	1.00	2072	2072
FUNCTIONAL	A+	0.31	2571	1589
NONFUNCTIONAL	A-	-0.32	3009	262
NONFUNCTIONAL	A-	-0.38	3009	262



# Item Statistics/ Analyses

## Difficulty

- Measures how difficult/hard the item is to students.
- **"P-Val"**
  - Proportion of students who answered item correctly
  - Ranges from 0.0 to 1.0
  - Lower values—hard item
  - Higher values—easy item

## Discrimination

- Measures item's ability to differentiate between high and low achievers
- **"Item Total Corr"**
  - Ranges from -1.0 to +1.0
  - High positive—high achievers outperformed low achievers
  - High negative—low achievers outperformed high achievers

## Distractors/ Steps

- **MC: Proportions and correlations** for incorrect response options
- **Proportions** for each score point

## DIF

- **Differential Item Functioning (DIF)**
  - Statistical analysis to determine if items are potentially **unfair or inappropriate** for assessing the knowledge of various subgroups (e.g., gender, ethnicity, and test mode).

# Item Difficulty

## Difficulty

It measures how difficult/hard the item is to students. P-value ("**P-Val**") is the most frequently used statistics to indicate the item difficulty.

- **"P-Val"** for MC items
  - Proportion (e.g., 0.5) of students who answered an item correctly
- **"P-Val"** for TE items
  - Average score ("**Mean**", e.g. 2) obtained by the students
  - $P\text{-Val} = \text{Mean} / \text{Max Score Point}$

### MC Item

#### Traditional Statistics

N	P-Val	Mean	Item Total Corr
4349	0.73		0.49

### TE Item

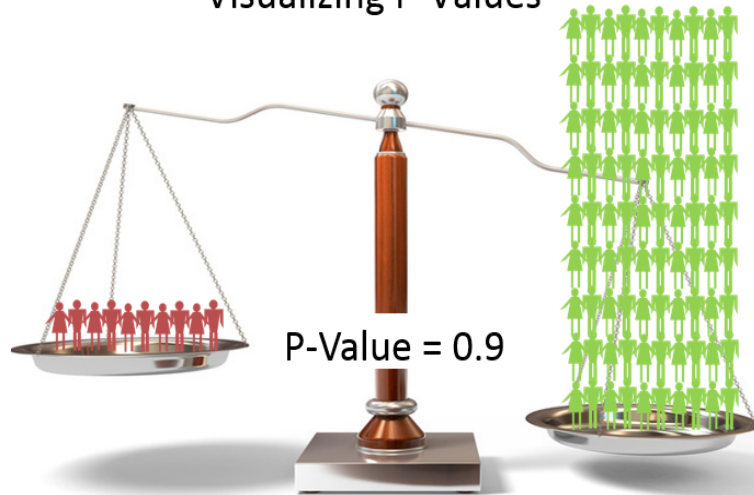
#### Traditional Statistics

N	P-Val	Mean	Item Total Corr
2008	0.25	1.01	0.65



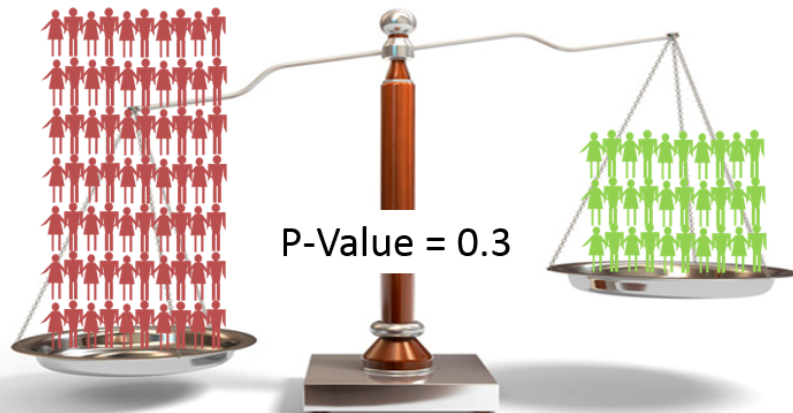
Item Difficulty

Visualizing P-Values



Item Difficulty

Visualizing P-Values



## Item Difficulty

### Range

**P-Value:** Ranges from **0 to 1**

- 0 = no students answered item correctly
- 1 = all students answered item correctly
- Lower numbers = more difficult (hard)
- Higher numbers = less difficult (easy)

### Guidelines

- Accept items in this range: **0.30 to 0.90**
  - Items below 0.30 may be too hard
  - Items above 0.90 may be too easy

### Guideline Questions

- Is this a difficult or an easy item?
- Why did most students answer this item correctly or incorrectly?
- Should I **ACCEPT** this item so there are enough items to assess the corresponding content standard and/or assess students at different performance levels (i.e., Below Basic, Basic, Proficient, and Advanced)?
- Are there any other reasons other than item difficulty to support my decision on ACCEPTING or REJECTING this item?

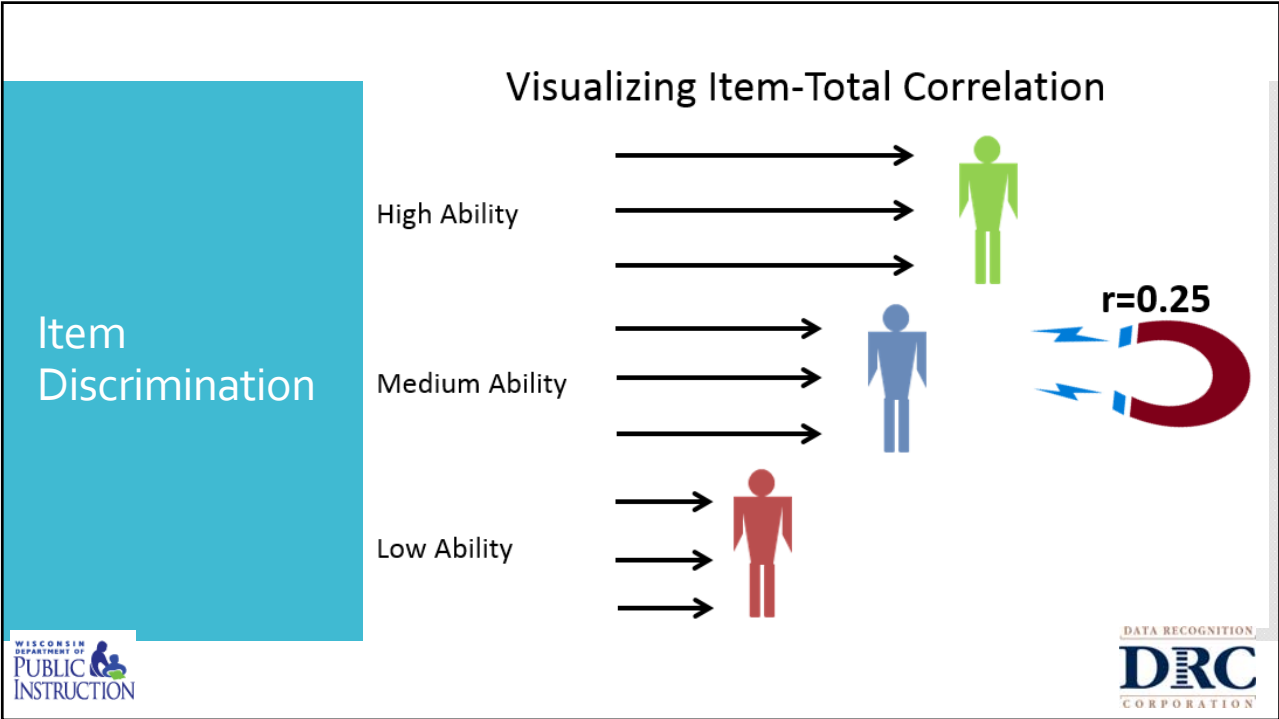
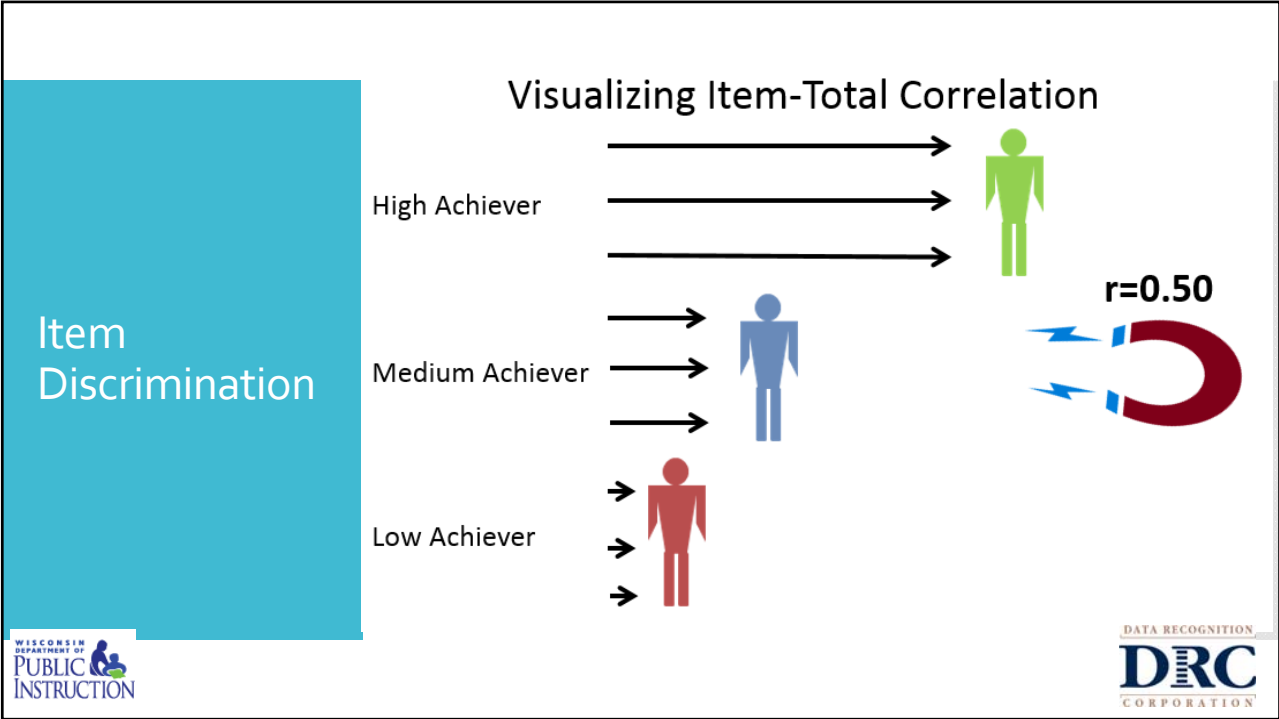
## Item Discrimination

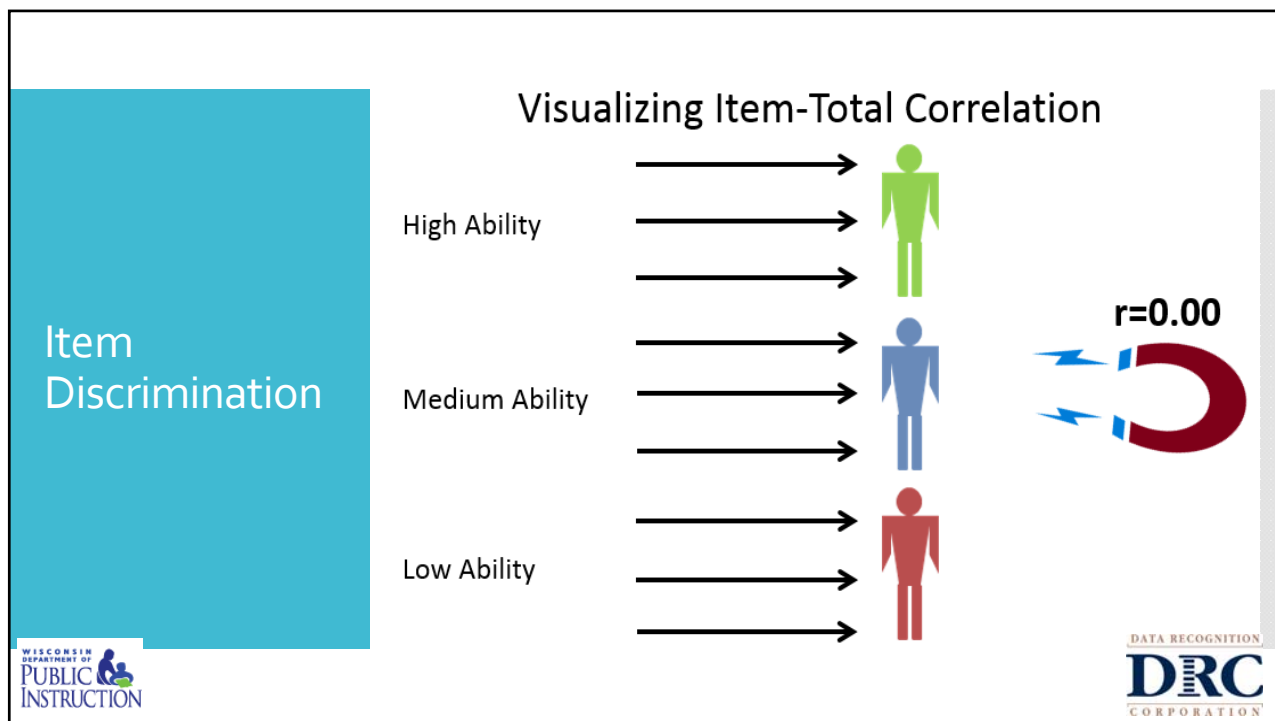
### Discrimination

- Measures item's ability to **differentiate** between **high** and **low** achievers
- **Item Total Corr** (Item Total Correlation)

### Traditional Statistics

N	P-Val	Mean	Item Total Corr
4349	0.73		0.49





### Item Discrimination: Item Total Corr

#### Range



- Ranges from -1 to +1
- 1 = "perfect" **negative** relationship
- 0 = no linear relationship
- +1 = "perfect" **positive** relationship

#### Guideline

- **at or above 0.25**
- Smaller is okay, depending on difficulty
- Items with **negative** or around **0.0** item total correlations are **very poor** items that should be rejected most of the time.

#### Guideline Questions

- Why is this item less able to differentiate high and low achievers?
- Is the low discrimination associated with extreme low or high P-Values (item difficulty)?
- Should I **ACCEPT** this item so there are enough items to assess the corresponding content standard and/or assess students at different performance levels (i.e., Below Basic, Basic, Proficient, and Advanced)?
- Are there any other reasons other than item discrimination to support my decision on **ACCEPTING** or **REJECTING** this item?

# Dif Analysis



- Procedure used to identify items that function differently for particular groups of students (e.g., gender, ethnicity, and test administration mode).
- Hypothesis is that test takers with **similar** knowledge or ability should perform in **similar** ways on a test item.

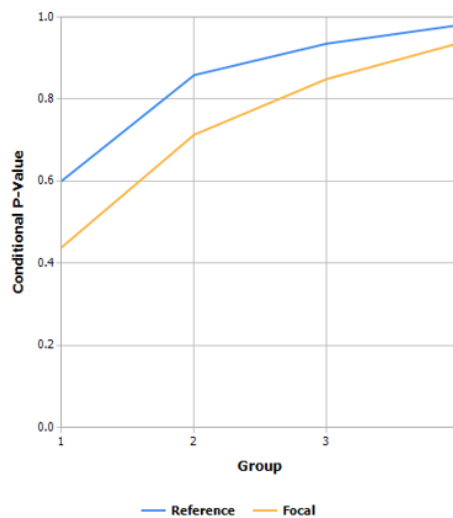
## DIF Analysis

Category	Bias Code	Num Value	N - Ref	N - Focal
MALEFEMALE	C-	1.66	2072	2079
PAPERONLINE	A+	0.31	2571	1590
WHITEBLACK	A-	-0.32	3608	262
WHITEHISPANIC	A-	-0.38	3608	200



# Dif Analysis

## Visualizing DIF



# Dif Analysis

## Procedure

- Compares “focal” vs. “reference” groups
- Reference groups: Males, Whites, and Students who took the Paper/Pencil Test
- Focal groups: Females, Blacks, Hispanics, and Students who took the computer-based Online Test

## DIF Analysis

Category	Bias Code	Num Value	N - Ref	N - Focal
MALEFEMALE	C-	1.66	2072	2079
PAPERONLINE	A+	0.31	2571	1590
WHITEBLACK	A-	-0.32	3608	262
WHITEHISPANIC	A-	-0.38	3608	200

### Letter A, B, and C:

- A-Minor
- B-Moderate
- C-Severe

### Sign:

- '-' favors Reference;
- '+' favors Focal

Reference/Focal

# Dif Analysis

## Guideline

- Only items with C+ or C- (i.e., severe) DIF require review
- Items with C+ or C- DIF may be acceptable if no potential bias causes the differential item functioning

## Guideline Question

- Is there anything in the content or format of the item that may interfere with, or advantage, one group of students over another based on:
  - Gender
  - Ethnicity
  - Mode of administration (paper/pencil or online, which will be reviewed toward the end of the data review meeting)?

## Guidelines

### Items have been flagged for review if:

- an MC item has:
  - $p$ -value  $< 0.20$  or  $p$ -value  $> 0.90$
  - item-total correlation for the correct response  $< 0.25$
  - item-total correlation for any incorrect response  $> 0.0$
  - proportion selecting any incorrect response  $> p$ -value
  - MALEFEMALE, WHITEBLACK, WHITEHISPANIC, and/or ONLINE bias code of either C- or C+
- a TE item has:
  - $p$ -value  $< 0.20$  or  $p$ -value  $> 0.90$
  - item-total correlation  $< 0.25$
  - score proportion  $< 0.05$
  - MALEFEMALE, WHITEBLACK, WHITEHISPANIC, and/or ONLINE bias code of either C- or C+

## In Content Areas

- DPI and DRC review flagged items from Spring 2016
- DPI and DRC review statistics of FT items from Spring 2016

• Questions?





## **Appendix B**

### **Spring 2016 Field Test Data Review Results**

Table B-1. Summary of the Spring 2016 Field Test Data Review Results

Content	Item Type	Number of Items in Spring 2016 FT	Flagged Items in Spring 2016 FT Examined at September 2016 Data Review		Flagged Items in Spring 2016 FT Rejected at September 2016 Data Review	
			Number of Items	% of Field Test	Number of Items	% of Field Test
English Language Arts	MC	123	37	30.1%	30	24.4%
	TE/MS	65	12	18.5%	4	6.2%
	ESR	26	9	34.6%	8	30.8%
Mathematics	MC	59	18	30.5%	9	15.3%
	TE/MS	37	10	27.0%	5	13.5%
Social Studies	MC	52	18	34.6%	4	7.7%
	TE	NA	NA	N/A	NA	N/A
Science	MC	27	16	59.3%	10	37.0%
	TE	5	2	40.0%	4	80.0%
Total		504	122		74	

## **Appendix C**

### **Spring 2017 English Language Arts Operational Test Maps**

Table C-1. English Language Arts, Grade 3 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Standard	Domain
3	1	1	MC	OP	1	3.L.4	Reading
3	1	2	MC	OP	1	3.RI.1	Reading
3	1	3	MC	OP	1	3.RI.2	Reading
3	1	4	TDA	OP	4	3.W.2	Writing
3	2	5	MC	OP	1	3.W.1c	Writing
3	2	6	Text Highlight	OP	1	3.W.3b	Writing
3	2	7	MS (Multi-select)	OP	2	3.W.1a	Writing
3	2	8	Text Highlight	OP	1	3.W.2c	Writing
3	2	10	MC	OP	1	3.W.8	Writing
3	2	11	TE	OP	2	3.W.8	Writing
3	2	12	MC	OP	1	3.W.8	Writing
3	2	13	MC	OP	1	3.W.8	Writing
3	2	14	MC	OP	1	3.L.1g	Writing
3	2	15	MC	OP	1	3.L.2d	Writing
3	2	16	Text Highlight	OP	2	3.L.2a	Writing
3	3	17	ESR	OP	2	3.SL.3	Listening
3	3	18	MC	OP	1	3.SL.3	Listening
3	3	22	MC	OP	1	3.SL.3	Listening
3	3	23	ESR	OP	2	3.SL.2	Listening
3	3	24	MC	OP	1	3.SL.3	Listening
3	4	25	MC	OP	1	3.L.4	Reading
3	4	26	ESR	OP	2	3.RI.6	Reading
3	4	27	MS (Multi-select)	OP	2	3.RI.8	Reading
3	4	28	MC	OP	1	3.RI.3	Reading
3	4	29	MC	OP	1	3.L.5	Reading
3	4	30	ESR	OP	2	3.RL.1	Reading
3	4	31	MC	OP	1	3.RL.3	Reading
3	4	32	MC	OP	1	3.RL.2	Reading
3	4	33	MC	OP	1	3.RL.6	Reading
3	4	38	Drag and Drop	OP	2	3.RL.1	Reading
3	4	39	MC	OP	1	3.RL.5	Reading
3	4	40	MC	OP	1	3.RL.3	Reading
3	4	41	MC	OP	1	3.RL.9	Reading

Table C-2. English Language Arts, Grade 4 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Standard	Domain
4	1	1	MC	OP	1	4.L.4	Reading
4	1	2	MC	OP	1	4.RL.3	Reading
4	1	3	Drag and Drop	OP	2	4.RL.3	Reading
4	1	4	Text Highlight	OP	1	4.RL.1	Reading
4	1	5	MC	OP	1	4.RL.2	Reading
4	1	6	TDA	OP	4	4.W.9	Writing
4	2	7	Text Highlight	OP	1	4.W.2e	Writing
4	2	8	Text Highlight	OP	1	4.W.3e	Writing
4	2	9	MC	OP	1	4.W.2d	Writing
4	2	10	MC	OP	1	4.W.1d	Writing
4	2	12	MC	OP	1	4.W.3a	Writing
4	2	13	MC	OP	1	4.W.8	Writing
4	2	14	MC	OP	1	4.W.8	Writing
4	2	15	MS (Multi-select)	OP	2	4.W.8	Writing
4	2	16	MC	OP	1	4.W.8	Writing
4	2	17	MC	OP	1	4.L.2b	Writing
4	2	18	MC	OP	1	4.L.2a	Writing
4	2	19	Text Highlight	OP	2	4.L.1b	Writing
4	2	20	Drop Down Menu	OP	2	4.L.3b	Writing
4	3	21	TE	OP	2	4.SL.3	Listening
4	3	22	MC	OP	1	4.SL.3	Listening
4	3	23	MC	OP	1	4.SL.2	Listening
4	3	27	ESR	OP	2	4.SL.2	Listening
4	3	28	MC	OP	1	4.SL.2	Listening
4	3	29	MC	OP	1	4.SL.3	Listening
4	4	30	MC	OP	1	4.RI.5	Reading
4	4	31	MC	OP	1	4.L.4	Reading
4	4	32	MC	OP	1	4.RI.5	Reading
4	4	33	MC	OP	1	4.L.5	Reading
4	4	34	Text Highlight	OP	1	4.L.4	Reading
4	4	35	MC	OP	1	4.RL.2	Reading
4	4	36	MS (Multi-select)	OP	2	4.RL.3	Reading
4	4	37	MC	OP	1	4.RL.6	Reading
4	4	42	MC	OP	1	4.L.4	Reading
4	4	43	MC	OP	1	4.RI.2	Reading
4	4	44	ESR	OP	2	4.RI.8	Reading
4	4	45	MC	OP	1	4.RI.5	Reading

Table C-3. English Language Arts, Grade 5 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Standard	Domain
5	1	1	MC	OP	1	5.L.4	Reading
5	1	2	MC	OP	1	5.RI.9	Reading
5	1	3	MC	OP	1	5.RI.6	Reading
5	1	4	TDA	OP	4	5.W.9	Writing
5	2	5	MC	OP	1	5.W.1c	Writing
5	2	6	MC	OP	1	5.W.2b	Writing
5	2	7	MC	OP	1	5.W.2d	Writing
5	2	8	MC	OP	1	5.W.1d	Writing
5	2	10	MC	OP	1	5.W.3b	Writing
5	2	11	MC	OP	1	5.W.3e	Writing
5	2	12	MC	OP	1	5.W.5	Writing
5	2	13	MS (Multi-select)	OP	2	5.W.8	Writing
5	2	14	MC	OP	1	5.W.8	Writing
5	2	15	MC	OP	1	5.L.2	Writing
5	2	16	MC	OP	1	5.L.2b	Writing
5	2	17	MC	OP	1	5.L.3a	Writing
5	2	18	MC	OP	1	5.W.5	Writing
5	2	19	Drag and Drop	OP	2	5.L.1b	Writing
5	3	20	MS (Multi-select)	OP	2	5.SL.3	Listening
5	3	21	MC	OP	1	5.SL.2	Listening
5	3	22	MC	OP	1	5.SL.3	Listening
5	3	26	MC	OP	1	5.SL.3	Listening
5	3	27	MC	OP	1	5.SL.2	Listening
5	3	28	ESR	OP	2	5.SL.2	Listening
5	4	29	MS (Multi-select)	OP	2	5.RI.1	Reading
5	4	30	MC	OP	1	5.RI.8	Reading
5	4	31	MC	OP	1	5.RI.1	Reading
5	4	32	MC	OP	1	5.L.4	Reading
5	4	33	ESR	OP	2	5.RL.1	Reading
5	4	34	MC	OP	1	5.RL.9	Reading
5	4	35	MC	OP	1	5.RL.9	Reading
5	4	36	ESR	OP	2	5.RL.2	Reading
5	4	41	MC	OP	1	5.RL.5	Reading
5	4	42	MC	OP	1	5.RL.1	Reading
5	4	43	MS (Multi-select)	OP	2	5.RL.1	Reading
5	4	44	MC	OP	1	5.RL.6	Reading
5	4	45	MC	OP	1	5.RL.5	Reading

Table C-4. English Language Arts, Grade 6 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Standard	Domain
6	1	1	MS (Multi-select)	OP	2	6.RL.3	Reading
6	1	2	MC	OP	1	6.RL.1	Reading
6	1	3	MC	OP	1	6.RL.5	Reading
6	1	4	TDA	OP	4	6.W.9	Writing
6	2	5	MC	OP	1	6.W.1c	Writing
6	2	6	MC	OP	1	6.W.2d	Writing
6	2	7	MC	OP	1	6.W.2e	Writing
6	2	8	MC	OP	1	6.W.3b	Writing
6	2	10	Text Highlight	OP	1	6.L.3a	Writing
6	2	11	MS (Multi-select)	OP	2	6.L.1d	Writing
6	2	12	Text Highlight	OP	1	6.L.3b	Writing
6	2	13	Text Input	OP	2	6.L.2b	Writing
6	2	14	MC	OP	1	6.L.2a	Writing
6	2	15	Text Highlight	OP	2	6.W.8	Writing
6	2	16	Text Highlight	OP	2	6.W.8	Writing
6	2	17	MC	OP	1	6.W.8	Writing
6	3	18	MC	OP	1	6.SL.2	Listening
6	3	19	MC	OP	1	6.SL.3	Listening
6	3	20	MS (Multi-select)	OP	2	6.SL.2	Listening
6	3	24	ESR	OP	2	6.SL.3	Listening
6	3	25	MC	OP	1	6.SL.3	Listening
6	3	26	MC	OP	1	6.SL.2	Listening
6	4	27	Text Highlight	OP	2	6.RI.4	Reading
6	4	28	MC	OP	1	6.RI.8	Reading
6	4	29	MC	OP	1	6.RI.3	Reading
6	4	30	MC	OP	1	6.RI.9	Reading
6	4	31	MC	OP	1	6.L.5	Reading
6	4	32	MC	OP	1	6.RL.4	Reading
6	4	33	MC	OP	1	6.RL.6	Reading
6	4	34	Drag and Drop	OP	2	6.RL.2	Reading
6	4	39	MC	OP	1	6.RI.1	Reading
6	4	40	MC	OP	1	6.RI.4	Reading
6	4	41	MS (Multi-select)	OP	2	6.RI.6	Reading
6	4	42	Drag and Drop	OP	1	6.RI.3	Reading
6	4	43	MC	OP	1	6.RI.5	Reading

Table C-5. English Language Arts, Grade 7 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Standard	Domain
7	1	1	MC	OP	1	7.RL.1	Reading
7	1	2	MC	OP	1	7.RL.3	Reading
7	1	3	MC	OP	1	7.RL.2	Reading
7	1	4	TDA	OP	4	7.W.9	Writing
7	2	5	Text Highlight	OP	1	7.W.3e	Writing
7	2	6	MC	OP	1	7.W.2e	Writing
7	2	7	MC	OP	1	7.W.1b	Writing
7	2	8	MC	OP	1	7.W.1e	Writing
7	2	10	ESR	OP	2	7.W.2b	Writing
7	2	11	Text Highlight	OP	2	7.L.3a	Writing
7	2	12	Text Highlight	OP	2	7.L.3a	Writing
7	2	13	MC	OP	1	7.L.2	Writing
7	2	14	MC	OP	1	7.L.1b	Writing
7	2	15	MC	OP	1	7.W.8	Writing
7	2	16	MC	OP	1	7.W.8	Writing
7	2	17	ESR	OP	2	7.W.8	Writing
7	3	18	MC	OP	1	7.SL.2	Listening
7	3	19	ESR	OP	2	7.SL.3	Listening
7	3	23	MC	OP	1	7.SL.3	Listening
7	3	24	ESR	OP	2	7.SL.2	Listening
7	3	25	MS (Multi-select)	OP	2	7.SL.3	Listening
7	4	26	MC	OP	1	7.RI.5	Reading
7	4	27	MC	OP	1	7.L.4	Reading
7	4	28	MC	OP	1	7.L.4	Reading
7	4	29	ESR	OP	2	7.RI.6	Reading
7	4	30	MC	OP	1	7.RI.6	Reading
7	4	31	MC	OP	1	7.RI.1	Reading
7	4	32	MC	OP	1	7.RI.2	Reading
7	4	33	Text Highlight	OP	1	7.RI.4	Reading
7	4	34	MC	OP	1	7.RI.6	Reading
7	4	35	MC	OP	1	7.RI.8	Reading
7	4	40	ESR	OP	2	7.RL.1	Reading
7	4	41	MC	OP	1	7.RL.3	Reading
7	4	42	MC	OP	1	7.RL.4	Reading
7	4	43	MC	OP	1	7.RL.3	Reading
7	4	44	MC	OP	1	7.RL.2	Reading



Table C-6. English Language Arts, Grade 8 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Standard	Domain
8	1	1	Text Highlight	OP	1	8.RI.1	Reading
8	1	2	ESR	OP	2	8.RI.2	Reading
8	1	3	MC	OP	1	8.RI.3	Reading
8	1	4	MC	OP	1	8.RI.8	Reading
8	1	5	TDA	OP	4	8.W.9	Writing
8	2	6	MC	OP	1	8.W.3c	Writing
8	2	7	MC	OP	1	8.W.3d	Writing
8	2	8	MC	OP	1	8.W.1e	Writing
8	2	9	MC	OP	1	8.W.2b	Writing
8	2	11	MC	OP	1	8.W.1c	Writing
8	2	12	Text Input	OP	2	8.L.2c	Writing
8	2	13	MC	OP	1	8.L.2a	Writing
8	2	14	MC	OP	1	8.L.2b	Writing
8	2	15	Text Highlight	OP	2	8.W.2b	Writing
8	2	16	MC	OP	1	8.W.8	Writing
8	2	17	Text Highlight	OP	1	8.W.8	Writing
8	2	18	MC	OP	1	8.W.8	Writing
8	2	19	MS (Multi-select)	OP	2	8.W.8	Writing
8	3	20	MC	OP	1	8.SL.3	Listening
8	3	21	MC	OP	1	8.SL.3	Listening
8	3	22	ESR	OP	2	8.SL.2	Listening
8	3	26	MC	OP	1	8.SL.3	Listening
8	3	27	MC	OP	1	8.SL.2	Listening
8	3	28	ESR	OP	2	8.SL.2	Listening
8	4	29	MC	OP	1	8.RL.4	Reading
8	4	30	MC	OP	1	8.RL.1	Reading
8	4	31	MC	OP	1	8.RL.3	Reading
8	4	32	MC	OP	1	8.RL.6	Reading
8	4	33	MC	OP	1	8.RI.3	Reading
8	4	34	Text Highlight	OP	1	8.L.5	Reading
8	4	35	MS (Multi-select)	OP	2	8.RI.8	Reading
8	4	36	Drag and Drop	OP	2	8.RI.9	Reading
8	4	41	MC	OP	1	8.L.4	Reading
8	4	42	MS (Multiselect)	OP	2	8.RL.2	Reading
8	4	43	MC	OP	1	8.RL.3	Reading
8	4	44	MC	OP	1	8.RL.6	Reading

## **Appendix D**

### **Spring 2017 Mathematics Operational Test Maps**

Table D-1. Mathematics, Grade 3 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Standard	Domain
3	1	2	MC	OP	1	3.NBT.1	Number and Operations in Base Ten
3	1	3	MC	OP	1	3.NF.1	Number and Operations–Fractions
3	1	4	MC	OP	1	3.OA.1	Operations and Algebraic Thinking
3	1	6	MC	OP	1	3.MD.1	Measurement and Data
3	1	7	MC	OP	1	3.G.1	Geometry
3	1	8	Hot Spot	OP	1	3.NBT.1	Number and Operations in Base Ten
3	1	9	MC	OP	1	3.OA.4	Operations and Algebraic Thinking
3	1	11	Text Input	OP	1	3.G.1	Geometry
3	1	12	MC	OP	1	3.NF.2	Number and Operations–Fractions
3	1	13	Text Input	OP	1	3.MD.3	Measurement and Data
3	1	14	MC	OP	1	3.OA.6	Operations and Algebraic Thinking
3	1	16	Text Input	OP	1	3.MD.1	Measurement and Data
3	1	17	MC	OP	1	3.NBT.2	Number and Operations in Base Ten
3	1	18	MC	OP	1	3.MD.5	Measurement and Data
3	1	19	MC	OP	1	3.NF.3	Number and Operations–Fractions
3	1	20	MC	OP	1	3.MD.7	Measurement and Data
3	1	21	MC	OP	1	3.OA.8	Operations and Algebraic Thinking
3	1	22	Text Input	OP	1	3.NF.3	Number and Operations–Fractions
3	1	23	MC	OP	1	3.NBT.3	Number and Operations in Base Ten
3	1	24	Matching	OP	1	3.NF.3	Number and Operations–Fractions
3	1	25	MC	OP	1	3.G.2	Geometry
3	2	27	MC	OP	1	3.MD.8	Measurement and Data
3	2	28	MC	OP	1	3.NBT.1	Number and Operations in Base Ten
3	2	29	MC	OP	1	3.OA.2	Operations and Algebraic Thinking
3	2	31	MC	OP	1	3.G.1	Geometry
3	2	32	Text Input	OP	1	3.OA.3	Operations and Algebraic Thinking
3	2	33	Drag & Drop	OP	1	3.G.1	Geometry

Table D-1. Mathematics, Grade 3 Test Map (cont.)

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Standard	Domain
3	2	34	MC	OP	1	3.NF.2	Number and Operations– Fractions
3	2	36	MC	OP	1	3.NBT.2	Number and Operations in Base Ten
3	2	37	MC	OP	1	3.MD.2	Measurement and Data
3	2	38	MC	OP	1	3.G.2	Geometry
3	2	39	MC	OP	1	3.OA.5	Operations and Algebraic Thinking
3	2	41	Text Input	OP	1	3.NF.1	Number and Operations– Fractions
3	2	42	Text Input	OP	1	3.G.2	Geometry
3	2	43	MC	OP	1	3.NF.2	Number and Operations– Fractions
3	2	44	Text Input	OP	1	3.MD.4	Measurement and Data
3	2	45	MC	OP	1	3.OA.7	Operations and Algebraic Thinking
3	2	46	Text Input	OP	1	3.NBT.3	Number and Operations in Base Ten
3	2	47	MC	OP	1	3.NBT.2	Number and Operations in Base Ten
3	2	48	MC	OP	1	3.MD.6	Measurement and Data
3	2	49	MC	OP	1	3.MD.3	Measurement and Data
3	2	50	MC	OP	1	3.OA.9	Operations and Algebraic Thinking

Table D-2. Mathematics, Grade 4 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Standard	Domain
4	1	2	MC	OP	1	4.NBT.1	Number and Operations in Base Ten
4	1	3	MC	OP	1	4.OA.1	Operations and Algebraic Thinking
4	1	4	MC	OP	1	4.G.1	Geometry
4	1	6	MC	OP	1	4.NF.1	Number and Operations–Fractions
4	1	7	MC	OP	1	4.NBT.3	Number and Operations in Base Ten
4	1	8	Text Input	OP	1	4.MD.3	Measurement and Data
4	1	9	MC	OP	1	4.OA.2	Operations and Algebraic Thinking
4	1	11	MC	OP	1	4.MD.2	Measurement and Data
4	1	12	MC	OP	1	4.OA.3	Operations and Algebraic Thinking
4	1	13	MC	OP	1	4.G.2	Geometry
4	1	14	MC	OP	1	4.OA.4	Operations and Algebraic Thinking
4	1	16	Drag & Drop	OP	1	4.NBT.1	Number and Operations in Base Ten
4	1	17	MC	OP	1	4.NF.3	Number and Operations–Fractions
4	1	18	MC	OP	1	4.MD.4	Measurement and Data
4	1	19	MC	OP	1	4.NBT.5	Number and Operations in Base Ten
4	1	20	Text Input	OP	1	4.OA.4	Operations and Algebraic Thinking
4	1	21	Text Input	OP	1	4.MD.5	Measurement and Data
4	1	22	MC	OP	1	4.NF.4	Number and Operations–Fractions
4	1	23	MC	OP	1	4.OA.5	Operations and Algebraic Thinking
4	1	24	Number Line	OP	1	4.NF.6	Number and Operations–Fractions
4	1	25	MC	OP	1	4.MD.6	Measurement and Data
4	1	26	MC	OP	1	4.G.3	Geometry
4	1	27	MC	OP	1	4.NF.7	Number and Operations–Fractions
4	2	29	MC	OP	1	4.NBT.2	Number and Operations in Base Ten
4	2	30	MC	OP	1	4.OA.1	Operations and Algebraic Thinking
4	2	31	MC	OP	1	4.G.1	Geometry
4	2	33	MC	OP	1	4.MD.2	Measurement and Data
4	2	34	MC	OP	1	4.NF.1	Number and Operations–Fractions

Table D-2. Mathematics, Grade 4 Test Map (cont.)

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Standard	Domain
4	2	35	Drag & Drop	OP	1	4.NBT.4	Number and Operations in Base Ten
4	2	36	Text Input	OP	1	4.NBT.2	Number and Operations in Base Ten
4	2	38	MC	OP	1	4.MD.3	Measurement and Data
4	2	39	MC	OP	1	4.OA.3	Operations and Algebraic Thinking
4	2	40	MC	OP	1	4.NF.2	Number and Operations–Fractions
4	2	41	MC	OP	1	4.G.2	Geometry
4	2	43	MC	OP	1	4.MD.7	Measurement and Data
4	2	44	MC	OP	1	4.NF.3	Number and Operations–Fractions
4	2	45	Text Input	OP	1	4.NBT.5	Number and Operations in Base Ten
4	2	46	MC	OP	1	4.OA.5	Operations and Algebraic Thinking
4	2	47	MC	OP	1	4.MD.5	Measurement and Data
4	2	48	MC	OP	1	4.G.2	Geometry
4	2	49	MC	OP	1	4.NF.5	Number and Operations–Fractions
4	2	50	Text Input	OP	1	4.NBT.6	Number and Operations in Base Ten
4	2	51	MC	OP	1	4.OA.5	Operations and Algebraic Thinking
4	2	52	Text Input	OP	1	4.NF.6	Number and Operations–Fractions
4	2	53	MC	OP	1	4.MD.7	Measurement and Data
4	2	54	MC	OP	1	4.G.2	Geometry

Table D-3. Mathematics, Grade 5 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Standard	Domain
5	1	2	MC	OP	1	5.NBT.1	Number and Operations in Base Ten
5	1	3	MC	OP	1	5.NF.1	Number and Operations–Fractions
5	1	4	MC	OP	1	5.MD.1	Measurement and Data
5	1	6	MC	OP	1	5.G.1	Geometry
5	1	7	MC	OP	1	5.NBT.4	Number and Operations in Base Ten
5	1	8	Drag & Drop	OP	1	5.NF.2	Number and Operations–Fractions
5	1	9	Text Input	OP	1	5.OA.1	Operations and Algebraic Thinking
5	1	11	MC	OP	1	5.MD.2	Measurement and Data
5	1	12	MC	OP	1	5.NBT.5	Number and Operations in Base Ten
5	1	13	MC	OP	1	5.OA.2	Operations and Algebraic Thinking
5	1	14	MC	OP	1	5.NF.3	Number and Operations–Fractions
5	1	16	Coordinate Grid	OP	1	5.OA.3	Operations and Algebraic Thinking
5	1	17	Text Input	OP	1	5.MD.5	Measurement and Data
5	1	18	Text Input	OP	1	5.NBT.6	Number and Operations in Base Ten
5	1	19	MC	OP	1	5.G.2	Geometry
5	1	20	Text Input	OP	1	5.OA.2	Operations and Algebraic Thinking
5	1	21	MC	OP	1	5.NF.6	Number and Operations–Fractions
5	1	22	MC	OP	1	5.G.1	Geometry
5	1	23	MC	OP	1	5.OA.1	Operations and Algebraic Thinking
5	1	24	Drag & Drop	OP	1	5.NBT.5	Number and Operations in Base Ten
5	1	25	MC	OP	1	5.MD.3	Measurement and Data
5	1	26	Text Input	OP	1	5.NF.7	Number and Operations–Fractions
5	1	27	MC	OP	1	5.G.2	Geometry
5	2	29	MC	OP	1	5.NBT.2	Number and Operations in Base Ten
5	2	30	MC	OP	1	5.MD.1	Measurement and Data
5	2	31	MS (Multi-select)	OP	1	5.G.1	Geometry
5	2	33	Text Input	OP	1	5.NBT.3	Number and Operations in Base Ten
5	2	34	Text Input	OP	1	5.MD.1	Measurement and Data

Table D-3. Mathematics, Grade 5 Test Map (cont.)

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Standard	Domain
5	2	35	Drag & Drop	OP	1	5.OA.1	Operations and Algebraic Thinking
5	2	36	Text Input	OP	1	5.G.2	Geometry
5	2	38	MC	OP	1	5.NF.3	Number and Operations–Fractions
5	2	39	MC	OP	1	5.OA.2	Operations and Algebraic Thinking
5	2	40	MC	OP	1	5.NF.4	Number and Operations–Fractions
5	2	41	Text Input	OP	1	5.G.1	Geometry
5	2	43	Line Plot	OP	1	5.MD.2	Measurement and Data
5	2	44	MC	OP	1	5.NF.5	Number and Operations–Fractions
5	2	45	MS (Multi-select)	OP	1	5.G.4	Geometry
5	2	46	MC	OP	1	5.OA.3	Operations and Algebraic Thinking
5	2	47	MC	OP	1	5.NF.6	Number and Operations–Fractions
5	2	48	MC	OP	1	5.NBT.7	Number and Operations in Base Ten
5	2	49	MC	OP	1	5.MD.5	Measurement and Data
5	2	50	MC	OP	1	5.MD.3	Measurement and Data
5	2	51	MC	OP	1	5.OA.2	Operations and Algebraic Thinking
5	2	52	MC	OP	1	5.MD.4	Measurement and Data
5	2	53	MC	OP	1	5.NBT.7	Number and Operations in Base Ten
5	2	54	MS (Multi-select)	OP	1	5.G.3	Geometry



Table D-4. Mathematics, Grade 6 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Standard	Domain
6	1	2	MC	OP	1	6.EE.1	Expressions and Equations
6	1	3	MS (Multi-select)	OP	1	6.RP.1	Ratios and Proportional Relationships
6	1	4	MC	OP	1	6.NS.2	The Number System
6	1	6	Coordinate Grid	OP	1	6.RP.3	Ratios and Proportional Relationships
6	1	7	MC	OP	1	6.RP.3	Ratios and Proportional Relationships
6	1	8	Text Input	OP	1	6.NS.1	The Number System
6	1	9	MC	OP	1	6.RP.2	Ratios and Proportional Relationships
6	1	11	MC	OP	1	6.NS.3	The Number System
6	1	12	Matching	OP	1	6.EE.4	Expressions and Equations
6	1	13	MC	OP	1	6.RP.3	Ratios and Proportional Relationships
6	1	14	Text Input	OP	1	6.NS.4	The Number System
6	1	16	MC	OP	1	6.EE.2	Expressions and Equations
6	1	17	Text Input	OP	1	6.RP.1	Ratios and Proportional Relationships
6	1	18	MC	OP	1	6.RP.2	Ratios and Proportional Relationships
6	1	19	MC	OP	1	6.EE.2	Expressions and Equations
6	1	20	MC	OP	1	6.NS.3	The Number System
6	2	22	MC	OP	1	6.SP.1	Statistics and Probability
6	2	23	MC	OP	1	6.NS.5	The Number System
6	2	24	Text Input	OP	1	6.EE.7	Expressions and Equations
6	2	26	MC	OP	1	6.G.1	Geometry
6	2	27	MC	OP	1	6.SP.3	Statistics and Probability
6	2	28	MC	OP	1	6.EE.7	Expressions and Equations
6	2	29	MC	OP	1	6.NS.6	The Number System
6	2	31	Text Input	OP	1	6.G.1	Geometry
6	2	32	MC	OP	1	6.SP.4	Statistics and Probability
6	2	33	MC	OP	1	6.G.3	Geometry
6	2	34	Drag & Drop	OP	1	6.NS.7	The Number System
6	2	36	MC	OP	1	6.EE.8	Expressions and Equations
6	2	37	MS (Multi-select)	OP	1	6.SP.5	Statistics and Probability
6	2	38	MS (Multi-select)	OP	1	6.G.4	Geometry
6	2	39	MC	OP	1	6.EE.5	Expressions and Equations

Table D-4. Mathematics, Grade 6 Test Map (cont.)

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Standard	Domain
6	2	40	MC	OP	1	6.NS.8	The Number System
6	2	41	Text Input	OP	1	6.G.2	Geometry
6	2	42	MC	OP	1	6.SP.4	Statistics and Probability
6	2	43	MC	OP	1	6.EE.6	Expressions and Equations
6	2	44	MC	OP	1	6.SP.5	Statistics and Probability
6	2	45	Drag & Drop	OP	1	6.NS.6	The Number System
6	2	46	MC	OP	1	6.EE.5	Expressions and Equations
6	2	47	MC	OP	1	6.SP.2	Statistics and Probability
6	2	48	MC	OP	1	6.NS.8	The Number System
6	2	49	MC	OP	1	6.SP.3	Statistics and Probability
6	2	50	Text Input	OP	1	6.EE.9	Expressions and Equations
6	2	51	MC	OP	1	6.G.4	Geometry
6	2	52	MC	OP	1	6.SP.4	Statistics and Probability
6	2	53	MC	OP	1	6.G.2	Geometry
6	2	54	MC	OP	1	6.SP.5	Statistics and Probability

Table D-5. Mathematics, Grade 7 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Standard	Domain
7	1	2	MC	OP	1	7.EE.2	Expressions and Equations
7	1	3	MC	OP	1	7.NS.3	The Number System
7	1	4	Text Input	OP	1	7.NS.2	The Number System
7	1	6	Drag & Drop	OP	1	7.NS.2	The Number System
7	1	7	MC	OP	1	7.EE.1	Expressions and Equations
7	1	8	MC	OP	1	7.NS.2	The Number System
7	1	9	Text Input	OP	1	7.NS.1	The Number System
7	1	10	MC	OP	1	7.NS.1	The Number System
7	1	11	MC	OP	1	7.EE.2	Expressions and Equations
7	1	12	MC	OP	1	7.NS.3	The Number System
7	1	13	MC	OP	1	7.EE.1	Expressions and Equations
7	2	15	MC	OP	1	7.EE.3	Expressions and Equations
7	2	16	MC	OP	1	7.RP.2	Ratios and Proportional Relationships
7	2	17	Text Input	OP	1	7.G.5	Geometry
7	2	19	Drag & Drop	OP	1	7.EE.4	Expressions and Equations
7	2	20	MC	OP	1	7.EE.4	Expressions and Equations
7	2	21	MC	OP	1	7.G.1	Geometry
7	2	22	MC	OP	1	7.RP.3	Ratios and Proportional Relationships
7	2	24	Text Input	OP	1	7.G.5	Geometry
7	2	25	Coordinate Grid	OP	1	7.G.2	Geometry
7	2	26	MS (Multi-select)	OP	1	7.SP.8	Statistics and Probability
7	2	27	MC	OP	1	7.G.4	Geometry
7	2	29	MC	OP	1	7.SP.1	Statistics and Probability
7	2	30	MC	OP	1	7.RP.2	Ratios and Proportional Relationships
7	2	31	MC	OP	1	7.SP.5	Statistics and Probability
7	2	32	MC	OP	1	7.RP.3	Ratios and Proportional Relationships
7	2	34	Hot Spot	OP	1	7.G.6	Geometry
7	2	35	MC	OP	1	7.SP.1	Statistics and Probability
7	2	36	MS (Multi-select)	OP	1	7.EE.3	Expressions and Equations
7	2	37	Text Input	OP	1	7.SP.6	Statistics and Probability
7	2	39	MC	OP	1	7.RP.1	Ratios and Proportional Relationships
7	2	40	MC	OP	1	7.G.3	Geometry

Table D-5. Mathematics, Grade 7 Test Map (cont.)

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Standard	Domain
7	2	41	MC	OP	1	7.SP.2	Statistics and Probability
7	2	42	MC	OP	1	7.RP.3	Ratios and Proportional Relationships
7	2	43	Text Input	OP	1	7.EE.4	Expressions and Equations
7	2	44	MC	OP	1	7.G.6	Geometry
7	2	45	Drag & Drop	OP	1	7.SP.7	Statistics and Probability
7	2	46	MC	OP	1	7.SP.3	Statistics and Probability
7	2	47	MC	OP	1	7.G.4	Geometry
7	2	48	Coordinate Grid	OP	1	7.SP.6	Statistics and Probability
7	2	49	Text Input	OP	1	7.RP.2	Ratios and Proportional Relationships
7	2	50	MC	OP	1	7.EE.3	Expressions and Equations
7	2	51	MC	OP	1	7.SP.2	Statistics and Probability
7	2	52	Text Input	OP	1	7.RP.1	Ratios and Proportional Relationships
7	2	53	MC	OP	1	7.G.6	Geometry
7	2	54	MC	OP	1	7.SP.3	Statistics and Probability

Table D-6. Mathematics, Grade 8 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Standard	Domain
8	1	2	MC	OP	1	8.EE.1	Expressions and Equations
8	1	3	Text Input	OP	1	8.NS.2	The Number System
8	1	4	MC	OP	1	8.EE.2	Expressions and Equations
8	1	6	MC	OP	1	8.NS.1	The Number System
8	1	7	MC	OP	1	8.NS.2	The Number System
8	1	8	Text Input	OP	1	8.EE.4	Expressions and Equations
8	1	9	MC	OP	1	8.NS.1	The Number System
8	1	10	MC	OP	1	8.EE.2	Expressions and Equations
8	1	11	MC	OP	1	8.NS.1	The Number System
8	1	12	MC	OP	1	8.NS.1	The Number System
8	1	13	Text Input	OP	1	8.NS.1	The Number System
8	1	14	MC	OP	1	8.EE.3	Expressions and Equations
8	1	15	Text Input	OP	1	8.NS.2	The Number System
8	2	17	MC	OP	1	8.G.1	Geometry
8	2	18	MC	OP	1	8.EE.5	Expressions and Equations
8	2	19	Drag & Drop	OP	1	8.G.5	Geometry
8	2	21	MC	OP	1	8.SP.1	Statistics and Probability
8	2	22	Text Input	OP	1	8.G.9	Geometry
8	2	23	Text Input	OP	1	8.F.2	Functions
8	2	24	MC	OP	1	8.G.3	Geometry
8	2	26	Drag & Drop	OP	1	8.SP.4	Statistics and Probability
8	2	27	MC	OP	1	8.SP.2	Statistics and Probability
8	2	28	MC	OP	1	8.F.5	Functions
8	2	29	MC	OP	1	8.SP.3	Statistics and Probability
8	2	31	MC	OP	1	8.G.8	Geometry
8	2	32	Text Input	OP	1	8.F.4	Functions
8	2	33	Coordinate Grid	OP	1	8.EE.5	Expressions and Equations
8	2	34	MC	OP	1	8.F.4	Functions
8	2	36	MC	OP	1	8.F.2	Functions
8	2	37	MS (Multi-select)	OP	1	8.F.5	Functions
8	2	38	Text Input	OP	1	8.EE.8	Expressions and Equations
8	2	39	Hot Spot	OP	1	8.F.3	Functions
8	2	41	MC	OP	1	8.SP.1	Statistics and Probability
8	2	42	MC	OP	1	8.G.5	Geometry

Table D-6. Mathematics, Grade 8 Test Map (cont.)

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Standard	Domain
8	2	43	MC	OP	1	8.F.2	Functions
8	2	44	MC	OP	1	8.EE.6	Expressions and Equations
8	2	45	Drag & Drop	OP	1	8.G.2	Geometry
8	2	46	MC	OP	1	8.F.5	Functions
8	2	47	MC	OP	1	8.G.6	Geometry
8	2	48	MS (Multi-select)	OP	1	8.G.2	Geometry
8	2	49	MC	OP	1	8.F.1	Functions
8	2	50	Text Input	OP	1	8.G.3	Geometry
8	2	51	MC	OP	1	8.SP.3	Statistics and Probability
8	2	52	Text Input	OP	1	8.EE.7	Expressions and Equations
8	2	53	MC	OP	1	8.SP.4	Statistics and Probability
8	2	54	MC	OP	1	8.SP.3	Statistics and Probability

## **Appendix E**

### **Spring 2017 Science Operational Test Maps**

Table E-1. Science, Grade 4 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Standard	Domain
4	1	1	MC	OP	1	B.4.3	Science Connections & Nature of Science
4	1	2	MC	OP	1	C.4.4	Science Inquiry
4	1	3	MC	OP	1	D.4.4	Physical Science
4	1	4	MC	OP	1	F.4.1	Life and Environmental Science
4	1	5	MC	OP	1	A.4.2	Science Connections & Nature of Science
4	1	6	MC	OP	1	H.4.4	Science Applications & Science in Social and Personal Perspectives
4	1	7	MC	OP	1	C.4.2	Science Inquiry
4	1	8	MC	OP	1	E.4.8	Earth and Space Science
4	1	9	MC	OP	1	C.4.5	Science Inquiry
4	1	10	MC	OP	1	C.4.6	Science Inquiry
4	1	11	TE	OP	1	E.4.8	Earth and Space Science
4	1	12	MC	OP	1	H.4.1	Science Applications & Science in Social and Personal Perspectives
4	1	13	MC	OP	1	G.4.3	Science Applications & Science in Social and Personal Perspectives
4	1	14	MC	OP	1	C.4.7	Science Inquiry
4	1	15	MC	OP	1	D.4.3	Physical Science
4	1	16	MC	OP	1	F.4.2	Life and Environmental Science
4	1	17	MC	OP	1	A.4.4	Science Connections & Nature of Science
4	1	18	MC	OP	1	E.4.6	Earth and Space Science
4	1	19	MC	OP	1	G.4.5	Science Applications & Science in Social and Personal Perspectives
4	1	20	MC	OP	1	F.4.2	Life and Environmental Science



Table E-1. Science, Grade 4 Test Map (cont.)

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Standard	Domain
4	1	21	MC	OP	1	A.4.3	Science Connections & Nature of Science
4	2	25	MC	OP	1	F.4.4	Life and Environmental Science
4	2	26	MC	OP	1	H.4.2	Science Applications & Science in Social and Personal Perspectives
4	2	27	MC	OP	1	D.4.4	Physical Science
4	2	28	MC	OP	1	D.4.8	Physical Science
4	2	29	MC	OP	1	B.4.1	Science Connections & Nature of Science
4	2	30	MC	OP	1	C.4.5	Science Inquiry
4	2	31	MC	OP	1	F.4.1	Life and Environmental Science
4	2	32	MC	OP	1	E.4.5	Earth and Space Science
4	2	33	MC	OP	1	C.4.4	Science Inquiry
4	2	34	MC	OP	1	C.4.8	Science Inquiry
4	2	35	MC	OP	1	F.4.4	Life and Environmental Science
4	2	36	MC	OP	1	H.4.1	Science Applications & Science in Social and Personal Perspectives
4	2	37	MC	OP	1	B.4.1	Science Connections & Nature of Science
4	2	38	MC	OP	1	G.4.1	Science Applications & Science in Social and Personal Perspectives
4	2	39	MC	OP	1	E.4.4	Earth and Space Science
4	2	40	MC	OP	1	G.4.4	Science Applications & Science in Social and Personal Perspectives
4	2	41	MC	OP	1	D.4.8	Physical Science
4	2	42	MC	OP	1	C.4.1	Science Inquiry
4	2	43	MC	OP	1	B.4.2	Science Connections & Nature of Science

Table E-2. Science, Grade 8 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Standard	Domain
8	1	1	MC	OP	1	B.8.3	Science Connections & Nature of Science
8	1	2	MC	OP	1	G.8.6	Science Applications & Science in Social and Personal Perspectives
8	1	3	MC	OP	1	A.8.5	Science Connections & Nature of Science
8	1	4	MC	OP	1	C.8.3	Science Inquiry
8	1	5	MC	OP	1	C.8.4	Science Inquiry
8	1	6	MC	OP	1	D.8.8	Physical Science
8	1	7	MC	OP	1	G.8.1	Science Applications & Science in Social and Personal Perspectives
8	1	8	MC	OP	1	D.8.6	Physical Science
8	1	9	MC	OP	1	C.8.6	Science Inquiry
8	1	10	MC	OP	1	E.8.3	Earth and Space Science
8	1	11	MC	OP	1	F.8.8	Life and Environmental Science
8	1	12	MC	OP	1	F.8.9	Life and Environmental Science
8	1	13	MC	OP	1	D.8.2	Physical Science
8	1	14	MC	OP	1	G.8.7	Science Applications & Science in Social and Personal Perspectives
8	1	15	MC	OP	1	E.8.3	Earth and Space Science
8	1	16	MC	OP	1	F.8.8	Life and Environmental Science
8	1	17	MC	OP	1	E.8.1	Earth and Space Science
8	1	18	MC	OP	1	B.8.1	Science Connections & Nature of Science
8	1	19	MC	OP	1	C.8.1	Science Inquiry
8	1	20	MC	OP	1	C.8.2	Science Inquiry
8	1	21	MC	OP	1	D.8.6	Physical Science

Table E-2. Science, Grade 8 Test Map (cont.)

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Standard	Domain
8	2	25	MC	OP	1	E.8.5	Earth and Space Science
8	2	26	MC	OP	1	A.8.6	Science Connections & Nature of Science
8	2	27	MC	OP	1	E.8.2	Earth and Space Science
8	2	28	MC	OP	1	F.8.1	Life and Environmental Science
8	2	29	MC	OP	1	G.8.3	Science Applications & Science in Social and Personal Perspectives
8	2	30	MC	OP	1	F.8.8	Life and Environmental Science
8	2	31	MC	OP	1	D.8.8	Physical Science
8	2	32	MC	OP	1	A.8.3	Science Connections & Nature of Science
8	2	33	MC	OP	1	C.8.6	Science Inquiry
8	2	34	MC	OP	1	G.8.3	Science Applications & Science in Social and Personal Perspectives
8	2	35	MC	OP	1	G.8.4	Science Applications & Science in Social and Personal Perspectives
8	2	36	MC	OP	1	G.8.5	Science Applications & Science in Social and Personal Perspectives
8	2	37	MC	OP	1	H.8.3	Science Applications & Science in Social and Personal Perspectives
8	2	38	MC	OP	1	B.8.6	Science Connections & Nature of Science
8	2	39	MC	OP	1	C.8.6	Science Inquiry
8	2	40	MC	OP	1	C.8.6	Science Inquiry
8	2	41	MC	OP	1	B.8.4	Science Connections & Nature of Science
8	2	42	MC	OP	1	C.8.10	Science Inquiry
8	2	43	MC	OP	1	F.8.8	Life and Environmental Science

## **Appendix F**

### **Spring 2017 Social Studies Operational Test Maps**

Table F-1. Social Studies, Grade 4 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Standard	Domain
4	1	1	MC	OP	1	A.4.2	Geography
4	1	2	MC	OP	1	A.4.5	Geography
4	1	3	MC	OP	1	B.4.6	History
4	1	4	MC	OP	1	D.4.2	Economics
4	1	5	MC	OP	1	D.4.1	Economics
4	1	6	MC	OP	1	C.4.5	Political Science and Citizenship
4	1	7	MC	OP	1	C.4.4	Political Science and Citizenship
4	1	8	MC	OP	1	C.4.3	Political Science and Citizenship
4	1	9	MC	OP	1	A.4.9	Geography
4	1	10	MC	OP	1	E.4.15	Behavioral Sciences
4	1	11	MC	OP	1	B.4.2	History
4	1	12	MC	OP	1	C.4.2	Political Science and Citizenship
4	1	13	MC	OP	1	C.4.6	Political Science and Citizenship
4	1	14	MC	OP	1	E.4.12	Behavioral Sciences
4	1	19	MC	OP	1	B.4.3	History
4	1	20	MC	OP	1	B.4.8	History
4	1	21	MC	OP	1	B.4.1	History
4	1	22	MC	OP	1	A.4.4	Geography
4	1	23	MC	OP	1	E.4.3	Behavioral Sciences
4	2	24	MC	OP	1	B.4.7	History
4	2	25	MC	OP	1	C.4.1	Political Science and Citizenship
4	2	26	MC	OP	1	B.4.10	History
4	2	27	MC	OP	1	E.4.11	Behavioral Sciences
4	2	30	MC	OP	1	C.4.2	Political Science and Citizenship
4	2	31	MC	OP	1	A.4.9	Geography
4	2	32	MC	OP	1	A.4.5	Geography
4	2	33	MC	OP	1	A.4.1	Geography
4	2	34	MC	OP	1	A.4.2	Geography
4	2	35	MC	OP	1	E.4.10	Behavioral Sciences
4	2	36	MC	OP	1	D.4.5	Economics
4	2	38	MC	OP	1	B.4.2	History

Table F-1. Social Studies, Grade 4 Test Map (cont.)

<b>Grade</b>	<b>Session</b>	<b>Item Sequence</b>	<b>Item Type</b>	<b>Item Usage</b>	<b>Max Score Points</b>	<b>Standard</b>	<b>Domain</b>
4	2	41	MC	OP	1	E.4.15	Behavioral Sciences
4	2	42	MC	OP	1	D.4.4	Economics
4	2	43	MC	OP	1	D.4.7	Economics
4	2	44	MC	OP	1	E.4.15	Behavioral Sciences
4	2	45	MC	OP	1	D.4.4	Economics
4	2	46	MC	OP	1	E.4.15	Behavioral Sciences

Table F-2. Social Studies, Grade 8 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Standard	Domain
8	1	1	MC	OP	1	B.8.1	History
8	1	2	MC	OP	1	B.8.1	History
8	1	3	MC	OP	1	A.8.7	Geography
8	1	4	MC	OP	1	A.8.8	Geography
8	1	5	MC	OP	1	A.8.10	Geography
8	1	6	MC	OP	1	C.8.9	Political Science and Citizenship
8	1	7	MC	OP	1	D.8.2	Economics
8	1	8	MC	OP	1	D.8.2	Economics
8	1	9	MC	OP	1	D.8.8	Economics
8	1	10	MC	OP	1	D.8.8	Economics
8	1	15	MC	OP	1	B.8.5	History
8	1	16	MC	OP	1	C.8.9	Political Science and Citizenship
8	1	17	MC	OP	1	A.8.5	Geography
8	1	18	MC	OP	1	E.8.9	Behavioral Sciences
8	1	19	MC	OP	1	B.8.7	History
8	1	20	MC	OP	1	B.8.4	History
8	1	21	MC	OP	1	B.8.1	History
8	1	22	MC	OP	1	A.8.9	Geography
8	1	23	MC	OP	1	A.8.11	Geography
8	1	24	MC	OP	1	B.8.1	History
8	2	25	MC	OP	1	A.8.2	Geography
8	2	26	MC	OP	1	A.8.2	Geography
8	2	27	MC	OP	1	B.8.7	History
8	2	28	MC	OP	1	C.8.4	Political Science and Citizenship
8	2	29	MC	OP	1	B.8.2	History
8	2	34	MC	OP	1	C.8.1	Political Science and Citizenship
8	2	35	MC	OP	1	A.8.9	Geography
8	2	36	MC	OP	1	A.8.2	Geography
8	2	37	MC	OP	1	E.8.10	Behavioral Sciences
8	2	38	MC	OP	1	B.8.9	History
8	2	39	MC	OP	1	E.8.8	Behavioral Sciences
8	2	40	MC	OP	1	D.8.2	Economics
8	2	41	MC	OP	1	B.8.7	History
8	2	42	MC	OP	1	C.8.6	Political Science and Citizenship

Table F-2. Social Studies, Grade 8 Test Map (cont.)

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Standard	Domain
8	2	43	MC	OP	1	E.8.14	Behavioral Sciences
8	2	44	MC	OP	1	D.8.2	Economics
8	2	45	MC	OP	1	C.8.8	Political Science and Citizenship
8	2	46	MC	OP	1	D.8.7	Economics
8	2	47	MC	OP	1	B.8.10	History
8	2	48	MC	OP	1	E.8.4	Behavioral Sciences



Table F-3. Social Studies, Grade 10 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Standard	Domain
10	1	1	MC	OP	1	C.10.16	Political Science and Citizenship
10	1	2	MC	OP	1	B.10.7	History
10	1	3	MC	OP	1	E.10.6	Behavioral Sciences
10	1	4	MC	OP	1	B.10.14	History
10	1	5	MC	OP	1	A.10.3	Geography
10	1	6	MC	OP	1	B.10.8	History
10	1	7	MC	OP	1	B.10.14	History
10	1	8	MC	OP	1	A.10.7	Geography
10	1	9	MC	OP	1	B.10.6	History
10	1	10	MC	OP	1	A.10.6	Geography
10	1	11	MC	OP	1	D.10.10	Economics
10	1	12	MC	OP	1	D.10.8	Economics
10	1	13	MC	OP	1	B.10.16	History
10	1	14	MC	OP	1	C.10.10	Political Science and Citizenship
10	1	15	MC	OP	1	E.10.6	Behavioral Sciences
10	1	16	MC	OP	1	D.10.7	Economics
10	1	17	MC	OP	1	E.10.5	Behavioral Sciences
10	1	18	MC	OP	1	C.10.6	Political Science and Citizenship
10	1	24	MC	OP	1	C.10.1	Political Science and Citizenship
10	1	25	MC	OP	1	A.10.8	Geography
10	1	26	MC	OP	1	D.10.1	Economics
10	1	27	MC	OP	1	C.10.14	Political Science and Citizenship
10	1	28	MC	OP	1	C.10.13	Political Science and Citizenship
10	1	29	MC	OP	1	C.10.13	Political Science and Citizenship
10	1	30	MC	OP	1	C.10.13	Political Science and Citizenship
10	2	31	MC	OP	1	D.10.4	Economics
10	2	32	MC	OP	1	B.10.6	History
10	2	33	MC	OP	1	A.10.1	Geography

Table F-3. Social Studies, Grade 10 Test Map (cont.)

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Standard	Domain
10	2	34	MC	OP	1	A.10.1	Geography
10	2	35	MC	OP	1	A.10.4	Geography
10	2	36	MC	OP	1	D.10.2	Economics
10	2	37	MC	OP	1	D.10.2	Economics
10	2	38	MC	OP	1	E.10.8	Behavioral Sciences
10	2	44	MC	OP	1	E.10.12	Behavioral Sciences
10	2	45	MC	OP	1	C.10.2	Political Science and Citizenship
10	2	46	MC	OP	1	B.10.14	History
10	2	47	MC	OP	1	D.10.7	Economics
10	2	48	MC	OP	1	A.10.5	Geography
10	2	49	MC	OP	1	C.10.12	Political Science and Citizenship
10	2	50	MC	OP	1	C.10.6	Political Science and Citizenship
10	2	51	MC	OP	1	A.10.12	Geography
10	2	52	MC	OP	1	B.10.3	History
10	2	53	MC	OP	1	E.10.14	Behavioral Sciences
10	2	54	MC	OP	1	E.10.14	Behavioral Sciences
10	2	55	MC	OP	1	B.10.16	History
10	2	56	MC	OP	1	B.10.15	History
10	2	57	MC	OP	1	E.10.17	Behavioral Sciences
10	2	58	MC	OP	1	C.10.11	Political Science and Citizenship
10	2	59	MC	OP	1	A.10.8	Geography
10	2	60	MC	OP	1	B.10.12	History

## **Appendix G**

### **Classical Item Analysis Results**

Table G-1. Item Statistics, ELA Grade 3

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Percent Omit	Percent of Students					Item-Total Test Correlation				
							Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4	Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4
1	MC	1	63766	0.64	0.37	0.12		0.64	0.21	0.08	0.07		0.37	-0.24	-0.16	-0.15
2	MC	1	63773	0.83	0.28	0.11		0.08	0.83	0.05	0.05		-0.07	0.28	-0.22	-0.18
3	MC	1	63765	0.57	0.31	0.12		0.10	0.15	0.57	0.17		-0.24	-0.08	0.31	-0.14
4	TDA	4	48242	0.33	0.45	0.26		0.54	0.19	0.02	0.00		0.08	0.43	0.20	0.01
5	MC	1	63792	0.67	0.30	0.09		0.11	0.18	0.67	0.04		-0.15	-0.15	0.30	-0.18
6	TE	1	63494	0.58	0.34	0.56	0.41	0.58				-0.33	0.34			
7	TE	2	63779	0.74	0.49	0.11	0.09	0.35	0.56			-0.32	-0.28	0.46		
8	TE	1	63373	0.43	0.21	0.75	0.57	0.42				-0.20	0.22			
9	MC	1	63737	0.81	0.43	0.18		0.09	0.81	0.06	0.04		-0.26	0.43	-0.22	-0.20
10	TE	2	63756	0.51	0.44	0.15	0.22	0.54	0.24			-0.37	0.03	0.33		
11	MC	1	63755	0.64	0.32	0.15		0.64	0.17	0.11	0.08		0.32	-0.11	-0.18	-0.20
12	MC	1	63723	0.60	0.41	0.20		0.15	0.14	0.11	0.60		-0.23	-0.14	-0.22	0.41
13	MC	1	63731	0.53	0.30	0.18		0.13	0.53	0.15	0.19		-0.22	0.30	-0.14	-0.06
14	MC	1	63728	0.49	0.30	0.19		0.49	0.18	0.20	0.13		0.30	-0.14	-0.09	-0.17
15	TE	2	63622	0.40	0.29	0.36	0.34	0.51	0.15			-0.20	0.00	0.28		
16	ESR	2	63804	0.71	0.43	0.06	0.06	0.44	0.49			-0.24	-0.30	0.41		
17	MC	1	63751	0.69	0.49	0.15		0.08	0.69	0.04	0.18		-0.22	0.49	-0.21	-0.31
18	MC	1	63760	0.88	0.35	0.13		0.03	0.03	0.88	0.06		-0.19	-0.19	0.35	-0.20
19	ESR	2	63769	0.57	0.51	0.12	0.39	0.09	0.52			-0.46	-0.09	0.51		
20	MC	1	63718	0.56	0.28	0.20		0.30	0.10	0.04	0.56		-0.09	-0.18	-0.22	0.28

Note: TDA responses that received a condition code were not included in item analysis.

Table G-1. Item Statistics, ELA Grade 3 (cont.)

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Percent Omit	Percent of Students					Item-Total Test Correlation				
							Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4	Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4
21	MC	1	63757	0.56	0.28	0.12		0.18	0.56	0.18	0.08		-0.23	0.28	-0.01	-0.18
22	ESR	2	63775	0.65	0.53	0.09	0.22	0.26	0.52			-0.43	-0.17	0.50		
23	TE	2	63632	0.49	0.58	0.32	0.30	0.41	0.29			-0.46	-0.03	0.51		
24	MC	1	63675	0.50	0.33	0.25		0.27	0.12	0.50	0.11		-0.11	-0.18	0.33	-0.18
25	MC	1	63676	0.64	0.44	0.25		0.23	0.06	0.64	0.07		-0.21	-0.27	0.44	-0.22
26	ESR	2	63748	0.40	0.46	0.14	0.51	0.19	0.30			-0.40	-0.02	0.45		
27	MC	1	63686	0.61	0.45	0.23		0.19	0.61	0.06	0.14		-0.23	0.45	-0.27	-0.18
28	MC	1	63679	0.52	0.44	0.24		0.11	0.27	0.10	0.52		-0.28	-0.10	-0.28	0.44
29	MC	1	63686	0.64	0.37	0.23		0.64	0.15	0.15	0.06		0.37	-0.15	-0.18	-0.24
30	TE	2	63604	0.69	0.52	0.36	0.10	0.41	0.49			-0.27	-0.35	0.52		
31	MC	1	63651	0.41	0.33	0.29		0.34	0.41	0.08	0.16		-0.03	0.33	-0.25	-0.20
32	MC	1	63663	0.51	0.36	0.27		0.22	0.11	0.16	0.50		-0.15	-0.19	-0.15	0.36
33	MC	1	63663	0.37	0.24	0.27		0.20	0.36	0.21	0.22		-0.05	0.24	-0.26	0.04

Table G-2. Item Statistics, ELA Grade 4

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Percent Omit	Percent of Students					Item-Total Test Correlation				
							Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4	Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4
1	MC	1	64298	0.49	0.43	0.08		0.13	0.20	0.17	0.49		-0.35	-0.08	-0.17	0.43
2	MC	1	64276	0.80	0.47	0.11		0.08	0.06	0.80	0.05		-0.30	-0.25	0.47	-0.20
3	TE	2	64262	0.57	0.56	0.13	0.38	0.09	0.52			-0.51	-0.10	0.56		
4	TE	1	64072	0.26	0.44	0.43	0.73	0.26				-0.42	0.44			
5	MC	1	64278	0.61	0.42	0.11		0.17	0.61	0.14	0.08		-0.26	0.42	-0.22	-0.11
6	TDA	4	54551	0.37	0.54	0.30		0.53	0.24	0.07	0.00		-0.16	0.39	0.34	0.04
7	TE	1	64262	0.34	0.29	0.12	0.66	0.34				-0.29	0.29			
8	TE	1	64249	0.65	0.46	0.14	0.35	0.65				-0.46	0.46			
9	MC	1	64279	0.83	0.36	0.09		0.06	0.04	0.07	0.83		-0.21	-0.15	-0.22	0.36
10	MC	1	64250	0.65	0.32	0.14		0.20	0.65	0.11	0.04		-0.24	0.32	-0.08	-0.15
11	MC	1	64251	0.72	0.37	0.14		0.10	0.06	0.72	0.12		-0.26	-0.24	0.37	-0.09
12	MC	1	64241	0.45	0.26	0.15		0.45	0.16	0.13	0.25		0.26	-0.15	-0.19	-0.02
13	MC	1	64235	0.60	0.45	0.16		0.12	0.20	0.60	0.07		-0.25	-0.23	0.45	-0.17
14	TE	2	64251	0.54	0.47	0.14	0.18	0.56	0.25			-0.41	0.02	0.35		
15	MC	1	64247	0.71	0.46	0.14		0.16	0.09	0.71	0.04		-0.26	-0.24	0.46	-0.21
16	MC	1	64255	0.59	0.19	0.13		0.59	0.23	0.09	0.09		0.19	0.03	-0.17	-0.19
17	MC	1	64236	0.72	0.47	0.16		0.72	0.09	0.09	0.10		0.47	-0.26	-0.22	-0.25
18	TE	2	63990	0.70	0.38	0.54	0.06	0.48	0.45			-0.17	-0.29	0.38		
19	TE	2	64239	0.80	0.35	0.16	0.02	0.35	0.63			-0.20	-0.26	0.32		
20	TE	2	64275	0.69	0.48	0.07	0.14	0.36	0.51			-0.33	-0.24	0.46		

Note: TDA responses that received a condition code were not included in item analysis.

Table G-2. Item Statistics, ELA Grade 4 (cont.)

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Percent Omit	Percent of Students					Item-Total Test Correlation				
							Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4	Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4
21	MC	1	64229	0.60	0.36	0.14		0.14	0.15	0.60	0.10		-0.23	-0.14	0.36	-0.15
22	MC	1	64228	0.56	0.31	0.14		0.28	0.07	0.09	0.56		-0.06	-0.21	-0.25	0.31
23	ESR	2	64253	0.61	0.44	0.10	0.24	0.30	0.46			-0.40	-0.03	0.37		
24	MC	1	64199	0.73	0.40	0.19		0.08	0.73	0.08	0.11		-0.21	0.40	-0.24	-0.19
25	MC	1	64198	0.38	0.30	0.19		0.11	0.24	0.27	0.38		-0.11	-0.14	-0.11	0.30
26	MC	1	64226	0.57	0.36	0.14		0.08	0.12	0.23	0.57		-0.22	-0.31	-0.04	0.36
27	MC	1	64219	0.89	0.40	0.15		0.04	0.89	0.04	0.04		-0.24	0.41	-0.23	-0.20
28	MC	1	64223	0.81	0.42	0.14		0.08	0.81	0.05	0.06		-0.20	0.43	-0.26	-0.22
29	MC	1	64182	0.75	0.49	0.20		0.12	0.75	0.07	0.05		-0.27	0.49	-0.26	-0.25
30	TE	1	64030	0.28	0.47	0.44	0.71	0.28				-0.46	0.48			
31	MC	1	64180	0.47	0.30	0.21		0.47	0.06	0.19	0.28		0.30	-0.24	-0.25	0.02
32	TE	2	64209	0.74	0.50	0.16	0.07	0.37	0.55			-0.32	-0.31	0.47		
33	MC	1	64175	0.56	0.42	0.21		0.17	0.13	0.56	0.14		-0.13	-0.26	0.42	-0.20
34	MC	1	64150	0.55	0.35	0.25		0.55	0.17	0.17	0.11		0.36	-0.03	-0.23	-0.24
35	MC	1	64156	0.69	0.50	0.24		0.12	0.08	0.11	0.69		-0.21	-0.27	-0.29	0.50
36	ESR	2	64212	0.30	0.39	0.16	0.59	0.22	0.18			-0.31	0.01	0.39		
37	MC	1	64182	0.42	0.32	0.20		0.42	0.14	0.19	0.25		0.32	-0.19	-0.18	-0.04

Table G-3. Item Statistics, ELA Grade 5

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Percent Omit	Percent of Students					Item-Total Test Correlation				
							Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4	Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4
1	MC	1	62901	0.84	0.37	0.05		0.05	0.84	0.04	0.07		-0.23	0.37	-0.21	-0.17
2	MC	1	62904	0.80	0.37	0.05		0.80	0.08	0.04	0.08		0.37	-0.20	-0.20	-0.20
3	MC	1	62893	0.78	0.43	0.07		0.08	0.09	0.78	0.05		-0.25	-0.24	0.43	-0.19
4	TDA	4	57387	0.33	0.49	0.12		0.64	0.25	0.02	0.00		-0.20	0.45	0.20	0.01
5	MC	1	62910	0.66	0.22	0.05		0.05	0.06	0.66	0.24		-0.15	-0.07	0.22	-0.13
6	MC	1	62876	0.70	0.28	0.10		0.07	0.70	0.13	0.09		-0.15	0.28	-0.15	-0.14
7	MC	1	62882	0.84	0.42	0.09		0.84	0.04	0.07	0.05		0.42	-0.24	-0.21	-0.24
8	MC	1	62879	0.68	0.33	0.10		0.10	0.08	0.68	0.14		-0.10	-0.21	0.33	-0.18
9	MC	1	62873	0.68	0.45	0.11		0.68	0.14	0.10	0.08		0.45	-0.26	-0.20	-0.21
10	MC	1	62872	0.76	0.32	0.11		0.03	0.76	0.08	0.12		-0.21	0.32	-0.18	-0.14
11	MC	1	62874	0.82	0.42	0.10		0.07	0.05	0.82	0.05		-0.24	-0.25	0.43	-0.19
12	TE	2	62887	0.59	0.46	0.08	0.18	0.46	0.36			-0.34	-0.13	0.41		
13	MC	1	62888	0.85	0.38	0.08		0.07	0.05	0.04	0.85		-0.23	-0.14	-0.25	0.38
14	MC	1	62860	0.49	0.29	0.13		0.16	0.22	0.48	0.14		-0.23	-0.10	0.29	-0.05
15	MC	1	62841	0.76	0.44	0.16		0.10	0.06	0.76	0.09		-0.20	-0.25	0.44	-0.25
16	MC	1	62865	0.61	0.31	0.12		0.03	0.28	0.09	0.60		-0.24	-0.14	-0.18	0.31
17	MC	1	62873	0.64	0.40	0.11		0.19	0.64	0.11	0.05		-0.21	0.40	-0.19	-0.22
18	TE	2	62886	0.71	0.38	0.09	0.11	0.35	0.53			-0.34	-0.09	0.31		
19	TE	2	62905	0.61	0.42	0.05	0.13	0.51	0.36			-0.33	-0.11	0.34		
20	MC	1	62869	0.62	0.34	0.10		0.62	0.30	0.04	0.04		0.34	-0.22	-0.15	-0.19

Note: TDA responses that received a condition code were not included in item analysis.



Table G-3. Item Statistics, ELA Grade 5 (cont.)

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Percent Omit	Percent of Students					Item-Total Test Correlation				
							Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4	Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4
21	MC	1	62890	0.79	0.42	0.07		0.14	0.04	0.03	0.79		-0.28	-0.21	-0.19	0.42
22	MC	1	62890	0.33	0.27	0.07		0.08	0.33	0.28	0.31		-0.24	0.27	-0.17	0.03
23	MC	1	62861	0.56	0.33	0.12		0.22	0.13	0.56	0.08		-0.15	-0.19	0.33	-0.13
24	ESR	2	62893	0.39	0.45	0.07	0.53	0.17	0.30			-0.39	-0.01	0.44		
25	TE	2	62873	0.42	0.40	0.08	0.34	0.46	0.19			-0.29	-0.02	0.38		
26	MC	1	62821	0.55	0.33	0.16		0.26	0.55	0.12	0.07		-0.10	0.33	-0.17	-0.26
27	MC	1	62778	0.44	0.35	0.23		0.22	0.44	0.14	0.20		-0.16	0.35	-0.17	-0.12
28	MC	1	62803	0.56	0.46	0.19		0.19	0.15	0.10	0.56		-0.23	-0.23	-0.19	0.46
29	ESR	2	62870	0.23	0.30	0.08	0.59	0.35	0.05			-0.29	0.24	0.15		
30	MC	1	62794	0.61	0.39	0.21		0.18	0.09	0.61	0.12		-0.21	-0.21	0.39	-0.14
31	MC	1	62785	0.66	0.53	0.22		0.10	0.12	0.12	0.66		-0.26	-0.24	-0.30	0.54
32	ESR	2	62869	0.41	0.32	0.09	0.39	0.39	0.22			-0.36	0.24	0.15		
33	MC	1	62757	0.51	0.29	0.26		0.29	0.11	0.09	0.51		-0.02	-0.21	-0.24	0.29
34	MC	1	62773	0.53	0.45	0.24		0.53	0.15	0.23	0.09		0.45	-0.17	-0.24	-0.23
35	TE	2	62806	0.45	0.38	0.19	0.34	0.42	0.25			-0.30	0.00	0.34		
36	MC	1	62803	0.55	0.31	0.19		0.12	0.55	0.12	0.20		-0.12	0.31	-0.24	-0.09
37	MC	1	62797	0.40	0.21	0.20		0.15	0.21	0.39	0.25		-0.13	-0.06	0.21	-0.07

Table G-4. Item Statistics, ELA Grade 6

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Percent Omit	Percent of Students					Item-Total Test Correlation				
							Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4	Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4
1	TE	2	62647	0.72	0.46	0.05	0.10	0.35	0.54			-0.31	-0.26	0.44		
2	MC	1	62632	0.73	0.39	0.07		0.73	0.04	0.17	0.06		0.39	-0.24	-0.18	-0.26
3	MC	1	62631	0.48	0.34	0.08		0.48	0.29	0.10	0.13		0.34	-0.06	-0.23	-0.22
4	TDA	4	57843	0.39	0.57	0.24		0.49	0.35	0.08	0.01		-0.33	0.36	0.34	0.14
5	MC	1	62648	0.82	0.17	0.05		0.82	0.04	0.09	0.05		0.18	-0.18	-0.11	0.00
6	MC	1	62630	0.88	0.39	0.07		0.01	0.88	0.09	0.02		-0.16	0.39	-0.30	-0.15
7	MC	1	62622	0.76	0.39	0.09		0.76	0.14	0.08	0.02		0.39	-0.26	-0.19	-0.17
8	MC	1	62601	0.68	0.26	0.12		0.67	0.07	0.13	0.12		0.26	-0.14	-0.13	-0.13
9	TE	1	62533	0.54	0.30	0.23	0.46	0.54				-0.29	0.30			
10	TE	2	62606	0.69	0.41	0.11	0.09	0.44	0.47			-0.34	-0.14	0.34		
11	TE	1	62553	0.33	0.20	0.20	0.67	0.33				-0.19	0.20			
12	TE	2	62534	0.55	0.49	0.23	0.21	0.47	0.32			-0.41	-0.03	0.40		
13	MC	1	62613	0.81	0.39	0.10		0.07	0.81	0.06	0.06		-0.23	0.40	-0.24	-0.17
14	TE	2	62570	0.25	0.23	0.17	0.58	0.34	0.08			-0.17	0.07	0.21		
15	TE	2	62513	0.40	0.41	0.26	0.36	0.47	0.17			-0.35	0.11	0.31		
16	MC	1	62582	0.56	0.37	0.15		0.03	0.24	0.56	0.17		-0.24	-0.16	0.37	-0.19
17	MC	1	62619	0.82	0.37	0.07		0.82	0.06	0.09	0.03		0.37	-0.17	-0.25	-0.18
18	MC	1	62561	0.85	0.38	0.16		0.04	0.07	0.04	0.85		-0.18	-0.22	-0.23	0.38
19	TE	2	62591	0.63	0.44	0.11	0.11	0.53	0.36			-0.33	-0.15	0.37		
20	ESR	2	62616	0.37	0.25	0.07	0.48	0.31	0.21			-0.19	-0.02	0.25		

Note: TDA responses that received a condition code were not included in item analysis.

Table G-4. Item Statistics, ELA Grade 6 (cont.)

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Percent Omit	Percent of Students					Item-Total Test Correlation				
							Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4	Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4
21	MC	1	62572	0.50	0.32	0.14		0.50	0.14	0.31	0.05		0.32	-0.17	-0.09	-0.26
22	MC	1	62556	0.65	0.40	0.17		0.65	0.09	0.18	0.07		0.40	-0.23	-0.20	-0.17
23	TE	2	62441	0.69	0.48	0.35	0.13	0.34	0.52			-0.34	-0.22	0.45		
24	MC	1	62514	0.54	0.29	0.23		0.21	0.08	0.54	0.17		-0.09	-0.28	0.29	-0.09
25	MC	1	62488	0.51	0.38	0.27		0.12	0.51	0.17	0.19		-0.20	0.38	-0.19	-0.13
26	MC	1	62498	0.46	0.35	0.26		0.46	0.11	0.26	0.17		0.35	-0.28	-0.15	-0.06
27	MC	1	62544	0.40	0.33	0.18		0.13	0.05	0.41	0.40		-0.17	-0.17	-0.13	0.33
28	MC	1	62547	0.79	0.42	0.18		0.79	0.07	0.04	0.09		0.42	-0.26	-0.26	-0.17
29	MC	1	62520	0.78	0.40	0.22		0.08	0.05	0.78	0.09		-0.21	-0.22	0.41	-0.22
30	TE	2	62435	0.59	0.44	0.36	0.15	0.52	0.32			-0.36	-0.06	0.35		
31	MC	1	62483	0.75	0.48	0.28		0.75	0.07	0.13	0.04		0.49	-0.26	-0.29	-0.21
32	MC	1	62487	0.72	0.42	0.27		0.17	0.07	0.72	0.04		-0.21	-0.28	0.43	-0.20
33	TE	2	62503	0.68	0.47	0.25	0.13	0.38	0.49			-0.37	-0.16	0.41		
34	TE	1	62478	0.58	0.36	0.29	0.42	0.57				-0.35	0.36			
35	MC	1	62517	0.51	0.33	0.23		0.22	0.19	0.51	0.08		-0.12	-0.10	0.33	-0.26

Table G-5. Item Statistics, ELA Grade 7

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Percent Omit	Percent of Students					Item-Total Test Correlation				
							Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4	Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4
1	MC	1	62951	0.70	0.37	0.06		0.13	0.06	0.69	0.12		-0.18	-0.20	0.38	-0.20
2	MC	1	62945	0.75	0.46	0.07		0.03	0.04	0.17	0.75		-0.21	-0.21	-0.32	0.46
3	MC	1	62935	0.79	0.47	0.08		0.05	0.10	0.78	0.06		-0.22	-0.32	0.47	-0.21
4	TDA	4	59884	0.45	0.60	0.29		0.35	0.45	0.13	0.02		-0.42	0.25	0.37	0.21
5	TE	1	62849	0.65	0.28	0.22	0.35	0.65				-0.28	0.29			
6	MC	1	62896	0.53	0.23	0.15		0.19	0.17	0.12	0.53		-0.09	-0.10	-0.13	0.23
7	MC	1	62898	0.47	0.36	0.15		0.21	0.04	0.47	0.28		-0.16	-0.24	0.36	-0.16
8	MC	1	62899	0.72	0.30	0.14		0.07	0.18	0.03	0.72		-0.20	-0.12	-0.23	0.30
9	ESR	2	62919	0.57	0.46	0.11	0.41	0.06	0.54			-0.40	-0.20	0.49		
10	TE	2	62810	0.65	0.31	0.29	0.08	0.55	0.37			-0.17	-0.18	0.29		
11	TE	2	62852	0.46	0.25	0.22	0.18	0.72	0.10			-0.21	0.08	0.17		
12	MC	1	62906	0.80	0.37	0.13		0.80	0.08	0.08	0.04		0.37	-0.24	-0.27	-0.04
13	MC	1	62893	0.51	0.39	0.15		0.21	0.13	0.51	0.15		-0.14	-0.33	0.39	-0.08
14	MC	1	62876	0.64	0.35	0.18		0.09	0.04	0.64	0.23		-0.24	-0.23	0.36	-0.13
15	MC	1	62851	0.59	0.26	0.22		0.20	0.12	0.58	0.09		-0.10	-0.22	0.26	-0.06
16	ESR	2	62905	0.50	0.47	0.13	0.40	0.20	0.40			-0.40	-0.06	0.45		
17	MC	1	62916	0.76	0.38	0.07		0.17	0.03	0.76	0.04		-0.26	-0.15	0.39	-0.21
18	ESR	2	62929	0.70	0.55	0.05	0.24	0.11	0.65			-0.49	-0.17	0.55		
19	MC	1	62880	0.44	0.43	0.13		0.09	0.14	0.44	0.33		-0.19	-0.29	0.43	-0.12
20	ESR	2	62912	0.60	0.43	0.08	0.24	0.31	0.45			-0.38	-0.05	0.37		

Note: TDA responses that received a condition code were not included in item analysis.

Table G-5. Item Statistics, ELA Grade 7 (cont.)

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Percent Omit	Percent of Students					Item-Total Test Correlation				
							Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4	Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4
21	TE	2	62867	0.60	0.54	0.15	0.20	0.40	0.40			-0.38	-0.20	0.51		
22	MC	1	62909	0.64	0.34	0.08		0.09	0.04	0.64	0.22		-0.16	-0.25	0.34	-0.16
23	MC	1	62827	0.65	0.38	0.21		0.14	0.65	0.11	0.10		-0.19	0.39	-0.19	-0.19
24	MC	1	62849	0.80	0.47	0.17		0.07	0.06	0.79	0.08		-0.26	-0.27	0.47	-0.23
25	ESR	2	62905	0.65	0.49	0.09	0.26	0.18	0.56			-0.44	-0.09	0.46		
26	MC	1	62797	0.50	0.21	0.26		0.30	0.13	0.50	0.07		0.05	-0.22	0.21	-0.22
27	MC	1	62820	0.56	0.45	0.22		0.06	0.17	0.20	0.56		-0.24	-0.16	-0.25	0.45
28	MC	1	62705	0.46	0.27	0.40		0.46	0.07	0.26	0.21		0.27	-0.21	-0.06	-0.12
29	TE	1	61336	0.63	0.43	2.58	0.36	0.62				-0.39	0.45			
30	MC	1	62729	0.46	0.32	0.37		0.15	0.31	0.46	0.08		-0.12	-0.13	0.32	-0.18
31	MC	1	62797	0.58	0.38	0.26		0.11	0.23	0.58	0.08		-0.21	-0.15	0.38	-0.21
32	ESR	2	62835	0.55	0.54	0.20	0.39	0.10	0.50			-0.46	-0.16	0.55		
33	MC	1	62780	0.58	0.50	0.28		0.58	0.14	0.15	0.13		0.50	-0.23	-0.23	-0.24
34	MC	1	62790	0.78	0.43	0.27		0.07	0.78	0.07	0.08		-0.22	0.44	-0.28	-0.18
35	MC	1	62795	0.67	0.45	0.26		0.15	0.67	0.08	0.10		-0.18	0.45	-0.28	-0.23
36	MC	1	62800	0.54	0.48	0.25		0.08	0.24	0.14	0.54		-0.21	-0.18	-0.29	0.48

Table G-6. Item Statistics, ELA Grade 8

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Percent Omit	Percent of Students					Item-Total Test Correlation				
							Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4	Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4
1	TE	1	61687	0.48	0.36	0.48	0.52	0.48				-0.35	0.36			
2	ESR	2	61957	0.53	0.37	0.04	0.45	0.04	0.51			-0.33	-0.20	0.41		
3	MC	1	61928	0.66	0.45	0.09		0.16	0.07	0.66	0.11		-0.27	-0.26	0.45	-0.15
4	MC	1	61930	0.67	0.20	0.09		0.16	0.11	0.67	0.06		-0.10	-0.02	0.20	-0.22
5	TDA	4	55374	0.42	0.59	0.69		0.43	0.34	0.11	0.02		-0.30	0.31	0.37	0.21
6	MC	1	61946	0.62	0.34	0.05		0.17	0.08	0.14	0.62		-0.24	-0.19	-0.08	0.35
7	MC	1	61902	0.62	0.39	0.12		0.26	0.05	0.62	0.06		-0.28	-0.15	0.39	-0.14
8	MC	1	61887	0.49	0.19	0.14		0.06	0.26	0.18	0.49		-0.24	-0.07	-0.01	0.19
9	MC	1	61903	0.66	0.37	0.12		0.15	0.08	0.66	0.12		-0.17	-0.19	0.37	-0.19
10	MC	1	61913	0.47	0.43	0.10		0.47	0.26	0.16	0.11		0.43	-0.18	-0.23	-0.16
11	TE	2	61586	0.52	0.48	0.63	0.18	0.59	0.23			-0.38	0.00	0.37		
12	MC	1	61902	0.75	0.45	0.12		0.06	0.09	0.10	0.75		-0.24	-0.25	-0.23	0.46
13	MC	1	61915	0.70	0.36	0.10		0.09	0.08	0.70	0.12		-0.19	-0.21	0.36	-0.16
14	TE	2	61853	0.77	0.51	0.20	0.01	0.44	0.55			-0.15	-0.47	0.51		
15	MC	1	61887	0.59	0.41	0.14		0.11	0.19	0.11	0.59		-0.22	-0.17	-0.21	0.42
16	TE	1	61693	0.49	0.35	0.46	0.51	0.49				-0.34	0.35			
17	MC	1	61878	0.54	0.27	0.16		0.15	0.54	0.17	0.14		-0.16	0.27	-0.15	-0.06
18	TE	2	61888	0.52	0.41	0.14	0.20	0.56	0.24			-0.36	0.05	0.29		
19	MC	1	61894	0.78	0.31	0.06		0.03	0.15	0.78	0.04		-0.20	-0.17	0.31	-0.18
20	MC	1	61847	0.69	0.36	0.14		0.69	0.02	0.24	0.04		0.36	-0.23	-0.22	-0.18

Note: TDA responses that received a condition code were not included in item analysis.

Table G-6. Item Statistics, ELA Grade 8 (cont.)

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Percent Omit	Percent of Students					Item-Total Test Correlation				
							Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4	Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4
21	ESR	2	61904	0.47	0.38	0.05	0.45	0.15	0.40			-0.34	-0.03	0.37		
22	MC	1	61848	0.49	0.41	0.14		0.27	0.13	0.10	0.49		-0.09	-0.26	-0.24	0.41
23	MC	1	61831	0.74	0.42	0.17		0.06	0.73	0.17	0.04		-0.22	0.42	-0.25	-0.22
24	ESR	2	61877	0.66	0.49	0.09	0.26	0.16	0.58			-0.42	-0.14	0.48		
25	MC	1	61884	0.73	0.52	0.09		0.11	0.10	0.07	0.73		-0.25	-0.29	-0.27	0.52
26	MC	1	61840	0.48	0.25	0.16		0.08	0.48	0.06	0.39		-0.16	0.25	-0.25	-0.04
27	MC	1	61825	0.65	0.42	0.18		0.64	0.20	0.11	0.05		0.42	-0.18	-0.24	-0.24
28	MC	1	61794	0.67	0.50	0.23		0.06	0.14	0.13	0.67		-0.20	-0.28	-0.27	0.50
29	MC	1	61826	0.63	0.50	0.18		0.63	0.14	0.07	0.16		0.50	-0.30	-0.22	-0.22
30	TE	1	61642	0.56	0.39	0.48	0.44	0.55				-0.39	0.40			
31	TE	2	61809	0.50	0.45	0.21	0.21	0.57	0.22			-0.31	-0.07	0.40		
32	TE	2	61783	0.73	0.58	0.25	0.09	0.36	0.54			-0.29	-0.44	0.60		
33	MC	1	61788	0.64	0.35	0.24		0.04	0.64	0.05	0.27		-0.25	0.35	-0.27	-0.13
34	TE	2	61793	0.58	0.40	0.23	0.07	0.71	0.22			-0.30	-0.12	0.31		
35	MC	1	61783	0.77	0.48	0.25		0.07	0.77	0.06	0.09		-0.26	0.48	-0.27	-0.23
36	MC	1	61796	0.61	0.42	0.23		0.61	0.18	0.12	0.08		0.42	-0.17	-0.21	-0.25

Table G-7. Item Statistics, Mathematics Grade 3

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Percent Omit	Percent of Students					Item-Total Test Correlation				
							Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4	Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4
1	MC	1	63945	0.65	0.46	0.10		0.21	0.64	0.09	0.05		-0.26	0.46	-0.25	-0.19
2	MC	1	63953	0.71	0.44	0.08		0.03	0.04	0.71	0.22		-0.15	-0.15	0.44	-0.34
3	MC	1	63932	0.77	0.47	0.12		0.14	0.03	0.77	0.06		-0.42	-0.16	0.47	-0.11
4	MC	1	63939	0.76	0.45	0.10		0.04	0.08	0.76	0.12		-0.12	-0.33	0.45	-0.24
5	MC	1	63943	0.52	0.46	0.10		0.25	0.05	0.17	0.52		-0.28	-0.07	-0.24	0.46
6	TE	1	63822	0.43	0.58	0.29	0.56	0.43				-0.58	0.59			
7	MC	1	63748	0.82	0.43	0.40		0.06	0.82	0.06	0.07		-0.14	0.44	-0.24	-0.32
8	SA	1	63888	0.52	0.38	0.18	0.48	0.52				-0.38	0.39			
9	MC	1	63919	0.65	0.39	0.14		0.65	0.12	0.12	0.10		0.39	-0.25	-0.12	-0.21
10	SA	1	63935	0.60	0.58	0.11	0.40	0.60				-0.58	0.58			
11	MC	1	63904	0.46	0.34	0.16		0.18	0.18	0.45	0.18		-0.10	-0.14	0.34	-0.20
12	SA	1	63917	0.21	0.45	0.14	0.79	0.21				-0.45	0.45			
13	MC	1	63741	0.64	0.41	0.41		0.64	0.09	0.18	0.08		0.42	-0.13	-0.26	-0.21
14	MC	1	63666	0.44	0.29	0.53		0.34	0.08	0.14	0.44		-0.01	-0.26	-0.18	0.29
15	MC	1	63866	0.40	0.30	0.22		0.31	0.40	0.15	0.13		-0.20	0.30	-0.09	-0.05
16	MC	1	63909	0.17	0.25	0.15		0.04	0.66	0.17	0.13		-0.18	-0.10	0.25	-0.04
17	MC	1	63922	0.40	0.44	0.13		0.33	0.06	0.20	0.40		-0.20	-0.25	-0.14	0.44
18	SA	1	63907	0.29	0.38	0.15	0.71	0.29				-0.38	0.38			
19	MC	1	63925	0.72	0.49	0.13		0.05	0.10	0.13	0.72		-0.23	-0.28	-0.25	0.49
20	TE	1	63381	0.85	0.37	0.98	0.15	0.84				-0.36	0.38			



Table G-7. Item Statistics, Mathematics Grade 3 (cont.)

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Percent Omit	Percent of Students					Item-Total Test Correlation				
							Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4	Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4
21	MC	1	63910	0.80	0.43	0.15		0.80	0.10	0.07	0.03		0.43	-0.30	-0.19	-0.20
22	MC	1	63903	0.46	0.36	0.13		0.46	0.33	0.08	0.13		0.36	-0.03	-0.17	-0.36
23	MC	1	63897	0.45	0.39	0.14		0.20	0.45	0.09	0.25		-0.21	0.39	-0.20	-0.11
24	MC	1	63854	0.62	0.43	0.21		0.15	0.12	0.62	0.12		-0.21	-0.23	0.43	-0.19
25	MC	1	63874	0.60	0.37	0.18		0.14	0.15	0.60	0.11		-0.21	-0.18	0.37	-0.14
26	SA	1	63888	0.59	0.58	0.15	0.41	0.59				-0.58	0.58			
27	TE	1	63773	0.61	0.32	0.33	0.39	0.61				-0.32	0.33			
28	MC	1	63647	0.65	0.43	0.53		0.06	0.06	0.23	0.65		-0.24	-0.24	-0.21	0.43
29	MC	1	63882	0.77	0.43	0.16		0.06	0.07	0.10	0.77		-0.27	-0.20	-0.21	0.43
30	MC	1	63867	0.36	0.18	0.19		0.21	0.26	0.17	0.36		-0.09	0.02	-0.15	0.18
31	MC	1	63904	0.56	0.43	0.13		0.06	0.35	0.56	0.03		-0.26	-0.25	0.44	-0.20
32	MC	1	63853	0.41	0.36	0.21		0.29	0.15	0.15	0.41		-0.12	-0.19	-0.16	0.36
33	SA	1	63895	0.86	0.39	0.14	0.14	0.86				-0.38	0.39			
34	SA	1	63742	0.25	0.31	0.38	0.75	0.25				-0.30	0.31			
35	MC	1	63687	0.63	0.50	0.47		0.24	0.63	0.08	0.05		-0.38	0.50	-0.18	-0.12
36	SA	1	63861	0.46	0.57	0.20	0.54	0.46				-0.56	0.57			
37	MC	1	63853	0.67	0.48	0.21		0.05	0.67	0.11	0.17		-0.26	0.48	-0.21	-0.27
38	SA	1	63850	0.67	0.57	0.21	0.33	0.67				-0.56	0.57			
39	MC	1	63877	0.79	0.42	0.17		0.11	0.79	0.05	0.04		-0.26	0.43	-0.24	-0.18
40	MC	1	63869	0.76	0.38	0.18		0.09	0.02	0.13	0.76		-0.14	-0.09	-0.32	0.38
41	MC	1	63841	0.50	0.55	0.23		0.07	0.50	0.38	0.04		-0.17	0.55	-0.40	-0.18
42	MC	1	63866	0.67	0.48	0.19		0.16	0.67	0.12	0.04		-0.28	0.48	-0.21	-0.26

Table G-8. Item Statistics, Mathematics Grade 4

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Percent Omit	Percent of Students					Item-Total Test Correlation				
							Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4	Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4
1	MC	1	64293	0.30	0.49	0.28		0.02	0.30	0.15	0.52		-0.07	0.49	-0.09	-0.36
2	MC	1	64411	0.44	0.32	0.09		0.05	0.44	0.06	0.44		-0.21	-0.14	-0.17	0.32
3	MC	1	64426	0.56	0.49	0.07		0.56	0.11	0.16	0.17		0.49	-0.13	-0.26	-0.28
4	MC	1	64388	0.56	0.54	0.13		0.21	0.13	0.56	0.10		-0.41	-0.25	0.54	-0.05
5	MC	1	64411	0.72	0.46	0.09		0.72	0.09	0.14	0.05		0.46	-0.26	-0.22	-0.24
6	SA	1	64329	0.26	0.53	0.22	0.74	0.26				-0.52	0.53			
7	MC	1	64381	0.63	0.43	0.14		0.08	0.63	0.12	0.16		-0.10	0.43	-0.16	-0.34
8	MC	1	64319	0.63	0.50	0.24		0.22	0.07	0.62	0.08		-0.38	-0.14	0.50	-0.17
9	MC	1	64411	0.75	0.39	0.09		0.17	0.05	0.03	0.75		-0.28	-0.17	-0.16	0.39
10	MC	1	64382	0.35	0.31	0.14		0.26	0.27	0.35	0.11		-0.16	-0.12	0.31	-0.07
11	MC	1	64371	0.59	0.37	0.16		0.15	0.58	0.16	0.10		-0.18	0.37	-0.13	-0.23
12	TE	1	64397	0.70	0.37	0.12	0.30	0.70				-0.37	0.37			
13	MC	1	64393	0.47	0.57	0.12		0.22	0.12	0.20	0.47		-0.28	-0.28	-0.20	0.57
14	MC	1	64367	0.41	0.26	0.16		0.15	0.33	0.12	0.41		-0.01	-0.22	-0.06	0.26
15	MC	1	64185	0.38	0.31	0.45		0.38	0.22	0.22	0.17		0.31	-0.12	-0.15	-0.10
16	SA	1	64285	0.25	0.43	0.29	0.75	0.25				-0.42	0.43			
17	SA	1	64376	0.40	0.58	0.15	0.59	0.40				-0.58	0.58			
18	MC	1	64367	0.38	0.42	0.16		0.17	0.29	0.16	0.38		-0.27	-0.04	-0.23	0.42
19	MC	1	64379	0.58	0.33	0.14		0.58	0.16	0.17	0.08		0.33	-0.18	-0.23	-0.04
20	TE	1	64136	0.24	0.55	0.52	0.76	0.24				-0.54	0.55			
21	MC	1	64389	0.59	0.40	0.13		0.58	0.16	0.17	0.08		0.40	-0.19	-0.18	-0.21
22	MC	1	64385	0.82	0.29	0.13		0.06	0.82	0.04	0.08		-0.15	0.29	-0.18	-0.14
23	MC	1	64321	0.43	0.45	0.23		0.18	0.42	0.13	0.25		-0.14	0.45	-0.27	-0.17

Table G-8. Item Statistics, Mathematics Grade 4 (cont.)

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Percent Omit	Percent of Students					Item-Total Test Correlation				
							Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4	Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4
24	MC	1	64273	0.81	0.40	0.27		0.81	0.08	0.06	0.05		0.41	-0.25	-0.21	-0.19
25	MC	1	64384	0.84	0.35	0.09		0.05	0.05	0.84	0.05		-0.22	-0.17	0.35	-0.17
26	MC	1	64385	0.47	0.40	0.09		0.12	0.21	0.20	0.47		-0.19	-0.13	-0.21	0.40
27	MC	1	64325	0.48	0.49	0.18		0.27	0.48	0.14	0.10		-0.37	0.49	-0.15	-0.09
28	MC	1	64349	0.35	0.58	0.15		0.35	0.08	0.43	0.13		0.58	-0.09	-0.30	-0.31
29	TE	1	64336	0.26	0.45	0.17	0.74	0.26				-0.44	0.45			
30	SA	1	64324	0.58	0.48	0.19	0.42	0.58				-0.48	0.48			
31	MC	1	64227	0.34	0.40	0.34		0.34	0.12	0.14	0.41		0.40	-0.12	-0.07	-0.25
32	MC	1	64323	0.41	0.24	0.19		0.11	0.21	0.27	0.41		-0.14	-0.11	-0.06	0.24
33	MC	1	64343	0.52	0.60	0.16		0.52	0.32	0.08	0.08		0.60	-0.51	-0.11	-0.12
34	MC	1	64311	0.63	0.47	0.21		0.14	0.63	0.16	0.06		-0.22	0.47	-0.28	-0.18
35	MC	1	64362	0.61	0.48	0.13		0.61	0.14	0.13	0.12		0.48	-0.21	-0.24	-0.25
36	MC	1	64356	0.72	0.41	0.14		0.18	0.08	0.72	0.02		-0.23	-0.28	0.41	-0.14
37	SA	1	64303	0.41	0.49	0.22	0.59	0.41				-0.49	0.49			
38	MC	1	64238	0.66	0.44	0.32		0.11	0.10	0.66	0.13		-0.23	-0.25	0.44	-0.17
39	MC	1	64234	0.40	0.53	0.33		0.24	0.25	0.11	0.40		-0.29	-0.27	-0.05	0.53
40	MC	1	64337	0.46	0.48	0.17		0.43	0.46	0.07	0.04		-0.36	0.48	-0.13	-0.13
41	MC	1	64323	0.29	0.51	0.19		0.10	0.31	0.30	0.29		-0.21	-0.28	-0.07	0.51
42	SA	1	64291	0.41	0.58	0.24	0.59	0.41				-0.58	0.58			
43	MC	1	64335	0.26	0.14	0.17		0.46	0.18	0.09	0.26		-0.17	0.12	-0.08	0.14
44	SA	1	64307	0.38	0.52	0.21	0.62	0.38				-0.52	0.52			
45	MC	1	64317	0.55	0.39	0.20		0.23	0.54	0.15	0.08		-0.26	0.39	-0.15	-0.10
46	MC	1	64294	0.53	0.41	0.23		0.14	0.18	0.53	0.16		-0.16	-0.19	0.41	-0.20

Table G-9. Item Statistics, Mathematics Grade 5

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Percent Omit	Percent of Students					Item-Total Test Correlation				
							Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4	Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4
1	MC	1	62936	0.81	0.33	0.27		0.80	0.12	0.06	0.02		0.33	-0.19	-0.21	-0.13
2	MC	1	63056	0.46	0.51	0.08		0.46	0.48	0.04	0.02		0.51	-0.39	-0.23	-0.10
3	MC	1	63026	0.59	0.19	0.13		0.59	0.23	0.15	0.03		0.19	0.02	-0.23	-0.11
4	MC	1	63015	0.32	0.39	0.14		0.11	0.37	0.20	0.32		-0.23	-0.14	-0.10	0.39
5	MC	1	63047	0.67	0.41	0.09		0.06	0.67	0.11	0.17		-0.17	0.41	-0.20	-0.24
6	TE	1	62890	0.34	0.59	0.34	0.66	0.34				-0.59	0.59			
7	SA	1	63019	0.78	0.36	0.14	0.22	0.78				-0.35	0.36			
8	MC	1	62905	0.42	0.35	0.32		0.42	0.22	0.11	0.24		0.35	-0.14	-0.21	-0.10
9	MC	1	63029	0.68	0.45	0.12		0.13	0.08	0.11	0.68		-0.31	-0.19	-0.17	0.46
10	MC	1	63015	0.77	0.46	0.14		0.10	0.76	0.07	0.06		-0.28	0.46	-0.24	-0.20
11	MC	1	63007	0.64	0.12	0.16		0.64	0.18	0.13	0.05		0.12	-0.02	-0.10	-0.06
12	TE	1	62788	0.16	0.48	0.50	0.83	0.16				-0.46	0.48			
13	SA	1	62986	0.38	0.55	0.19	0.62	0.38				-0.54	0.55			
14	SA	1	62828	0.43	0.53	0.44	0.57	0.43				-0.52	0.53			
15	MC	1	62907	0.53	0.43	0.32		0.08	0.20	0.53	0.18		-0.21	-0.20	0.43	-0.19
16	SA	1	62898	0.28	0.55	0.33	0.72	0.28				-0.54	0.55			
17	MC	1	62982	0.47	0.32	0.20		0.46	0.25	0.13	0.15		0.32	-0.04	-0.23	-0.18
18	MC	1	62945	0.43	0.47	0.26		0.42	0.12	0.20	0.25		0.47	-0.15	-0.28	-0.15
19	MC	1	62979	0.43	0.46	0.20		0.32	0.16	0.09	0.43		-0.23	-0.20	-0.18	0.46
20	TE	1	62946	0.60	0.46	0.25	0.40	0.60				-0.45	0.46			
21	MC	1	62983	0.53	0.56	0.19		0.25	0.12	0.10	0.53		-0.34	-0.24	-0.17	0.56
22	SA	1	62929	0.45	0.58	0.28	0.54	0.45				-0.57	0.58			
23	MC	1	62893	0.65	0.40	0.34		0.06	0.14	0.64	0.15		-0.23	-0.20	0.41	-0.19

Table G-9. Item Statistics, Mathematics Grade 5 (cont.)

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Percent Omit	Percent of Students					Item-Total Test Correlation				
							Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4	Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4
24	MC	1	62936	0.55	0.32	0.24		0.11	0.20	0.55	0.14		-0.16	-0.09	0.32	-0.21
25	MC	1	63017	0.26	0.17	0.11		0.26	0.47	0.09	0.17		0.18	0.11	-0.12	-0.25
26	ESR	1	63003	0.17	0.46	0.13	0.83	0.17				-0.45	0.46			
27	SA	1	62806	0.15	0.45	0.44	0.85	0.15				-0.44	0.45			
28	SA	1	62958	0.14	0.42	0.20	0.86	0.14				-0.41	0.42			
29	TE	1	62890	0.46	0.49	0.31	0.54	0.46				-0.48	0.49			
30	SA	1	62902	0.54	0.52	0.29	0.46	0.53				-0.52	0.52			
31	MC	1	62912	0.56	0.31	0.28		0.20	0.07	0.56	0.16		-0.16	-0.17	0.31	-0.12
32	MC	1	62984	0.51	0.46	0.16		0.10	0.17	0.50	0.22		-0.20	-0.18	0.46	-0.25
33	MC	1	63000	0.51	0.58	0.14		0.51	0.18	0.11	0.20		0.58	-0.35	-0.25	-0.18
34	SA	1	62951	0.34	0.50	0.21	0.66	0.34				-0.49	0.50			
35	TE	1	62312	0.23	0.50	1.23	0.76	0.23				-0.46	0.50			
36	MC	1	62944	0.51	0.48	0.23		0.51	0.12	0.26	0.11		0.48	-0.29	-0.17	-0.23
37	ESR	1	62965	0.10	0.26	0.19	0.90	0.10				-0.25	0.26			
38	MC	1	62879	0.39	0.25	0.33		0.27	0.39	0.16	0.17		-0.21	0.25	-0.19	0.12
39	MC	1	62804	0.14	0.30	0.45		0.46	0.24	0.16	0.14		-0.12	-0.09	-0.01	0.30
40	MC	1	62967	0.60	0.45	0.19		0.07	0.15	0.18	0.60		-0.20	-0.20	-0.25	0.45
41	MC	1	62967	0.27	0.39	0.19		0.45	0.18	0.27	0.10		-0.35	0.02	0.39	-0.02
42	MC	1	62979	0.43	0.41	0.17		0.35	0.12	0.10	0.43		-0.12	-0.26	-0.20	0.41
43	MC	1	62954	0.55	0.46	0.21		0.25	0.54	0.08	0.12		-0.20	0.46	-0.21	-0.25
44	MC	1	62974	0.80	0.24	0.18		0.08	0.09	0.80	0.03		-0.11	-0.13	0.24	-0.15
45	MC	1	62954	0.65	0.37	0.21		0.05	0.12	0.18	0.65		-0.18	-0.22	-0.17	0.37
46	ESR	1	62931	0.07	0.26	0.25	0.93	0.07				-0.25	0.26			

Table G-10. Item Statistics, Mathematics Grade 6

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Percent Omit	Percent of Students					Item-Total Test Correlation				
							Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4	Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4
1	MC	1	62719	0.51	0.46	0.10		0.19	0.22	0.51	0.08		-0.26	-0.21	0.46	-0.15
2	ESR	1	62710	0.63	0.54	0.11	0.37	0.63				-0.53	0.54			
3	MC	1	62712	0.84	0.38	0.11		0.02	0.04	0.11	0.83		-0.18	-0.16	-0.28	0.38
4	TE	1	62525	0.58	0.51	0.41	0.42	0.57				-0.51	0.52			
5	MC	1	62719	0.55	0.34	0.10		0.24	0.11	0.55	0.09		-0.10	-0.18	0.35	-0.25
6	SA	1	62627	0.39	0.61	0.25	0.61	0.39				-0.61	0.61			
7	MC	1	62719	0.92	0.36	0.10		0.92	0.02	0.02	0.05		0.36	-0.14	-0.15	-0.28
8	MC	1	62695	0.54	0.31	0.14		0.15	0.54	0.19	0.11		-0.11	0.32	-0.19	-0.13
9	TE	1	62564	0.12	0.44	0.35	0.88	0.12				-0.42	0.44			
10	MC	1	62718	0.33	0.15	0.10		0.23	0.33	0.33	0.10		-0.36	0.15	0.25	-0.11
11	SA	1	62697	0.41	0.50	0.13	0.59	0.41				-0.50	0.50			
12	MC	1	62669	0.39	0.27	0.18		0.08	0.39	0.42	0.10		-0.08	0.27	-0.10	-0.20
13	SA	1	62672	0.18	0.48	0.17	0.82	0.18				-0.47	0.48			
14	MC	1	62664	0.49	0.36	0.19		0.17	0.49	0.18	0.15		-0.28	0.36	-0.11	-0.08
15	MC	1	62679	0.65	0.33	0.16		0.65	0.19	0.11	0.05		0.33	-0.22	-0.13	-0.13
16	MC	1	62673	0.66	0.48	0.17		0.07	0.12	0.15	0.66		-0.17	-0.28	-0.25	0.48
17	MC	1	62646	0.77	0.40	0.20		0.02	0.17	0.04	0.77		-0.16	-0.29	-0.18	0.40
18	MC	1	62642	0.69	0.38	0.20		0.06	0.69	0.12	0.12		-0.15	0.38	-0.18	-0.23
19	SA	1	62543	0.71	0.50	0.36	0.28	0.71				-0.49	0.50			
20	MC	1	62681	0.35	0.50	0.14		0.18	0.35	0.31	0.16		-0.29	0.50	-0.06	-0.26
21	MC	1	62660	0.71	0.30	0.17		0.09	0.71	0.13	0.07		-0.02	0.30	-0.19	-0.24
22	MC	1	62634	0.54	0.53	0.22		0.54	0.25	0.09	0.12		0.53	-0.36	-0.20	-0.15
23	MC	1	62619	0.52	0.42	0.24		0.29	0.52	0.15	0.04		-0.28	0.42	-0.17	-0.11

Table G-10. Item Statistics, Mathematics Grade 6 (cont.)

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Percent Omit	Percent of Students					Item-Total Test Correlation				
							Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4	Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4
24	SA	1	62635	0.25	0.61	0.21	0.75	0.25				-0.60	0.61			
25	MC	1	62642	0.91	0.32	0.20		0.05	0.90	0.02	0.03		-0.20	0.32	-0.14	-0.18
26	MC	1	62566	0.52	0.24	0.32		0.12	0.23	0.52	0.13		-0.12	-0.17	0.24	-0.01
27	TE	1	62587	0.20	0.50	0.29	0.80	0.20				-0.49	0.50			
28	MC	1	62542	0.32	0.33	0.36		0.46	0.32	0.12	0.09		-0.19	0.33	-0.21	0.04
29	ESR	1	62498	0.06	0.22	0.43	0.93	0.06				-0.20	0.22			
30	ESR	1	62678	0.41	0.38	0.14	0.58	0.41				-0.37	0.38			
31	MC	1	62648	0.66	0.49	0.19		0.10	0.65	0.11	0.13		-0.14	0.49	-0.26	-0.32
32	MC	1	62630	0.63	0.46	0.22		0.13	0.16	0.63	0.07		-0.24	-0.24	0.46	-0.18
33	SA	1	62448	0.11	0.50	0.51	0.88	0.11				-0.48	0.50			
34	MC	1	62610	0.43	0.47	0.25		0.38	0.08	0.11	0.43		-0.16	-0.26	-0.26	0.47
35	MC	1	62628	0.62	0.48	0.22		0.13	0.09	0.17	0.62		-0.19	-0.23	-0.28	0.48
36	MC	1	62633	0.44	0.45	0.22		0.27	0.20	0.44	0.08		-0.23	-0.12	0.45	-0.25
37	TE	1	62426	0.39	0.50	0.55	0.61	0.38				-0.48	0.50			
38	MC	1	62548	0.39	0.24	0.35		0.15	0.39	0.39	0.06		-0.17	-0.02	0.25	-0.19
39	MC	1	62527	0.40	0.24	0.39		0.12	0.16	0.32	0.40		-0.15	-0.15	-0.02	0.24
40	MC	1	62472	0.35	0.53	0.47		0.27	0.34	0.08	0.30		-0.26	0.53	-0.10	-0.23
41	MC	1	62508	0.34	0.12	0.42		0.24	0.25	0.34	0.16		-0.01	-0.03	0.12	-0.09
42	SA	1	62451	0.24	0.53	0.51	0.76	0.24				-0.52	0.53			
43	MC	1	62597	0.34	0.22	0.27		0.23	0.29	0.34	0.15		-0.24	-0.03	0.22	0.02
44	MC	1	62549	0.37	0.25	0.35		0.26	0.37	0.21	0.16		-0.21	0.25	0.05	-0.13
45	MC	1	62590	0.57	0.45	0.29		0.14	0.14	0.14	0.57		-0.22	-0.22	-0.19	0.45
46	MC	1	62594	0.35	0.49	0.28		0.35	0.13	0.11	0.41		0.49	-0.05	-0.12	-0.36

Table G-11. Item Statistics, Mathematics Grade 7

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Percent Omit	Percent of Students					Item-Total Test Correlation				
							Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4	Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4
1	MC	1	63075	0.23	0.41	0.06		0.18	0.23	0.41	0.18		-0.06	0.41	-0.30	-0.01
2	MC	1	63075	0.46	0.55	0.06		0.31	0.46	0.21	0.02		-0.24	0.55	-0.35	-0.12
3	SA	1	63038	0.57	0.60	0.12	0.43	0.57				-0.60	0.60			
4	TE	1	63045	0.21	0.44	0.11	0.79	0.21				-0.44	0.44			
5	MC	1	63053	0.49	0.42	0.09		0.07	0.49	0.14	0.30		-0.12	0.42	-0.28	-0.18
6	MC	1	63057	0.38	0.44	0.09		0.37	0.08	0.30	0.24		0.44	-0.23	-0.20	-0.13
7	SA	1	62857	0.14	0.49	0.40	0.85	0.14				-0.48	0.49			
8	MC	1	63069	0.49	0.37	0.07		0.27	0.10	0.13	0.49		-0.01	-0.29	-0.27	0.37
9	MC	1	63047	0.52	0.33	0.10		0.52	0.16	0.20	0.12		0.33	-0.14	-0.18	-0.11
10	MC	1	63044	0.50	0.39	0.11		0.50	0.19	0.25	0.06		0.40	-0.33	-0.05	-0.20
11	MC	1	63034	0.33	0.25	0.12		0.33	0.12	0.22	0.33		0.07	-0.24	-0.17	0.25
12	MC	1	62991	0.50	0.42	0.15		0.50	0.16	0.19	0.15		0.42	-0.16	-0.10	-0.31
13	MC	1	62910	0.63	0.48	0.28		0.63	0.11	0.14	0.11		0.48	-0.18	-0.27	-0.25
14	SA	1	62834	0.62	0.53	0.40	0.38	0.61				-0.52	0.53			
15	TE	1	62629	0.15	0.39	0.73	0.84	0.15				-0.37	0.39			
16	MC	1	62888	0.29	0.52	0.32		0.29	0.14	0.24	0.33		0.52	-0.12	-0.06	-0.35
17	MC	1	62917	0.36	0.47	0.27		0.28	0.24	0.12	0.36		-0.22	-0.14	-0.19	0.47
18	MC	1	62950	0.43	0.22	0.22		0.04	0.19	0.42	0.34		-0.09	-0.03	0.22	-0.16
19	SA	1	62638	0.26	0.62	0.71	0.73	0.26				-0.60	0.62			
20	TE	1	62590	0.64	0.18	0.79	0.35	0.64				-0.17	0.19			
21	ESR	1	62812	0.18	0.42	0.44	0.81	0.18				-0.41	0.42			
22	MC	1	62860	0.29	0.22	0.36		0.37	0.29	0.29	0.05		-0.06	-0.13	0.22	-0.05
23	MC	1	62818	0.65	0.33	0.43		0.06	0.65	0.08	0.21		-0.11	0.33	-0.23	-0.16



Table G-11. Item Statistics, Mathematics Grade 7 (cont.)

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Percent Omit	Percent of Students					Item-Total Test Correlation				
							Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4	Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4
24	MC	1	62804	0.78	0.43	0.45		0.05	0.07	0.09	0.78		-0.19	-0.24	-0.25	0.44
25	MC	1	62805	0.34	0.31	0.45		0.28	0.22	0.34	0.16		-0.11	-0.08	0.31	-0.17
26	MC	1	62859	0.65	0.39	0.36		0.06	0.13	0.65	0.17		-0.17	-0.27	0.39	-0.15
27	TE	1	62755	0.09	0.34	0.53	0.90	0.09				-0.31	0.34			
28	MC	1	62854	0.51	0.46	0.37		0.50	0.14	0.15	0.20		0.47	-0.29	-0.25	-0.09
29	ESR	1	62837	0.14	0.35	0.40	0.86	0.13				-0.33	0.35			
30	SA	1	62718	0.26	0.57	0.58	0.73	0.26				-0.55	0.57			
31	MC	1	62845	0.63	0.55	0.38		0.08	0.16	0.13	0.63		-0.13	-0.36	-0.28	0.55
32	MC	1	62830	0.56	0.18	0.41		0.09	0.13	0.21	0.56		-0.04	-0.07	-0.12	0.18
33	MC	1	62821	0.68	0.37	0.42		0.10	0.68	0.14	0.07		-0.19	0.38	-0.23	-0.14
34	MC	1	62780	0.51	0.49	0.49		0.51	0.21	0.21	0.08		0.49	-0.27	-0.16	-0.25
35	SA	1	62721	0.38	0.55	0.58	0.62	0.38				-0.53	0.55			
36	MC	1	62750	0.56	0.34	0.53		0.12	0.55	0.18	0.14		-0.22	0.34	-0.22	-0.02
37	TE	1	62551	0.13	0.51	0.85	0.86	0.13				-0.47	0.51			
38	MC	1	62706	0.53	0.22	0.60		0.03	0.53	0.39	0.04		-0.15	0.22	-0.08	-0.18
39	MC	1	62702	0.30	0.22	0.61		0.17	0.30	0.30	0.22		0.00	0.22	0.05	-0.28
40	TE	1	62675	0.30	0.62	0.65	0.70	0.30				-0.60	0.62			
41	SA	1	62568	0.49	0.24	0.82	0.51	0.48				-0.23	0.24			
42	MC	1	62745	0.37	0.45	0.54		0.36	0.21	0.20	0.21		0.45	-0.17	-0.24	-0.11
43	MC	1	62758	0.54	0.51	0.52		0.11	0.14	0.20	0.54		-0.20	-0.24	-0.25	0.51
44	SA	1	62550	0.33	0.62	0.85	0.66	0.33				-0.60	0.62			
45	MC	1	62783	0.30	0.35	0.48		0.21	0.17	0.32	0.29		0.10	-0.27	-0.20	0.35
46	MC	1	62733	0.44	0.32	0.56		0.12	0.18	0.26	0.44		-0.10	-0.09	-0.19	0.32

Table G-12. Item Statistics, Mathematics Grade 8

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Percent Omit	Percent of Students					Item-Total Test Correlation				
							Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4	Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4
1	MC	1	62019	0.22	0.16	0.11		0.12	0.18	0.22	0.48		0.00	-0.11	0.16	-0.05
2	SA	1	61747	0.38	0.61	0.55	0.61	0.38				-0.60	0.61			
3	MC	1	62012	0.52	0.39	0.13		0.04	0.33	0.11	0.52		-0.17	-0.27	-0.11	0.39
4	MC	1	62012	0.46	0.20	0.13		0.46	0.17	0.24	0.13		0.20	-0.02	-0.15	-0.09
5	MC	1	62014	0.48	0.50	0.12		0.14	0.48	0.19	0.18		0.01	0.50	-0.30	-0.34
6	SA	1	61801	0.07	0.36	0.47	0.92	0.07				-0.34	0.36			
7	MC	1	62017	0.37	0.27	0.12		0.37	0.27	0.14	0.21		0.27	-0.15	-0.13	-0.04
8	MC	1	62036	0.50	0.49	0.09		0.11	0.28	0.50	0.12		-0.27	-0.16	0.49	-0.28
9	MC	1	62004	0.45	0.31	0.14		0.11	0.21	0.23	0.45		-0.01	-0.24	-0.13	0.31
10	MC	1	61979	0.42	0.29	0.18		0.42	0.27	0.19	0.11		0.29	-0.01	-0.23	-0.14
11	SA	1	61623	0.05	0.31	0.75	0.94	0.05				-0.28	0.31			
12	MC	1	62004	0.44	0.37	0.14		0.44	0.32	0.16	0.08		0.37	-0.13	-0.20	-0.18
13	SA	1	61722	0.17	0.50	0.59	0.83	0.17				-0.48	0.50			
14	MC	1	61956	0.37	0.40	0.19		0.15	0.37	0.05	0.43		-0.21	0.40	-0.14	-0.17
15	MC	1	61896	0.63	0.28	0.28		0.15	0.13	0.63	0.09		-0.09	-0.17	0.28	-0.15
16	TE	1	61766	0.49	0.50	0.49	0.51	0.49				-0.49	0.50			
17	MC	1	61885	0.59	0.46	0.30		0.11	0.59	0.15	0.14		-0.20	0.46	-0.22	-0.23
18	SA	1	61791	0.10	0.40	0.45	0.90	0.10				-0.38	0.40			
19	SA	1	61643	0.46	0.59	0.69	0.53	0.46				-0.58	0.59			
20	MC	1	61901	0.56	0.38	0.27		0.56	0.22	0.15	0.06		0.39	-0.16	-0.24	-0.15
21	TE	1	61771	0.10	0.40	0.48	0.89	0.10				-0.38	0.40			
22	MC	1	61898	0.41	0.43	0.28		0.21	0.24	0.13	0.41		-0.13	-0.15	-0.27	0.43
23	MC	1	61911	0.35	0.26	0.26		0.09	0.35	0.16	0.39		-0.11	0.26	-0.25	0.00

Table G-12. Item Statistics, Mathematics Grade 8 (cont.)

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Percent Omit	Percent of Students					Item-Total Test Correlation				
							Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4	Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4
24	MC	1	61837	0.65	0.41	0.38		0.09	0.65	0.11	0.14		-0.13	0.41	-0.26	-0.21
25	MC	1	61880	0.23	0.17	0.31		0.10	0.35	0.23	0.31		-0.20	0.02	0.17	-0.03
26	SA	1	61095	0.20	0.48	1.57	0.79	0.20				-0.44	0.48			
27	TE	1	61295	0.57	0.50	1.25	0.43	0.56				-0.48	0.51			
28	MC	1	61805	0.55	0.57	0.43		0.07	0.14	0.23	0.55		-0.22	-0.25	-0.32	0.57
29	MC	1	61880	0.48	0.24	0.31		0.20	0.48	0.22	0.09		-0.06	0.24	-0.07	-0.23
30	ESR	1	61881	0.31	0.38	0.31	0.69	0.31				-0.37	0.38			
31	SA	1	61323	0.30	0.58	1.21	0.69	0.29				-0.55	0.58			
32	TE	1	61512	0.34	0.49	0.90	0.65	0.34				-0.47	0.49			
33	MC	1	61847	0.66	0.31	0.36		0.25	0.05	0.65	0.04		-0.13	-0.23	0.32	-0.20
34	MC	1	61814	0.51	0.31	0.41		0.33	0.51	0.10	0.06		-0.09	0.32	-0.25	-0.17
35	MC	1	61784	0.73	0.46	0.46		0.07	0.72	0.13	0.08		-0.24	0.47	-0.24	-0.23
36	MC	1	61744	0.55	0.46	0.53		0.14	0.55	0.21	0.09		-0.13	0.47	-0.33	-0.17
37	TE	1	61490	0.35	0.37	0.94	0.65	0.34				-0.35	0.37			
38	MC	1	61730	0.76	0.44	0.55		0.11	0.05	0.76	0.07		-0.22	-0.22	0.45	-0.26
39	MC	1	61673	0.40	0.36	0.64		0.39	0.18	0.25	0.17		0.36	-0.25	-0.12	-0.06
40	ESR	1	61732	0.17	0.43	0.55	0.83	0.17				-0.41	0.43			
41	MC	1	61717	0.36	0.28	0.57		0.36	0.17	0.23	0.24		0.28	-0.14	-0.20	0.02
42	SA	1	61393	0.52	0.43	1.09	0.47	0.51				-0.41	0.44			
43	MC	1	61765	0.66	0.54	0.49		0.66	0.12	0.18	0.04		0.54	-0.28	-0.33	-0.18
44	SA	1	61263	0.23	0.56	1.30	0.76	0.23				-0.53	0.56			
45	MC	1	61803	0.66	0.39	0.43		0.66	0.11	0.13	0.10		0.40	-0.19	-0.27	-0.11
46	MC	1	61785	0.57	0.48	0.46		0.56	0.21	0.14	0.08		0.48	-0.16	-0.31	-0.21

Table G-13. Item Statistics, Science Grade 4

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Percent Omit	Percent of Students					Item-Total Test Correlation				
							Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4	Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4
1	MC	1	64375	0.76	0.39	0.13		0.76	0.12	0.05	0.07		0.39	-0.26	-0.21	-0.14
2	MC	1	64413	0.87	0.46	0.07		0.07	0.04	0.03	0.87		-0.26	-0.27	-0.24	0.46
3	MC	1	64418	0.92	0.18	0.07		0.04	0.92	0.02	0.02		-0.13	0.19	-0.06	-0.11
4	MC	1	64416	0.94	0.32	0.07		0.03	0.01	0.01	0.94		-0.20	-0.16	-0.18	0.32
5	MC	1	64404	0.57	0.41	0.09		0.02	0.12	0.28	0.57		-0.20	-0.24	-0.21	0.41
6	MC	1	64383	0.79	0.44	0.12		0.10	0.07	0.79	0.04		-0.26	-0.23	0.44	-0.20
7	MC	1	64402	0.78	0.47	0.09		0.07	0.10	0.78	0.04		-0.28	-0.28	0.47	-0.19
8	MC	1	64374	0.73	0.46	0.13		0.11	0.05	0.11	0.73		-0.22	-0.20	-0.30	0.46
9	MC	1	64202	0.35	0.21	0.40		0.25	0.13	0.35	0.27		0.02	-0.18	0.21	-0.11
10	MC	1	64367	0.66	0.50	0.14		0.12	0.09	0.65	0.13		-0.24	-0.24	0.50	-0.26
11	TE	1	64140	0.43	0.24	0.50	0.57	0.43				-0.23	0.24			
12	MC	1	64382	0.61	0.49	0.12		0.22	0.10	0.07	0.61		-0.21	-0.28	-0.26	0.49
13	MC	1	64406	0.89	0.47	0.08		0.04	0.04	0.04	0.89		-0.26	-0.26	-0.25	0.47
14	MC	1	64391	0.61	0.37	0.11		0.19	0.09	0.61	0.11		-0.13	-0.27	0.37	-0.17
15	MC	1	64413	0.87	0.28	0.07		0.87	0.03	0.02	0.07		0.28	-0.13	-0.12	-0.20
16	MC	1	64405	0.79	0.38	0.09		0.05	0.11	0.05	0.79		-0.22	-0.17	-0.24	0.38
17	MC	1	64312	0.79	0.45	0.23		0.10	0.79	0.03	0.08		-0.28	0.45	-0.15	-0.26
18	MC	1	64367	0.58	0.37	0.14		0.18	0.06	0.57	0.19		-0.12	-0.23	0.37	-0.21
19	MC	1	64391	0.77	0.29	0.11		0.77	0.05	0.08	0.10		0.29	-0.24	-0.18	-0.08
20	MC	1	64384	0.53	0.42	0.12		0.13	0.10	0.24	0.53		-0.22	-0.25	-0.14	0.42

Table G-13. Item Statistics, Science Grade 4 (cont.)

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Percent Omit	Percent of Students					Item-Total Test Correlation				
							Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4	Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4
21	MC	1	64384	0.85	0.37	0.12		0.08	0.85	0.03	0.04		-0.13	0.37	-0.22	-0.28
22	MC	1	64375	0.91	0.25	0.10		0.91	0.01	0.05	0.03		0.25	-0.15	-0.15	-0.13
23	MC	1	64392	0.84	0.42	0.07		0.84	0.06	0.07	0.04		0.42	-0.26	-0.23	-0.19
24	MC	1	64391	0.63	0.35	0.07		0.13	0.14	0.10	0.63		-0.33	-0.08	-0.11	0.35
25	MC	1	64372	0.47	0.16	0.10		0.05	0.13	0.47	0.35		-0.23	-0.20	0.16	0.08
26	MC	1	64360	0.60	0.34	0.12		0.10	0.10	0.60	0.20		-0.22	-0.20	0.34	-0.10
27	MC	1	64323	0.64	0.39	0.18		0.22	0.06	0.08	0.64		-0.17	-0.21	-0.25	0.39
28	MC	1	64389	0.56	0.29	0.08		0.03	0.24	0.17	0.56		-0.11	-0.11	-0.20	0.29
29	MC	1	64351	0.76	0.31	0.14		0.11	0.75	0.02	0.12		-0.09	0.31	-0.19	-0.24
30	MC	1	64250	0.67	0.41	0.29		0.67	0.10	0.14	0.09		0.41	-0.15	-0.23	-0.23
31	MC	1	64351	0.71	0.36	0.14		0.71	0.10	0.08	0.11		0.36	-0.20	-0.15	-0.21
32	MC	1	64351	0.75	0.49	0.14		0.13	0.05	0.07	0.75		-0.26	-0.25	-0.27	0.49
33	MC	1	64336	0.71	0.48	0.16		0.11	0.08	0.10	0.71		-0.22	-0.24	-0.28	0.48
34	MC	1	64371	0.78	0.45	0.10		0.78	0.10	0.05	0.07		0.45	-0.24	-0.26	-0.22
35	MC	1	64352	0.82	0.46	0.13		0.04	0.05	0.09	0.82		-0.26	-0.25	-0.24	0.46
36	MC	1	64378	0.41	0.20	0.09		0.23	0.11	0.41	0.25		-0.06	-0.13	0.20	-0.08
37	MC	1	64375	0.69	0.35	0.10		0.11	0.17	0.69	0.03		-0.18	-0.20	0.35	-0.18
38	MC	1	64284	0.44	0.33	0.24		0.12	0.26	0.19	0.43		-0.19	-0.21	-0.02	0.33
39	MC	1	64321	0.56	0.45	0.18		0.16	0.15	0.56	0.13		-0.21	-0.28	0.45	-0.13
40	MC	1	64357	0.77	0.47	0.13		0.12	0.06	0.05	0.77		-0.23	-0.27	-0.26	0.47

Table G-14. Item Statistics, Science Grade 8

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Percent Omit	Percent of Students				Item-Total Test Correlation					
							Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4	Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4
1	MC	1	62020	0.86	0.41	0.05		0.04	0.03	0.86	0.08		-0.20	-0.25	0.41	-0.25
2	MC	1	62012	0.87	0.40	0.06		0.03	0.87	0.05	0.05		-0.19	0.40	-0.21	-0.25
3	MC	1	62001	0.91	0.40	0.08		0.02	0.91	0.04	0.03		-0.18	0.40	-0.21	-0.27
4	MC	1	61993	0.78	0.38	0.09		0.04	0.06	0.12	0.78		-0.22	-0.25	-0.17	0.38
5	MC	1	62000	0.86	0.30	0.08		0.04	0.07	0.03	0.86		-0.16	-0.20	-0.13	0.30
6	MC	1	61999	0.85	0.40	0.08		0.02	0.85	0.08	0.05		-0.15	0.40	-0.24	-0.25
7	MC	1	62008	0.77	0.44	0.07		0.77	0.08	0.12	0.03		0.44	-0.29	-0.23	-0.18
8	MC	1	61967	0.84	0.42	0.13		0.04	0.84	0.06	0.06		-0.28	0.42	-0.28	-0.13
9	MC	1	61795	0.62	0.45	0.41		0.62	0.21	0.12	0.04		0.45	-0.23	-0.28	-0.13
10	MC	1	61922	0.45	0.19	0.20		0.19	0.08	0.28	0.45		-0.08	-0.23	0.00	0.19
11	MC	1	61935	0.66	0.43	0.18		0.66	0.13	0.13	0.09		0.43	-0.22	-0.20	-0.23
12	MC	1	61922	0.45	0.24	0.20		0.15	0.21	0.45	0.19		-0.10	-0.11	0.24	-0.09
13	MC	1	62005	0.76	0.29	0.07		0.12	0.09	0.76	0.02		-0.22	-0.09	0.30	-0.18
14	MC	1	61977	0.91	0.33	0.12		0.02	0.03	0.91	0.04		-0.16	-0.26	0.34	-0.15
15	MC	1	61992	0.75	0.35	0.09		0.07	0.09	0.09	0.75		-0.14	-0.14	-0.27	0.35
16	MC	1	62009	0.75	0.36	0.06		0.09	0.75	0.07	0.09		-0.19	0.36	-0.26	-0.11
17	MC	1	61938	0.62	0.29	0.18		0.62	0.21	0.07	0.10		0.29	-0.20	-0.14	-0.07
18	MC	1	61961	0.67	0.42	0.14		0.07	0.16	0.10	0.66		-0.24	-0.19	-0.21	0.42
19	MC	1	61965	0.86	0.51	0.14		0.86	0.05	0.05	0.04		0.51	-0.29	-0.30	-0.24
20	MC	1	61955	0.66	0.45	0.15		0.05	0.04	0.25	0.66		-0.24	-0.30	-0.23	0.45

Table G-14. Item Statistics, Science Grade 8 (cont.)

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Percent Omit	Percent of Students					Item-Total Test Correlation				
							Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4	Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4
21	MC	1	61971	0.77	0.35	0.13		0.16	0.77	0.03	0.03		-0.16	0.35	-0.25	-0.22
22	MC	1	61912	0.71	0.36	0.12		0.71	0.04	0.05	0.20		0.36	-0.21	-0.17	-0.21
23	MC	1	61878	0.74	0.48	0.18		0.74	0.13	0.09	0.05		0.48	-0.26	-0.32	-0.17
24	MC	1	61898	0.54	0.24	0.15		0.06	0.15	0.54	0.24		-0.04	-0.08	0.24	-0.18
25	MC	1	61884	0.54	0.27	0.17		0.18	0.10	0.54	0.17		-0.06	-0.25	0.27	-0.09
26	MC	1	61897	0.78	0.50	0.15		0.06	0.08	0.08	0.78		-0.25	-0.27	-0.27	0.50
27	MC	1	61895	0.62	0.33	0.15		0.06	0.09	0.23	0.61		-0.11	-0.23	-0.15	0.33
28	MC	1	61916	0.64	0.46	0.12		0.64	0.11	0.11	0.15		0.46	-0.26	-0.22	-0.20
29	MC	1	61883	0.71	0.45	0.17		0.06	0.13	0.11	0.71		-0.27	-0.24	-0.19	0.45
30	MC	1	61766	0.66	0.51	0.36		0.66	0.16	0.14	0.04		0.51	-0.21	-0.33	-0.25
31	MC	1	61808	0.49	0.32	0.29		0.12	0.21	0.49	0.18		-0.16	-0.19	0.32	-0.08
32	MC	1	61839	0.29	0.17	0.24		0.24	0.22	0.29	0.25		-0.10	-0.07	0.17	-0.01
33	MC	1	61844	0.36	0.23	0.23		0.14	0.14	0.36	0.36		-0.13	-0.21	0.02	0.23
34	MC	1	61878	0.85	0.46	0.18		0.04	0.85	0.08	0.03		-0.26	0.46	-0.28	-0.21
35	MC	1	61878	0.72	0.42	0.18		0.12	0.72	0.10	0.06		-0.13	0.42	-0.29	-0.24
36	MC	1	61876	0.66	0.49	0.18		0.66	0.12	0.10	0.12		0.49	-0.33	-0.28	-0.12
37	MC	1	61881	0.84	0.52	0.17		0.05	0.06	0.05	0.84		-0.27	-0.28	-0.29	0.52
38	MC	1	61820	0.67	0.36	0.27		0.03	0.24	0.06	0.67		-0.21	-0.15	-0.29	0.36
39	MC	1	61858	0.70	0.46	0.21		0.11	0.15	0.70	0.04		-0.28	-0.22	0.46	-0.22
40	MC	1	61861	0.74	0.45	0.21		0.09	0.08	0.74	0.09		-0.18	-0.26	0.46	-0.27

Table G-15. Item Statistics, Social Studies Grade 4

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Percent Omit	Percent of Students				Item-Total Test Correlation					
							Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4	Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4
1	MC	1	64425	0.89	0.34	0.04		0.03	0.02	0.89	0.05		-0.18	-0.17	0.34	-0.22
2	MC	1	64365	0.81	0.44	0.13		0.10	0.81	0.04	0.04		-0.30	0.44	-0.18	-0.23
3	MC	1	64413	0.86	0.33	0.06		0.03	0.06	0.86	0.05		-0.16	-0.20	0.33	-0.19
4	MC	1	64353	0.57	0.40	0.15		0.05	0.29	0.57	0.09		-0.20	-0.29	0.40	-0.08
5	MC	1	64352	0.52	0.39	0.15		0.52	0.18	0.10	0.20		0.39	-0.15	-0.21	-0.19
6	MC	1	64361	0.75	0.38	0.14		0.15	0.75	0.07	0.03		-0.22	0.39	-0.20	-0.23
7	MC	1	64391	0.84	0.42	0.09		0.08	0.01	0.84	0.07		-0.22	-0.17	0.42	-0.30
8	MC	1	64255	0.84	0.35	0.30		0.03	0.07	0.05	0.84		-0.21	-0.14	-0.23	0.35
9	MC	1	64398	0.81	0.45	0.08		0.05	0.81	0.10	0.04		-0.26	0.45	-0.24	-0.24
10	MC	1	64382	0.69	0.41	0.11		0.11	0.10	0.10	0.69		-0.19	-0.24	-0.20	0.42
11	MC	1	64392	0.79	0.45	0.09		0.06	0.10	0.79	0.06		-0.26	-0.25	0.45	-0.21
12	MC	1	64373	0.44	0.24	0.12		0.45	0.06	0.44	0.05		-0.02	-0.23	0.24	-0.24
13	MC	1	64391	0.55	0.25	0.09		0.16	0.55	0.12	0.17		-0.09	0.25	-0.13	-0.13
14	MC	1	64377	0.42	0.23	0.11		0.42	0.30	0.18	0.10		0.23	0.01	-0.21	-0.12
15	MC	1	64383	0.51	0.36	0.10		0.20	0.15	0.51	0.14		-0.10	-0.19	0.36	-0.21
16	MC	1	64277	0.74	0.45	0.27		0.74	0.08	0.11	0.06		0.45	-0.23	-0.23	-0.24
17	MC	1	64352	0.78	0.50	0.15		0.09	0.78	0.05	0.08		-0.21	0.50	-0.30	-0.31
18	MC	1	64390	0.86	0.46	0.09		0.07	0.86	0.03	0.04		-0.27	0.46	-0.24	-0.27
19	MC	1	64388	0.74	0.31	0.10		0.74	0.07	0.13	0.06		0.31	-0.08	-0.20	-0.19
20	MC	1	64381	0.81	0.31	0.05		0.81	0.03	0.14	0.02		0.31	-0.18	-0.18	-0.21



Table G-15. Item Statistics, Social Studies Grade 4 (cont.)

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Percent Omit	Percent of Students					Item-Total Test Correlation				
							Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4	Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4
21	MC	1	64357	0.83	0.45	0.09		0.06	0.83	0.03	0.08		-0.26	0.45	-0.25	-0.24
22	MC	1	64353	0.59	0.36	0.10		0.21	0.15	0.59	0.06		-0.11	-0.21	0.36	-0.24
23	MC	1	64349	0.73	0.46	0.10		0.12	0.07	0.73	0.08		-0.23	-0.28	0.46	-0.21
24	MC	1	64339	0.54	0.36	0.12		0.54	0.09	0.22	0.15		0.36	-0.23	-0.12	-0.19
25	MC	1	64167	0.59	0.39	0.39		0.11	0.59	0.11	0.18		-0.16	0.39	-0.25	-0.15
26	MC	1	64326	0.56	0.46	0.14		0.10	0.24	0.10	0.55		-0.25	-0.19	-0.24	0.46
27	MC	1	64306	0.81	0.42	0.17		0.07	0.06	0.06	0.81		-0.14	-0.29	-0.26	0.42
28	MC	1	64342	0.68	0.19	0.11		0.68	0.17	0.07	0.08		0.19	-0.03	-0.16	-0.12
29	MC	1	64195	0.85	0.44	0.34		0.85	0.06	0.05	0.04		0.43	-0.21	-0.26	-0.24
30	MC	1	64326	0.70	0.44	0.14		0.12	0.06	0.70	0.11		-0.24	-0.20	0.44	-0.23
31	MC	1	64324	0.64	0.46	0.14		0.64	0.13	0.13	0.10		0.46	-0.28	-0.20	-0.20
32	MC	1	64342	0.64	0.45	0.11		0.07	0.64	0.18	0.10		-0.28	0.45	-0.19	-0.23
33	MC	1	64320	0.67	0.40	0.15		0.09	0.67	0.10	0.14		-0.22	0.40	-0.26	-0.13
34	MC	1	64292	0.48	0.39	0.19		0.16	0.13	0.23	0.48		-0.17	-0.27	-0.10	0.39
35	MC	1	64293	0.50	0.39	0.19		0.24	0.12	0.14	0.50		-0.20	-0.11	-0.21	0.39
36	MC	1	64329	0.82	0.46	0.13		0.82	0.03	0.07	0.07		0.46	-0.23	-0.25	-0.27
37	MC	1	64333	0.54	0.34	0.13		0.18	0.12	0.16	0.54		-0.12	-0.23	-0.14	0.34
38	MC	1	64334	0.82	0.47	0.13		0.10	0.04	0.05	0.82		-0.25	-0.26	-0.28	0.47

Table G-16. Item Statistics, Social Studies Grade 8

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Percent Omit	Percent of Students					Item-Total Test Correlation				
							Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4	Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4
1	MC	1	61958	0.84	0.47	0.09		0.06	0.03	0.07	0.84		-0.24	-0.24	-0.29	0.47
2	MC	1	61947	0.78	0.40	0.10		0.14	0.05	0.78	0.03		-0.22	-0.21	0.40	-0.23
3	MC	1	61950	0.82	0.44	0.10		0.12	0.04	0.82	0.02		-0.27	-0.27	0.44	-0.20
4	MC	1	61948	0.82	0.41	0.10		0.05	0.08	0.05	0.81		-0.19	-0.25	-0.22	0.41
5	MC	1	61923	0.85	0.47	0.14		0.03	0.85	0.07	0.05		-0.19	0.47	-0.32	-0.25
6	MC	1	61941	0.75	0.28	0.11		0.75	0.10	0.11	0.03		0.28	-0.21	-0.07	-0.20
7	MC	1	61931	0.91	0.38	0.13		0.06	0.02	0.01	0.91		-0.27	-0.19	-0.19	0.38
8	MC	1	61919	0.78	0.42	0.15		0.08	0.08	0.78	0.05		-0.21	-0.21	0.42	-0.26
9	MC	1	61920	0.56	0.36	0.15		0.26	0.56	0.09	0.09		-0.15	0.36	-0.28	-0.11
10	MC	1	61948	0.74	0.47	0.10		0.05	0.74	0.16	0.04		-0.19	0.48	-0.32	-0.24
11	MC	1	61941	0.63	0.30	0.11		0.21	0.63	0.14	0.03		-0.05	0.30	-0.28	-0.15
12	MC	1	61920	0.67	0.51	0.15		0.66	0.12	0.12	0.10		0.52	-0.28	-0.31	-0.17
13	MC	1	61892	0.81	0.50	0.19		0.80	0.03	0.12	0.04		0.50	-0.21	-0.33	-0.25
14	MC	1	61902	0.69	0.42	0.18		0.12	0.12	0.69	0.06		-0.16	-0.26	0.42	-0.22
15	MC	1	61894	0.73	0.43	0.19		0.08	0.09	0.09	0.73		-0.18	-0.24	-0.24	0.43
16	MC	1	61866	0.61	0.56	0.24		0.61	0.14	0.15	0.10		0.56	-0.31	-0.30	-0.18
17	MC	1	61916	0.49	0.34	0.15		0.10	0.49	0.19	0.21		-0.22	0.34	-0.27	0.02
18	MC	1	61921	0.66	0.42	0.15		0.15	0.11	0.08	0.66		-0.22	-0.22	-0.20	0.43
19	MC	1	61926	0.58	0.46	0.14		0.27	0.08	0.07	0.58		-0.21	-0.24	-0.27	0.47
20	MC	1	61897	0.65	0.50	0.19		0.65	0.13	0.12	0.10		0.50	-0.23	-0.26	-0.25

Table G-16. Item Statistics, Social Studies Grade 8 (cont.)

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Percent Omit	Percent of Students					Item-Total Test Correlation				
							Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4	Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4
21	MC	1	61832	0.62	0.38	0.15		0.62	0.09	0.09	0.20		0.38	-0.26	-0.30	-0.05
22	MC	1	61845	0.45	0.36	0.13		0.11	0.11	0.45	0.32		0.04	-0.24	0.36	-0.24
23	MC	1	61831	0.41	0.26	0.16		0.23	0.12	0.41	0.24		-0.06	-0.16	0.26	-0.11
24	MC	1	61856	0.67	0.56	0.11		0.08	0.67	0.16	0.09		-0.22	0.56	-0.37	-0.23
25	MC	1	61847	0.74	0.47	0.13		0.05	0.16	0.74	0.06		-0.29	-0.25	0.47	-0.23
26	MC	1	61827	0.70	0.37	0.16		0.06	0.11	0.69	0.13		-0.20	-0.25	0.37	-0.12
27	MC	1	61843	0.85	0.45	0.14		0.02	0.85	0.07	0.05		-0.20	0.45	-0.27	-0.27
28	MC	1	61828	0.66	0.40	0.16		0.12	0.09	0.12	0.66		-0.14	-0.25	-0.21	0.40
29	MC	1	61820	0.59	0.40	0.17		0.11	0.17	0.59	0.12		-0.16	-0.24	0.40	-0.16
30	MC	1	61791	0.59	0.42	0.22		0.07	0.59	0.18	0.15		-0.19	0.42	-0.21	-0.20
31	MC	1	61823	0.76	0.34	0.17		0.76	0.04	0.05	0.15		0.34	-0.22	-0.27	-0.11
32	MC	1	61827	0.52	0.44	0.16		0.52	0.22	0.13	0.13		0.44	-0.22	-0.28	-0.10
33	MC	1	61771	0.71	0.39	0.25		0.06	0.16	0.71	0.07		-0.23	-0.14	0.39	-0.28
34	MC	1	61816	0.56	0.33	0.18		0.56	0.08	0.18	0.17		0.34	-0.29	-0.15	-0.07
35	MC	1	61810	0.63	0.48	0.19		0.12	0.63	0.15	0.09		-0.08	0.48	-0.33	-0.29
36	MC	1	61816	0.69	0.55	0.18		0.16	0.08	0.07	0.69		-0.24	-0.31	-0.33	0.56
37	MC	1	61811	0.71	0.49	0.19		0.11	0.10	0.07	0.71		-0.21	-0.23	-0.31	0.49
38	MC	1	61808	0.59	0.42	0.19		0.09	0.59	0.17	0.14		-0.11	0.42	-0.28	-0.18
39	MC	1	61821	0.51	0.32	0.17		0.26	0.51	0.13	0.09		-0.03	0.32	-0.27	-0.19
40	MC	1	61804	0.81	0.49	0.20		0.81	0.07	0.06	0.06		0.49	-0.29	-0.27	-0.23

Table G-17. Item Statistics, Social Studies Grade 10

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Percent Omit	Percent of Students				Item-Total Test Correlation					
							Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4	Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4
1	MC	1	63519	0.85	0.35	0.08		0.01	0.07	0.85	0.06		-0.14	-0.23	0.35	-0.19
2	MC	1	63461	0.80	0.43	0.18		0.06	0.06	0.80	0.09		-0.22	-0.26	0.43	-0.22
3	MC	1	63426	0.67	0.39	0.23		0.07	0.14	0.12	0.67		-0.14	-0.16	-0.28	0.39
4	MC	1	63492	0.78	0.35	0.13		0.78	0.09	0.04	0.09		0.35	-0.22	-0.20	-0.15
5	MC	1	63512	0.55	0.39	0.10		0.09	0.26	0.55	0.10		-0.17	-0.18	0.39	-0.21
6	MC	1	63513	0.86	0.44	0.09		0.05	0.04	0.05	0.86		-0.21	-0.26	-0.25	0.44
7	MC	1	63490	0.77	0.41	0.13		0.06	0.77	0.06	0.11		-0.13	0.41	-0.28	-0.23
8	MC	1	63462	0.66	0.33	0.17		0.16	0.07	0.10	0.66		-0.15	-0.17	-0.18	0.33
9	MC	1	63479	0.67	0.30	0.15		0.21	0.67	0.05	0.07		-0.08	0.30	-0.21	-0.22
10	MC	1	63470	0.71	0.30	0.16		0.11	0.11	0.07	0.71		-0.12	-0.21	-0.12	0.30
11	MC	1	63447	0.64	0.40	0.20		0.64	0.18	0.12	0.06		0.40	-0.21	-0.23	-0.14
12	MC	1	63379	0.50	0.30	0.31		0.19	0.50	0.21	0.10		-0.14	0.30	-0.17	-0.08
13	MC	1	63369	0.58	0.38	0.32		0.14	0.58	0.24	0.04		-0.19	0.38	-0.17	-0.23
14	MC	1	63446	0.65	0.37	0.20		0.05	0.06	0.65	0.24		-0.19	-0.28	0.37	-0.16
15	MC	1	63454	0.75	0.33	0.19		0.04	0.75	0.10	0.11		-0.22	0.34	-0.15	-0.17
16	MC	1	63421	0.55	0.29	0.24		0.20	0.55	0.11	0.14		-0.04	0.29	-0.25	-0.14
17	MC	1	63419	0.55	0.39	0.24		0.55	0.14	0.20	0.11		0.40	-0.16	-0.22	-0.15
18	MC	1	63399	0.23	0.13	0.27		0.22	0.26	0.29	0.23		0.02	-0.03	-0.10	0.13
19	MC	1	63391	0.73	0.49	0.29		0.07	0.14	0.73	0.05		-0.22	-0.30	0.49	-0.23
20	MC	1	63379	0.67	0.40	0.31		0.17	0.06	0.67	0.09		-0.13	-0.26	0.40	-0.24

Table G-17. Item Statistics, Social Studies Grade 10 (cont.)

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Percent Omit	Percent of Students					Item-Total Test Correlation				
							Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4	Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4
21	MC	1	63379	0.27	0.33	0.31		0.31	0.24	0.18	0.27		-0.01	-0.11	-0.23	0.33
22	MC	1	63363	0.48	0.26	0.33		0.33	0.48	0.12	0.07		-0.02	0.26	-0.24	-0.15
23	MC	1	63350	0.66	0.51	0.35		0.11	0.10	0.13	0.66		-0.20	-0.32	-0.23	0.51
24	MC	1	63345	0.52	0.44	0.36		0.52	0.11	0.20	0.17		0.44	-0.25	-0.13	-0.21
25	MC	1	63324	0.60	0.53	0.39		0.60	0.12	0.17	0.11		0.53	-0.29	-0.24	-0.23
26	MC	1	63042	0.77	0.47	0.19		0.11	0.07	0.77	0.05		-0.25	-0.32	0.47	-0.17
27	MC	1	62983	0.67	0.47	0.28		0.67	0.16	0.09	0.08		0.47	-0.27	-0.31	-0.12
28	MC	1	62995	0.85	0.35	0.26		0.03	0.85	0.09	0.03		-0.19	0.35	-0.24	-0.13
29	MC	1	62970	0.70	0.47	0.30		0.69	0.12	0.11	0.07		0.48	-0.24	-0.27	-0.20
30	MC	1	62949	0.63	0.46	0.34		0.63	0.14	0.10	0.12		0.46	-0.20	-0.26	-0.21
31	MC	1	62994	0.45	0.39	0.27		0.45	0.07	0.38	0.10		0.39	-0.29	-0.15	-0.15
32	MC	1	62983	0.69	0.49	0.28		0.12	0.11	0.08	0.69		-0.18	-0.28	-0.29	0.49
33	MC	1	62975	0.67	0.50	0.30		0.04	0.67	0.15	0.14		-0.21	0.50	-0.27	-0.28
34	MC	1	62913	0.81	0.52	0.39		0.06	0.09	0.81	0.05		-0.27	-0.32	0.52	-0.24
35	MC	1	62942	0.80	0.52	0.35		0.04	0.08	0.80	0.08		-0.24	-0.26	0.52	-0.32
36	MC	1	62933	0.73	0.34	0.36		0.07	0.72	0.09	0.11		-0.23	0.34	-0.12	-0.18
37	MC	1	62898	0.51	0.41	0.42		0.50	0.21	0.15	0.13		0.41	-0.15	-0.24	-0.15
38	MC	1	62901	0.66	0.43	0.41		0.03	0.11	0.20	0.66		-0.21	-0.17	-0.28	0.44
39	MC	1	62928	0.67	0.50	0.37		0.17	0.09	0.08	0.67		-0.24	-0.24	-0.28	0.51
40	MC	1	62908	0.53	0.34	0.40		0.14	0.22	0.53	0.11		-0.11	-0.17	0.34	-0.18

Table G-17. Item Statistics, Social Studies Grade 10 (cont.)

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Percent Omit	Percent of Students					Item-Total Test Correlation				
							Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4	Score Point 0	Score Point 1 or Option 1	Score Point 2 or Option 2	Score Point 3 or Option 3	Score Point 4 or Option 4
41	MC	1	62800	0.56	0.42	0.57		0.17	0.55	0.20	0.08		-0.12	0.42	-0.25	-0.22
42	MC	1	62858	0.51	0.25	0.48		0.06	0.50	0.36	0.07		-0.16	0.26	-0.05	-0.23
43	MC	1	62811	0.58	0.46	0.56		0.58	0.17	0.15	0.10		0.46	-0.25	-0.20	-0.19
44	MC	1	62811	0.53	0.35	0.56		0.19	0.11	0.52	0.17		-0.15	-0.19	0.35	-0.14
45	MC	1	62870	0.67	0.47	0.46		0.14	0.10	0.08	0.66		-0.14	-0.25	-0.33	0.48
46	MC	1	62850	0.71	0.43	0.49		0.71	0.09	0.13	0.06		0.43	-0.26	-0.20	-0.18
47	MC	1	62869	0.45	0.24	0.46		0.31	0.44	0.16	0.08		0.02	0.24	-0.21	-0.17
48	MC	1	62872	0.68	0.48	0.46		0.09	0.68	0.12	0.11		-0.23	0.48	-0.28	-0.20
49	MC	1	62869	0.71	0.44	0.46		0.12	0.10	0.06	0.71		-0.31	-0.16	-0.18	0.44
50	MC	1	62850	0.80	0.38	0.49		0.09	0.05	0.06	0.79		-0.14	-0.19	-0.28	0.38

**Appendix H**  
**Wisconsin Standard Performance Index Score Computation**

## Technical Details of Wisconsin Standard Performance Index Score Computation

Technical details of the Standard Performance Index (SPI) estimation procedure described in this Appendix are based on description of the SPI computation methodology included in the *TerraNova 2nd Edition Technical Report* (CTB/McGraw-Hill, 2000).

The Standard Performance Index (SPI) is an estimate of the true score (estimated proportion of total, or maximum, points possible) for a content standard based on the performance of a given student. Because most standards are measured by a relatively small number of items, a Bayesian procedure that takes into account the overall test performance is used to improve the reliability of the standard scores. Given a student's scale score on the test, item response theory (IRT) is used, via the 3-parameter logistic (3PL) model for MC items and the 2-parameter-partial credit (2PPC) model for CR items, to compute the estimated proportion of the maximum points obtained for that standard.

The estimated proportion of the maximum points obtained for the standard provides the initial (Bayesian prior) estimate of the student's mastery score. If this initial estimate is consistent with the student's observed proportion, as indicated by a chi-square test, the two scores are combined as a weighted average to obtain the SPI score (the estimated true score). The appropriate weight for the Bayesian prior estimate is computed as a function of the standard error (SE) of the scale score on which it is based: the smaller the SE, the larger the weight. If the prior estimate and the observed proportion differ significantly, the observed proportion of the maximum score is used without the prior estimate to compute the student's score on that objective.

### Standard Performance Index Computation

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 strand. Assume a  $k$ -item test is composed of  $j$  strands with a maximum possible raw score of  $n$ . Also assume that each item contributes to, at most, one strand, and the  $k_j$  items in strand  $j$  contribute a maximum of  $n_j$  points. Define  $X_j$  as the observed raw score on strand  $j$ . 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 strand score, and this information provides a prior distribution for  $T_j$ . This prior distribution of  $T_j$  for a given examinee is assumed to be  $\beta(r_j, s_j)$ :

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

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

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



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

where

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

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

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

with

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

and

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

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

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

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

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

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

The 3PL model is

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

where

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

A generalization of Master's (1982) partial credit (2PPC) model was used for the CR items. The 2PPC model, the same as Muraki's (1992) "generalized partial credit model," has been shown to fit response data obtained from a wide variety of mixed-item type achievement tests (Fitzpatrick, Link, Yen, Burket, Ito,

& Sykes, 1996). For a CR item with  $l_i$  score levels, integer scores were assigned that ranged from 0 to  $l_i - 1$ :

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

where

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

and

$$\gamma_{i0} = 0.$$

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

Item parameters estimated from the national standardization sample are used to obtain SPI values.

$T_{ij}(\theta)$  is the expected score for item  $i$  in strand  $j$ , and  $\theta$  is the common trait value to which the items are scaled:

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

where

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

$T_j$ , the expected proportion of maximum score for strand  $j$ , 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 strand  $j$  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  $\mu(\hat{T}_j | \hat{\theta})$  and variance  $\sigma^2(\hat{T}_j | \hat{\theta})$ . 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  $[\mu(\hat{T}_j | \hat{\theta})]$  and variance  $[\sigma^2(\hat{T}_j | \hat{\theta})]$  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 (Novick & Jackson, 1974, p. 113) produces

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

and

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

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

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

and

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

where

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

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

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

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

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

where

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

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

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

and

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

Thus,

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

and the parameters of the prior beta distribution for  $T_j$  can be expressed in terms of the parameters of the 3PL IRT and 2PPC models. Furthermore, the parameters of the posterior distribution of  $T_j$  also can be expressed in terms of the IRT parameters:

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

and

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

The SPI 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 Estimate

The item responses are assumed to be described by  $P_i(\hat{\theta})$  or  $P_{im}(\hat{\theta})$ , depending on the type of item.

Even if the IRT model accurately described item performance over examinees, their item responses grouped by strand 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 strands 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 strand performance.

As previously noted, the prior is estimated using an ability estimate based on responses to all the items (including the items of strand  $j$ ) 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 strand, 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 strand 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 strand score.

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

The SPI procedure assumes that  $p(X_j | T_j)$  is a binomial distribution. This assumption is appropriate only when all the items in a strand have the same Bernoulli item response function. Not only do real items differ in difficulty, but when there are mixed-item types,  $X_j$  is not the sum of  $n_j$  independent Bernoulli variables. It is instead the total raw score. In essence, the simplifying assumption has been made that each CR item with a maximum score of  $1_j - 1$  is the sum of  $1_j - 1$  independent Bernoulli variables. Thus,

a complex compound distribution is theoretically more applicable than the binomial. Given the complexity of working with such a model, it appears valuable to determine if the simpler model described here is sufficiently accurate to be useful.

Finally, because the prior estimate of  $T_j, \hat{T}_j$ , is based on performance on the entire test, including strand  $j$ , 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 strands that contained varying proportions of the total test points.

## References

- CTB/McGraw-Hill. (2000). *TerraNova* 2nd Edition. Monterey, CA.
- Fitzpatrick, A. R., V. Link, W. M. Yen, G. Burket, K. Ito & R. Sykes (1996). Scaling performance assessments: A comparison between one-parameter and two-parameter partial credit models. *Journal of Educational Measurement*, 33, 291–314.
- Johnson, N. L. & S. Kotz (1970). *Distributions in statistics: Continuous univariate distributions* (Vol. 2). New York: John Wiley.
- Lord, F. M. (1980). *Application of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159–176.
- Novick, M. R. & P. H. Jackson (1974). *Statistical methods for educational and psychological research*. New York: McGraw-Hill.
- 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., R. C. Sykes, K. Ito & M. Julian (1997). *A Bayesian/IRT index of objective performance for tests with mixed-item types*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

## **Appendix I**

### **Conditional Standard Error of Measurement with Cut Scores**

Figure I-1 CSEM with cut scores, ELA Grade 3

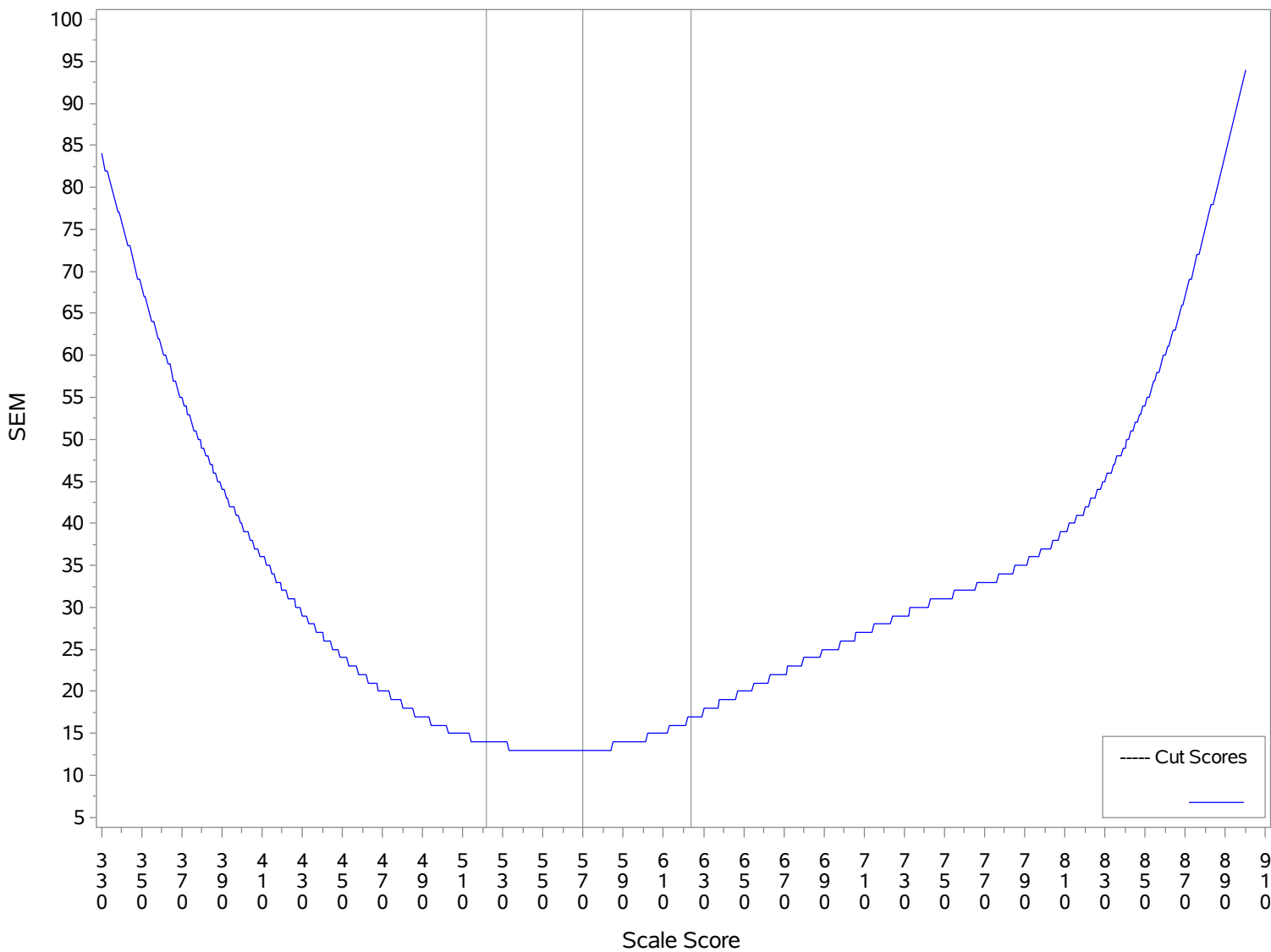




Figure I-2 CSEM with cut scores, ELA Grade 4

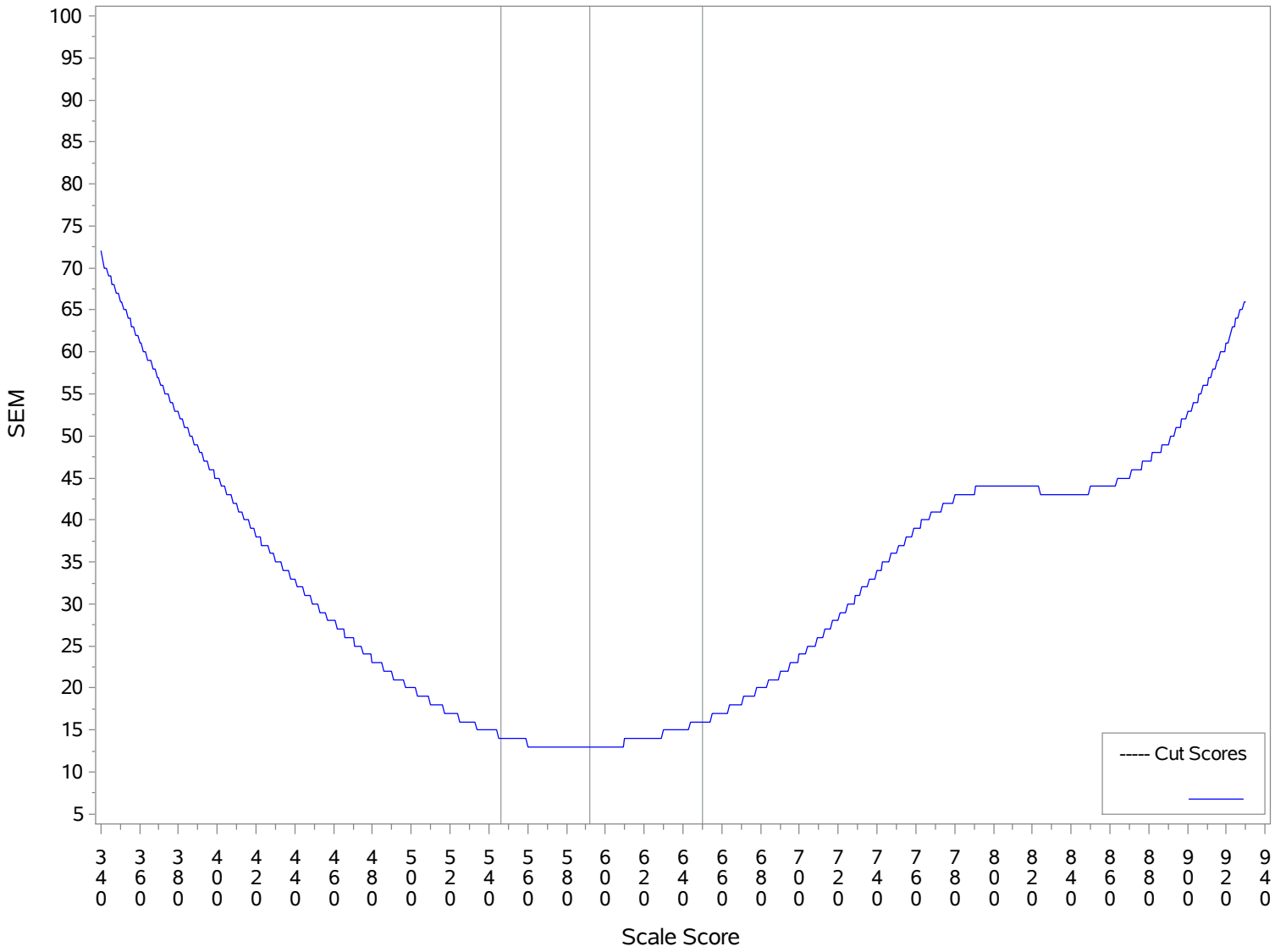


Figure I-3 CSEM with cut scores, ELA Grade 5

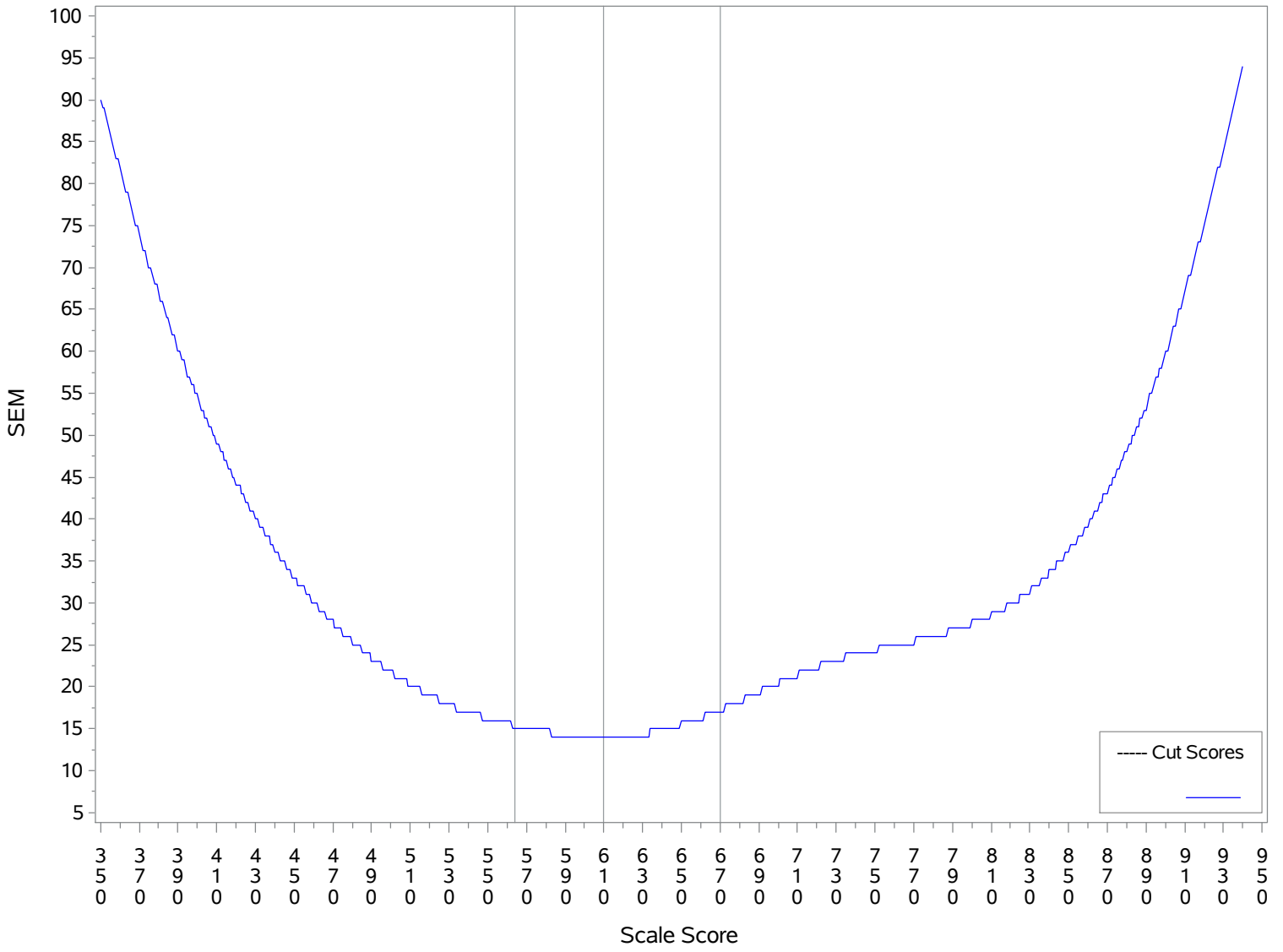


Figure I-4 CSEM with cut scores, ELA Grade 6

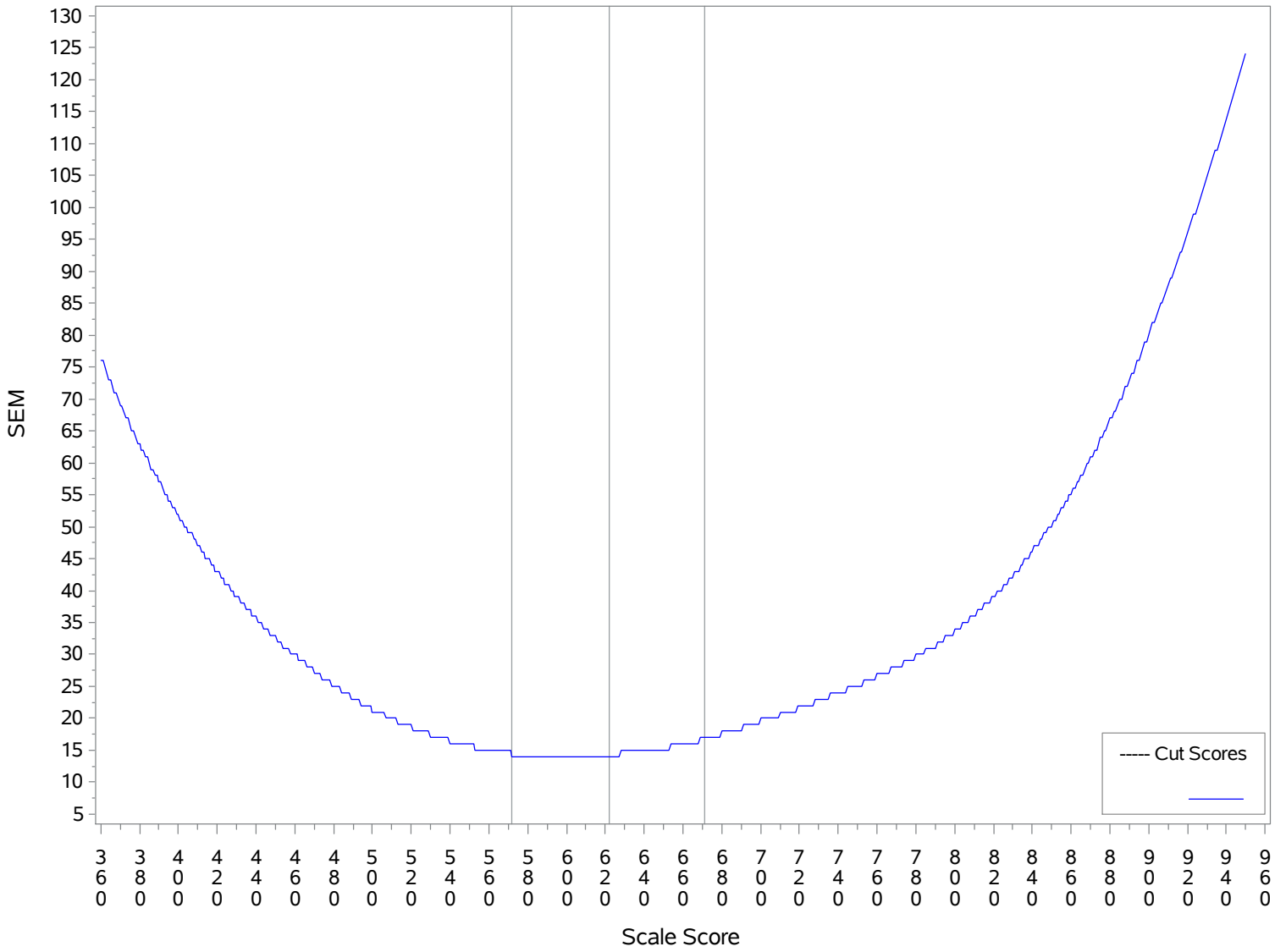


Figure I-5 CSEM with cut scores, ELA Grade 7

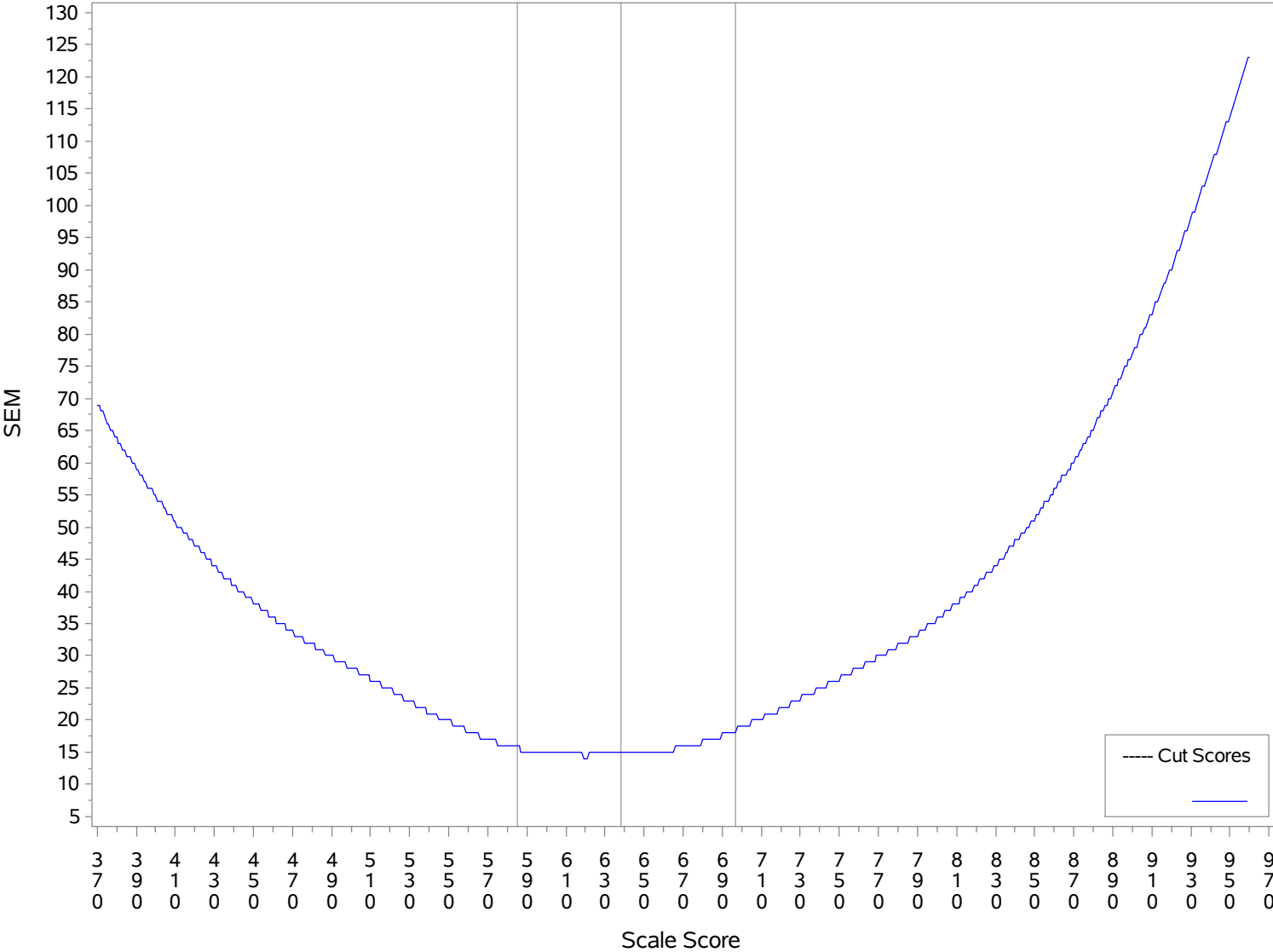


Figure I-6 CSEM with cut scores, ELA Grade 8

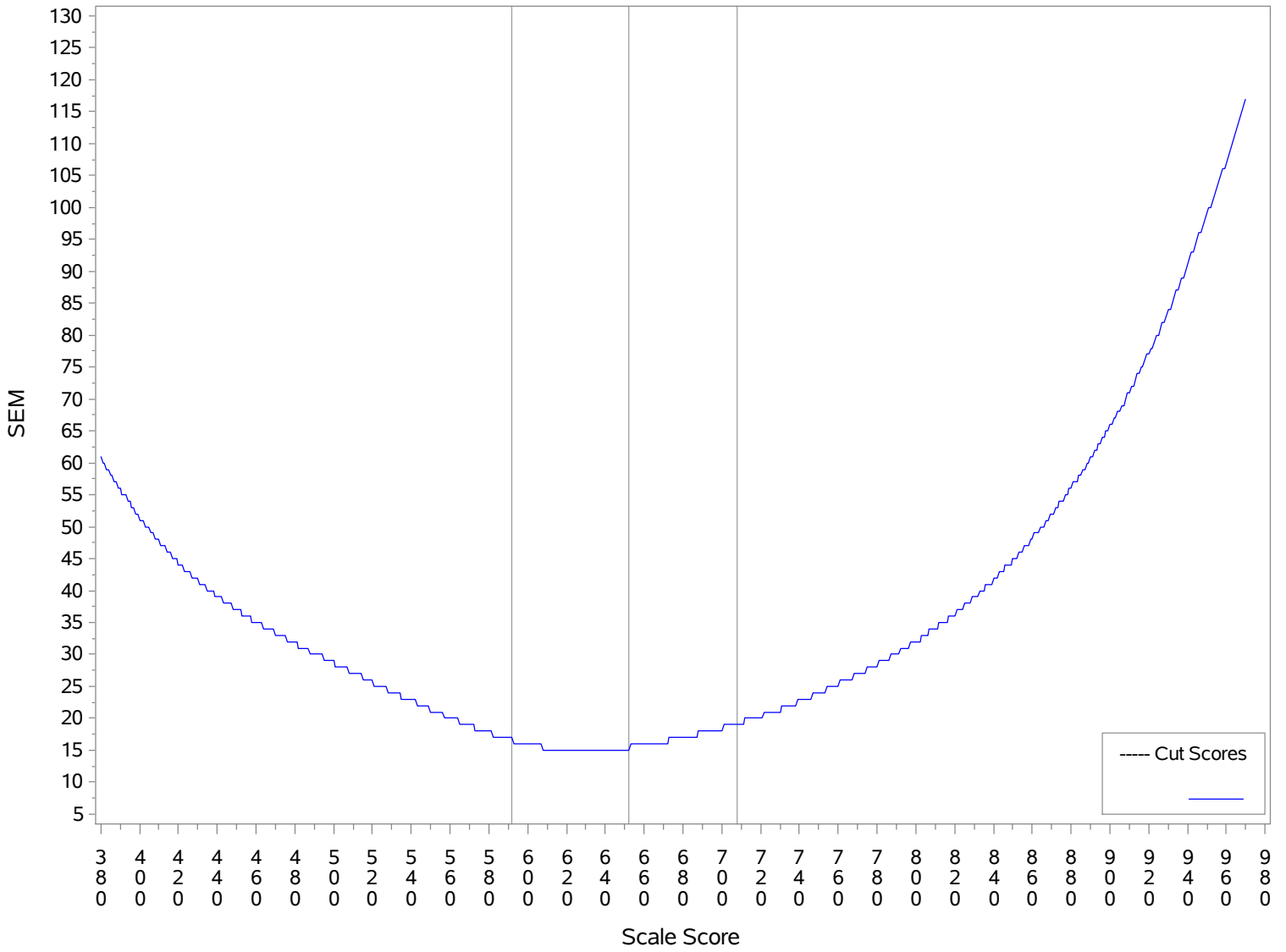


Figure I-7 CSEM with cut scores, Mathematics Grade 3

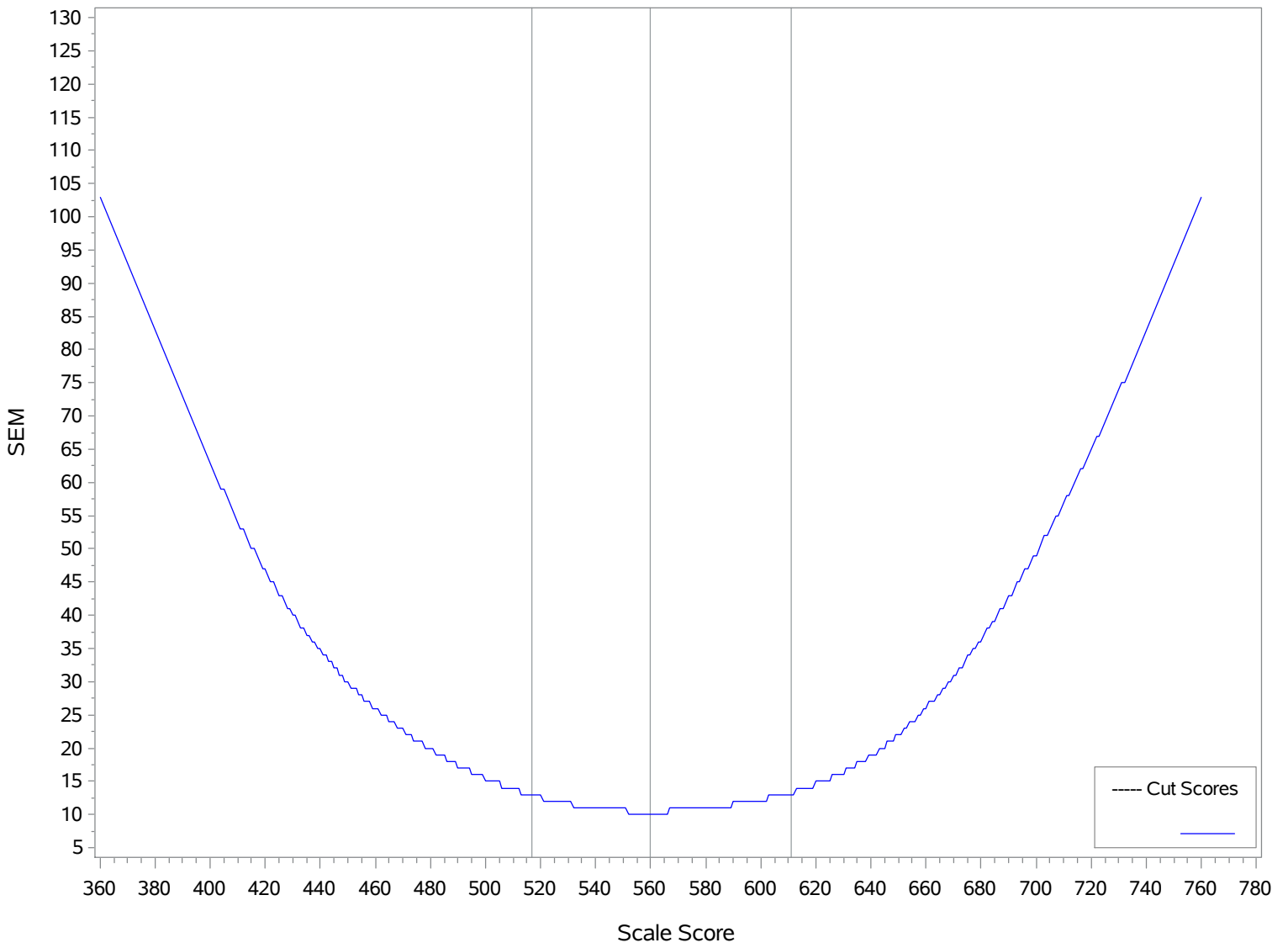


Figure I-8 CSEM with cut scores, Mathematics Grade 4

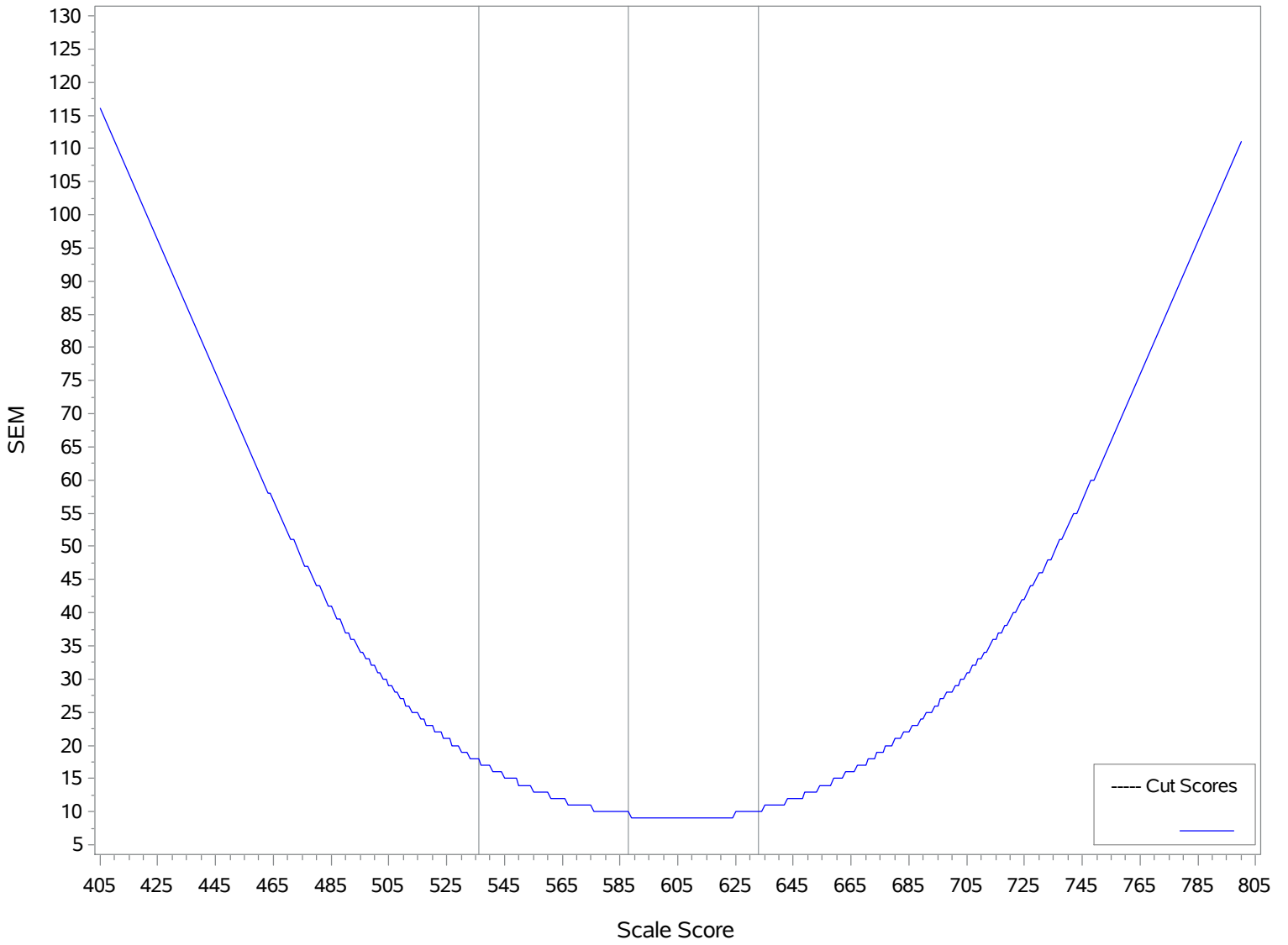


Figure I-9 CSEM with cut scores, Mathematics Grade 5

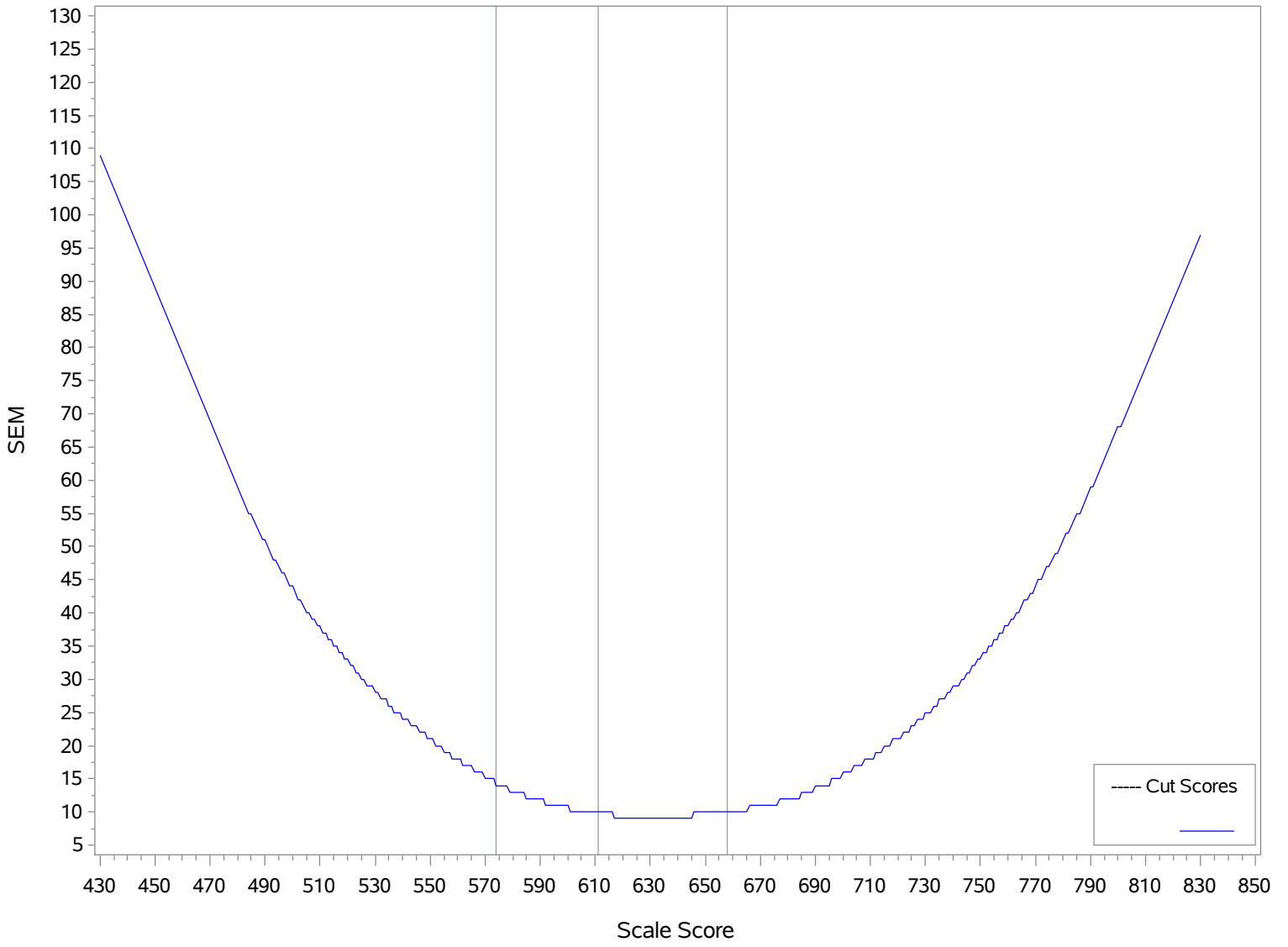




Figure I-10 CSEM with cut scores, Mathematics Grade 6

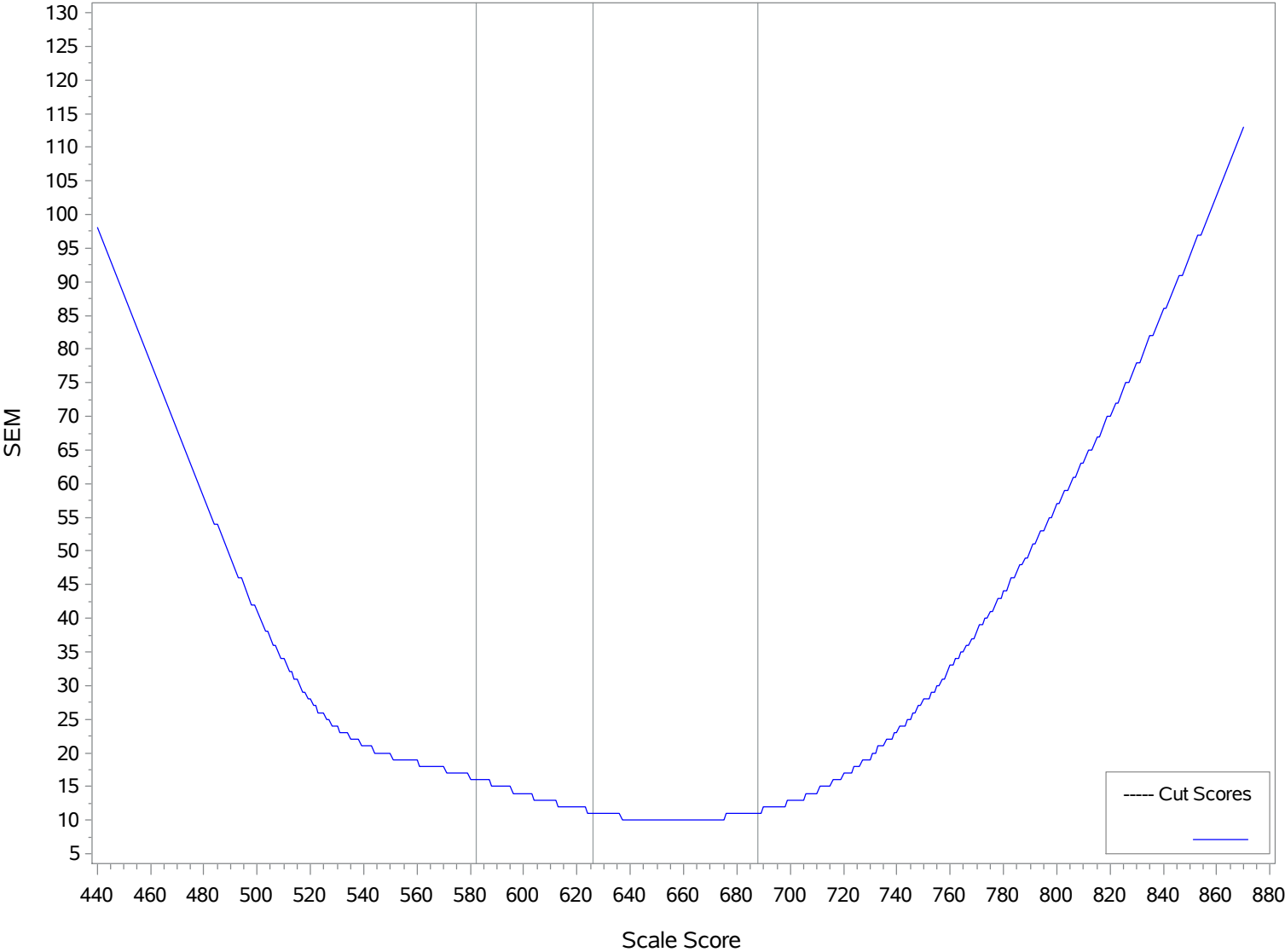


Figure I-11 CSEM with cut scores, Mathematics Grade 7

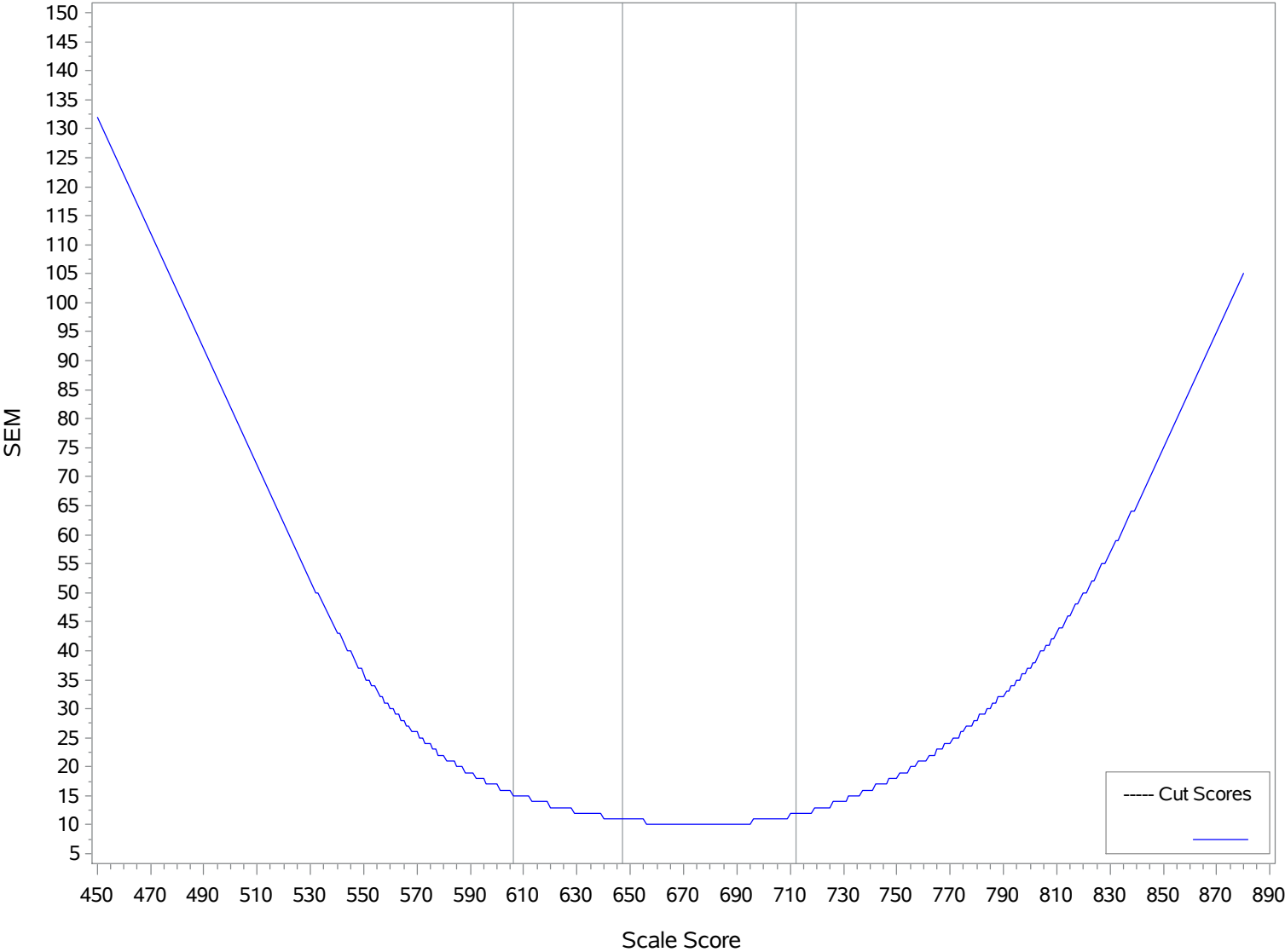


Figure I-12 CSEM with cut scores, Mathematics Grade 8

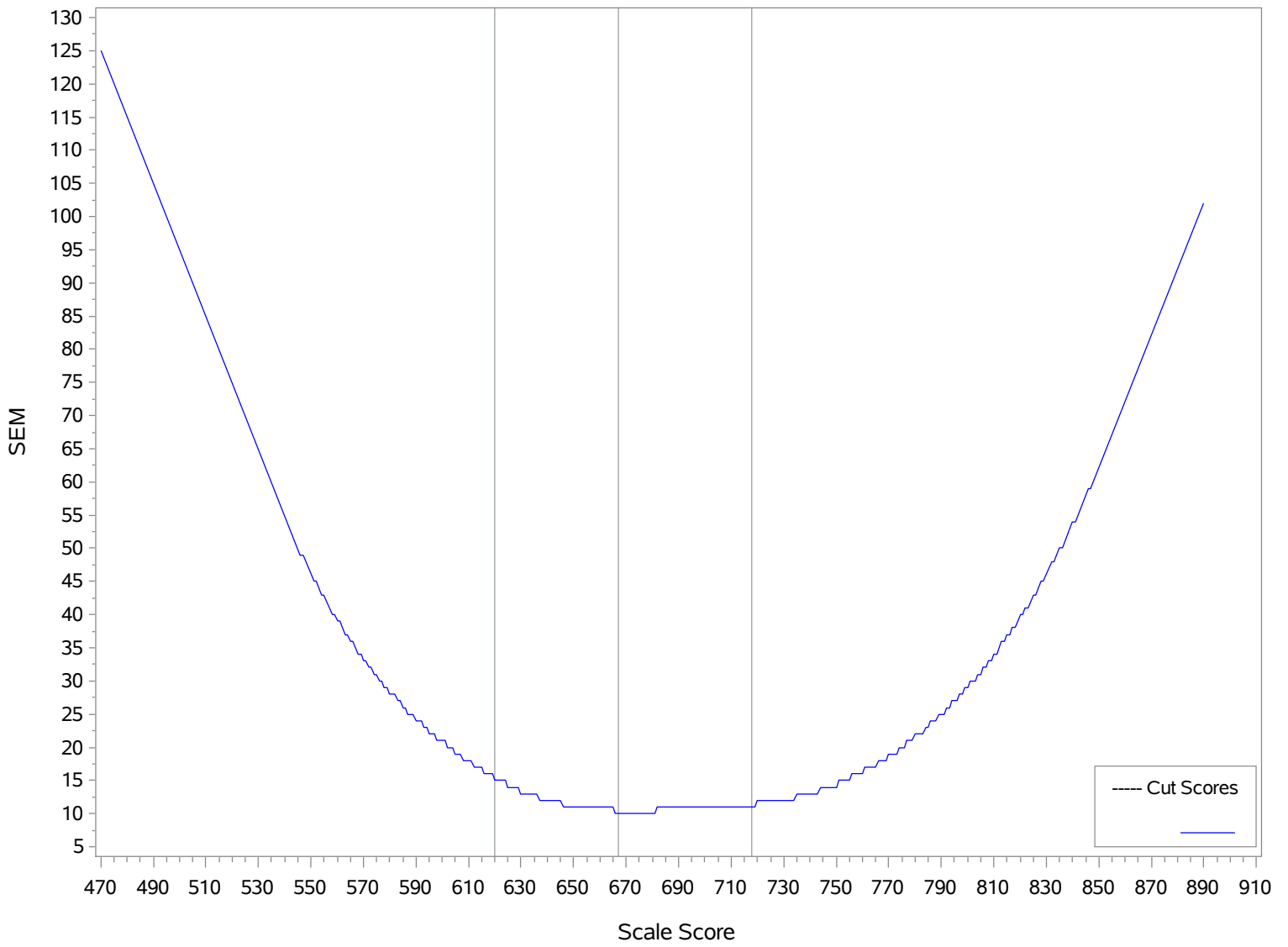


Figure I-13 CSEM with cut scores, Science Grade 4

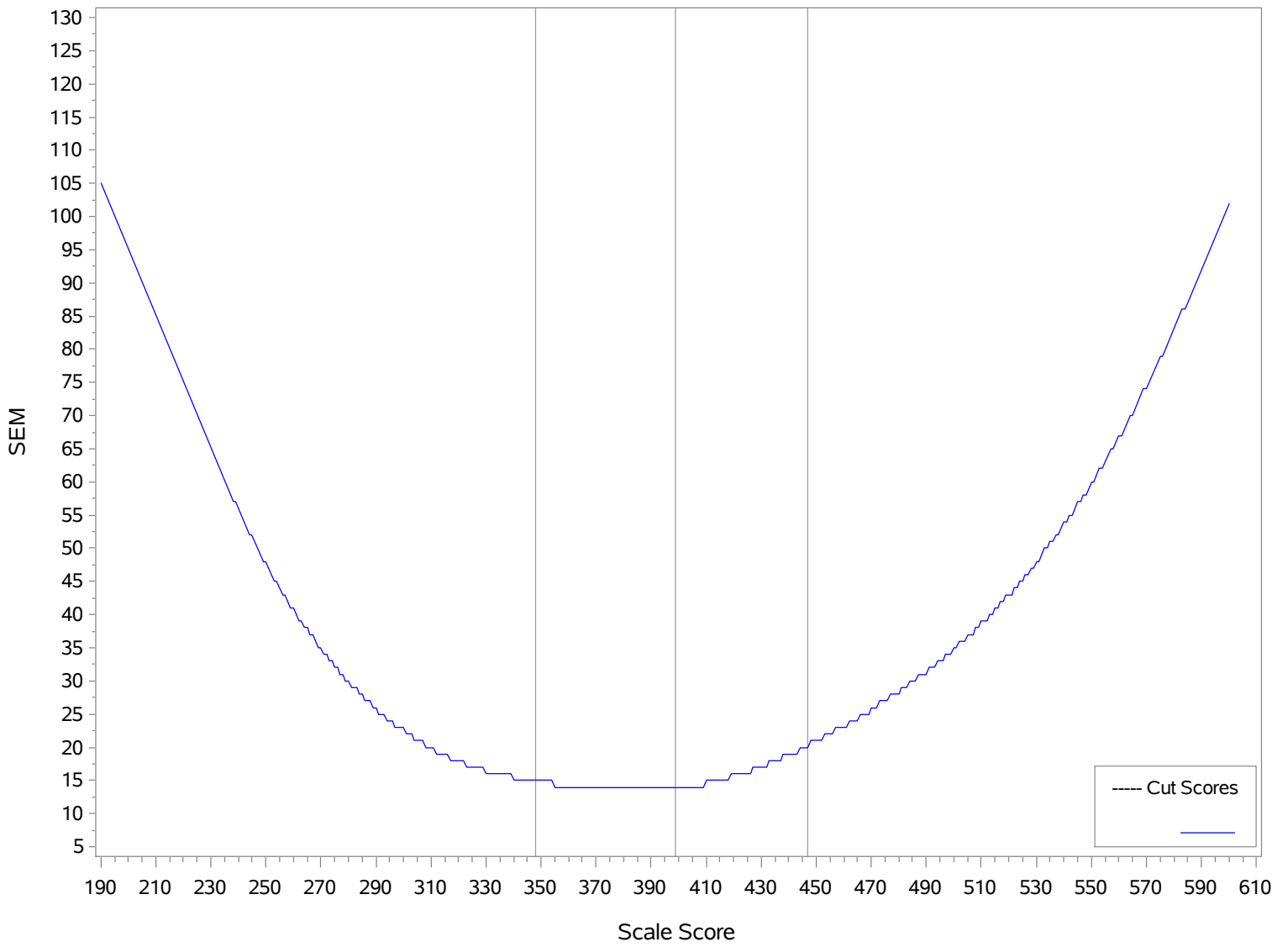


Figure I-14 CSEM with cut scores, Science Grade 8

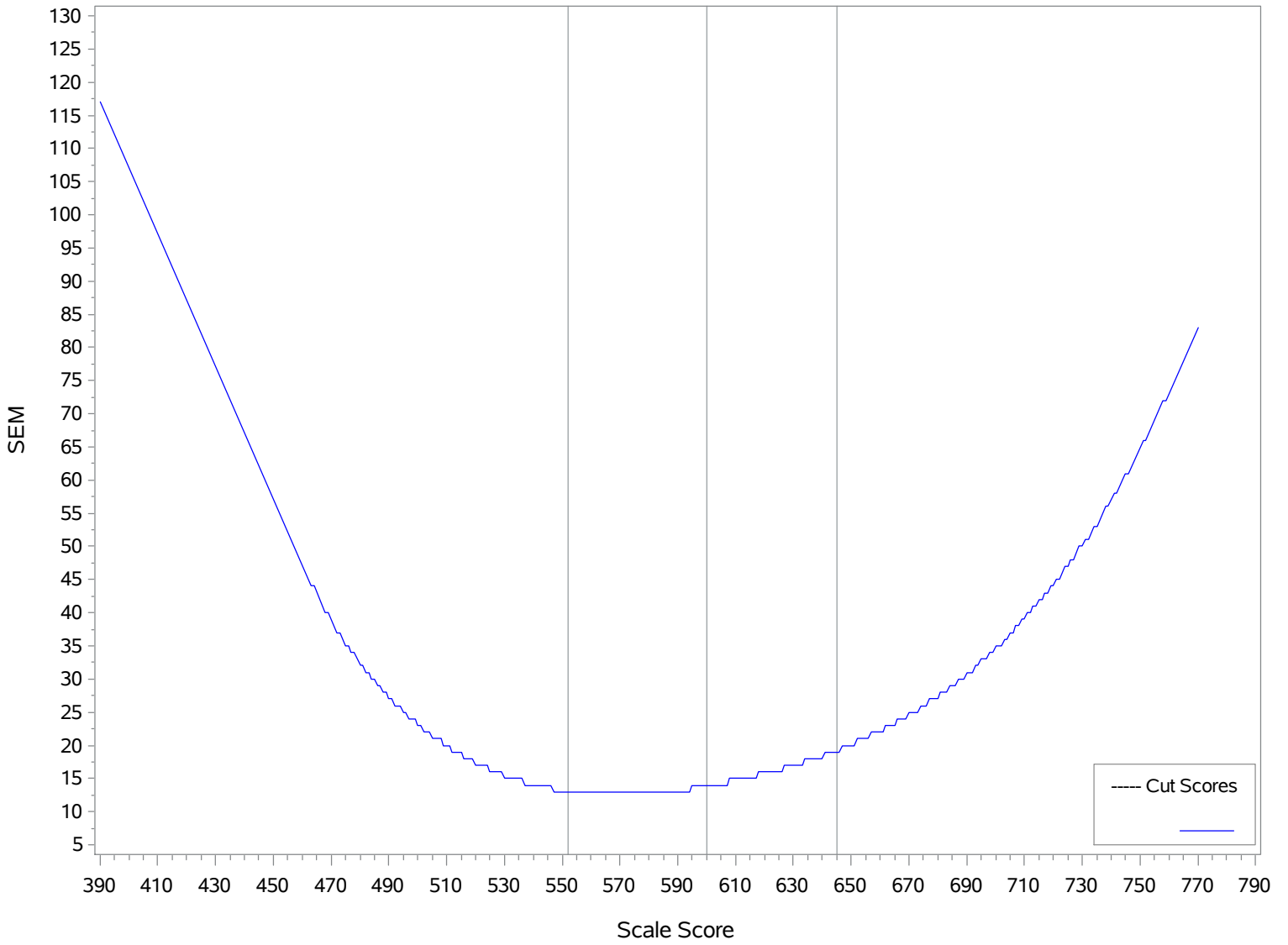


Figure I-15 CSEM with cut scores, Social Studies Grade 4

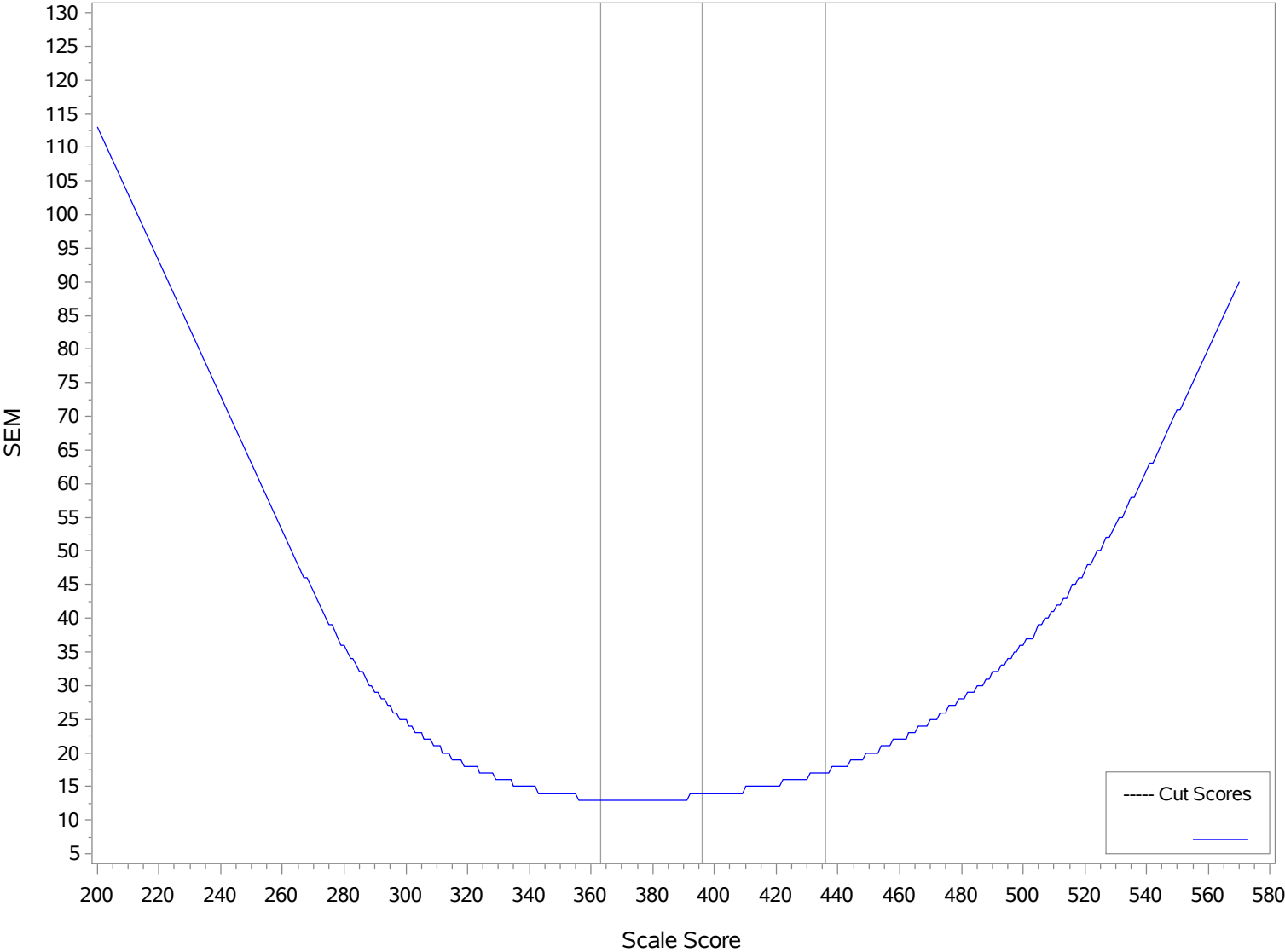


Figure I-16 CSEM with cut scores, Social Studies Grade 8

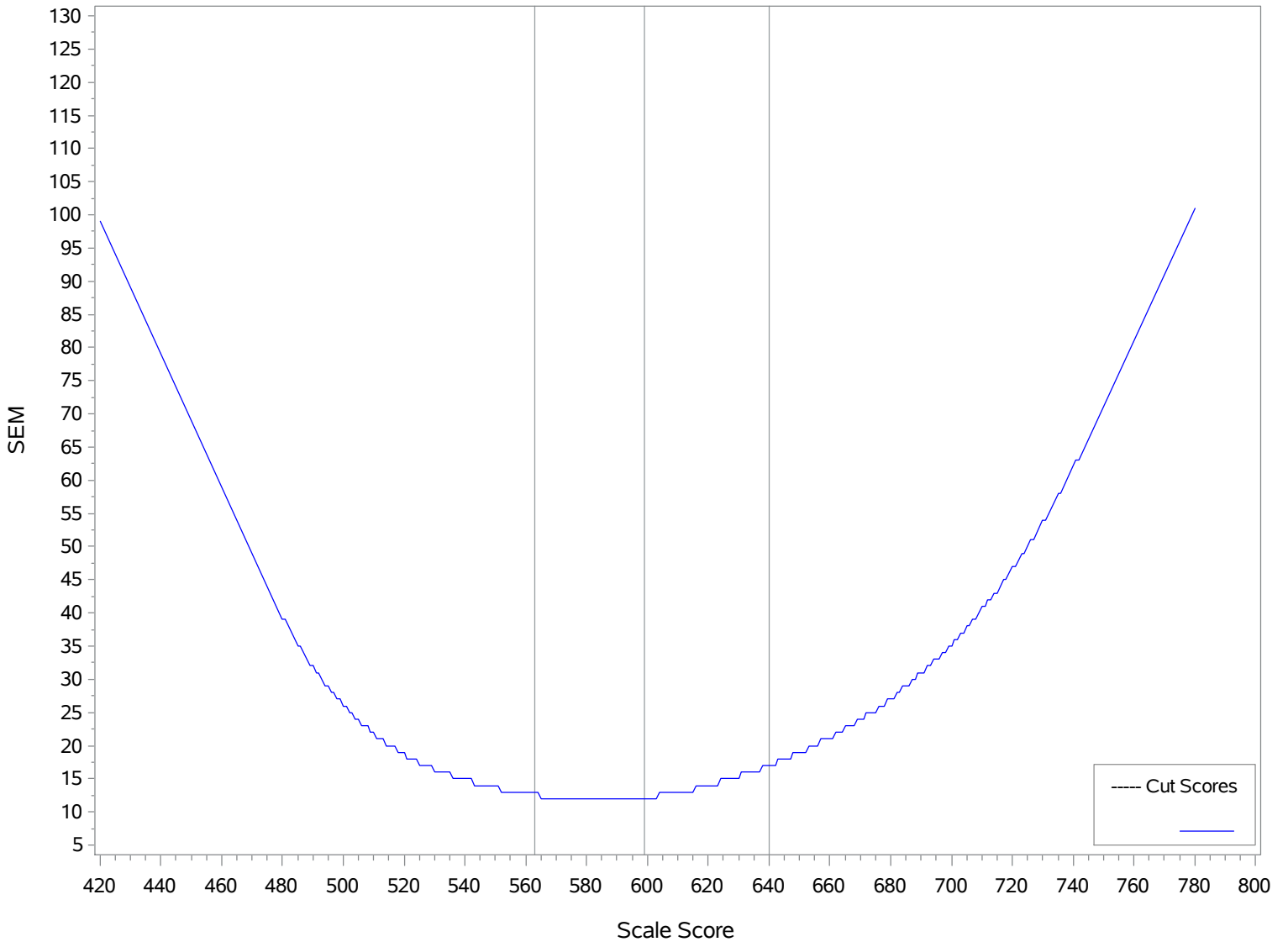
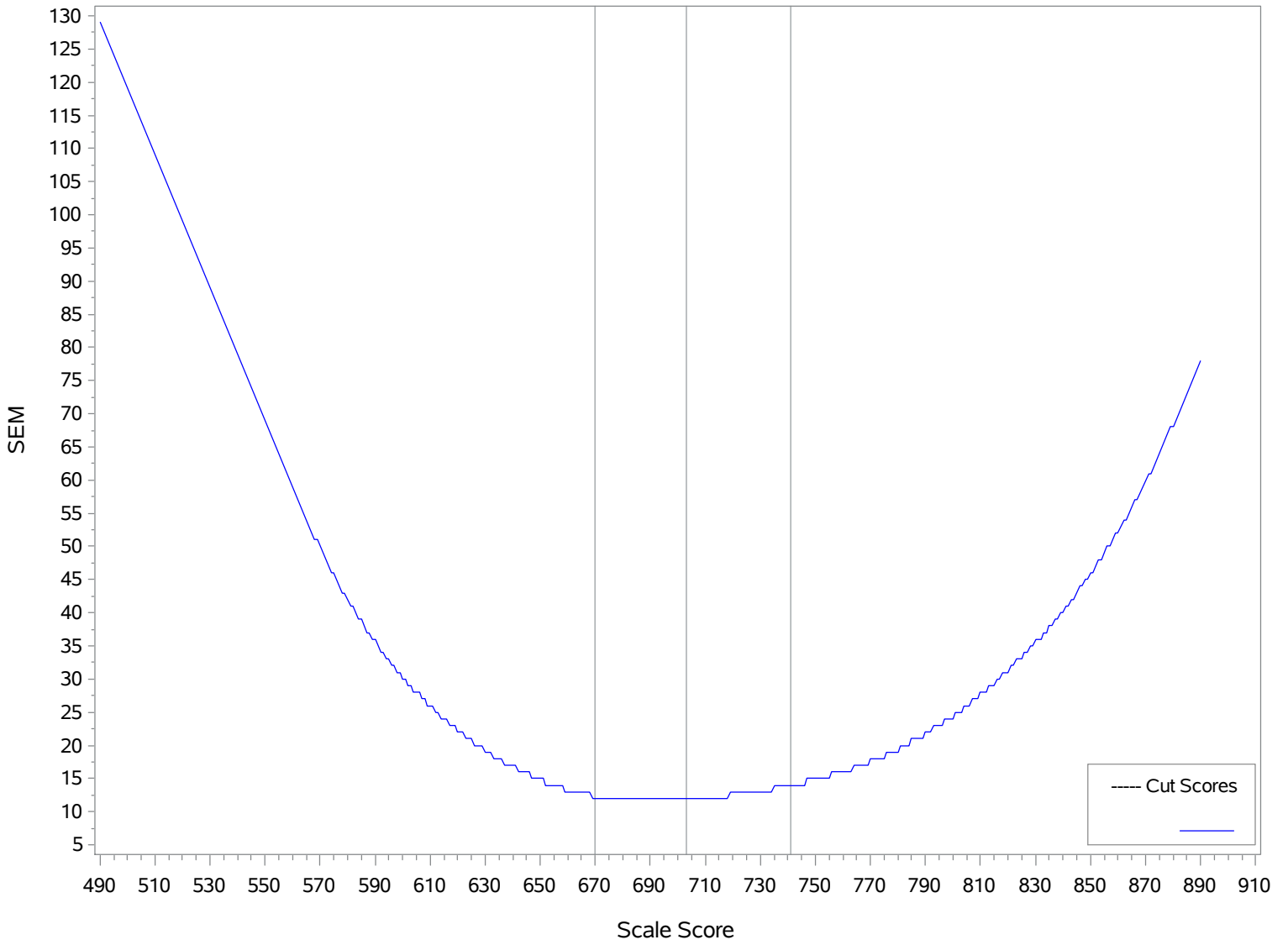


Figure I-17 CSEM with cut scores, Social Studies Grade 10





## **Appendix J**

### **Classification Consistency and Accuracy Analysis by Subgroup**

Table J-1 Indexes for Classification Consistency and Accuracy, ELA Grade 3

Category	Group	Indexes	Cut 1	Cut 2	Cut 3	All Cuts	
Gender	Female	Classification Consistency (P)	0.92	0.91	0.91	0.74	
		Probability of Chance	0.66	0.50	0.70	0.26	
		Kappa (k)	0.78	0.82	0.70	0.65	
		Classification Accuracy	0.94	0.93	0.90	0.77	
	Male	Classification Consistency (P)	0.90	0.89	0.94	0.73	
		Probability of Chance	0.63	0.53	0.86	0.30	
		Kappa (k)	0.72	0.76	0.58	0.61	
		Classification Accuracy	0.93	0.92	0.96	0.81	
Race/Ethnicity	White	Classification Consistency (P)	0.92	0.88	0.92	0.71	
		Probability of Chance	0.75	0.50	0.79	0.31	
		Kappa (k)	0.69	0.75	0.59	0.59	
		Classification Accuracy	0.95	0.91	0.94	0.80	
	African-American	Classification Consistency (P)	0.86	0.94	0.98	0.78	
		Probability of Chance	0.50	0.75	0.97	0.40	
		Kappa (k)	0.71	0.75	0.54	0.63	
		Classification Accuracy	0.90	0.96	0.99	0.85	
	Hispanic	Classification Consistency (P)	0.87	0.90	0.97	0.73	
		Probability of Chance	0.56	0.61	0.92	0.32	
		Kappa (k)	0.70	0.74	0.58	0.61	
		Classification Accuracy	0.91	0.93	0.98	0.81	
	Asian	Classification Consistency (P)	0.89	0.89	0.94	0.72	
		Probability of Chance	0.65	0.52	0.82	0.29	
		Kappa (k)	0.70	0.77	0.65	0.61	
		Classification Accuracy	0.93	0.92	0.95	0.80	
	American Indian	Classification Consistency (P)	0.87	0.91	0.97	0.74	
		Probability of Chance	0.54	0.62	0.94	0.33	
		Kappa (k)	0.71	0.75	0.52	0.62	
		Classification Accuracy	0.91	0.93	0.98	0.82	
	Two or More	Classification Consistency (P)	0.90	0.89	0.94	0.72	
		Probability of Chance	0.64	0.52	0.83	0.29	
		Kappa (k)	0.72	0.77	0.62	0.61	
		Classification Accuracy	0.93	0.92	0.96	0.81	
	Limited English Proficiency	Yes	Classification Consistency (P)	0.85	0.91	0.98	0.74
			Probability of Chance	0.53	0.69	0.96	0.36
			Kappa (k)	0.69	0.70	0.51	0.60
			Classification Accuracy	0.90	0.93	0.99	0.82
Disability Status	Yes	Classification Consistency (P)	0.86	0.92	0.98	0.76	
		Probability of Chance	0.50	0.69	0.95	0.36	
		Kappa (k)	0.72	0.76	0.57	0.63	
		Classification Accuracy	0.90	0.95	0.98	0.83	

Table J-1 Indexes for Classification Consistency and Accuracy, ELA Grade 3 (cont.)

Category	Group	Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Accommodation Use	Yes	Classification Consistency (P)	0.85	0.94	0.99	0.78
		Probability of Chance	0.51	0.80	0.99	0.43
		Kappa (k)	0.70	0.69	0.44	0.62
		Classification Accuracy	0.90	0.96	0.99	0.85
SES Disadvantaged	Yes	Classification Consistency (P)	0.87	0.90	0.97	0.74
		Probability of Chance	0.55	0.61	0.92	0.32
		Kappa (k)	0.71	0.75	0.57	0.62
		Classification Accuracy	0.91	0.93	0.98	0.82

Table J-2 Indexes for Classification Consistency and Accuracy, ELA Grade 4

Category	Group	Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Gender	Female	Classification Consistency (P)	0.92	0.89	0.92	0.73
		Probability of Chance	0.69	0.50	0.77	0.29
		Kappa (k)	0.74	0.77	0.68	0.62
		Classification Accuracy	0.95	0.92	0.95	0.81
	Male	Classification Consistency (P)	0.91	0.89	0.94	0.74
		Probability of Chance	0.63	0.51	0.83	0.29
		Kappa (k)	0.76	0.77	0.65	0.63
		Classification Accuracy	0.94	0.92	0.96	0.82
Race/Ethnicity	White	Classification Consistency (P)	0.93	0.88	0.92	0.72
		Probability of Chance	0.76	0.51	0.76	0.30
		Kappa (k)	0.71	0.75	0.66	0.60
		Classification Accuracy	0.95	0.91	0.94	0.81
	African-American	Classification Consistency (P)	0.87	0.93	0.98	0.79
		Probability of Chance	0.50	0.72	0.96	0.40
		Kappa (k)	0.75	0.76	0.59	0.65
		Classification Accuracy	0.91	0.95	0.99	0.86
	Hispanic	Classification Consistency (P)	0.88	0.90	0.97	0.74
		Probability of Chance	0.55	0.58	0.91	0.31
		Kappa (k)	0.73	0.75	0.63	0.62
		Classification Accuracy	0.92	0.93	0.98	0.82
	Asian	Classification Consistency (P)	0.91	0.88	0.94	0.73
		Probability of Chance	0.65	0.51	0.80	0.28
		Kappa (k)	0.73	0.77	0.69	0.62
		Classification Accuracy	0.93	0.92	0.96	0.81
	American Indian	Classification Consistency (P)	0.88	0.90	0.97	0.74
		Probability of Chance	0.54	0.60	0.93	0.32
		Kappa (k)	0.73	0.75	0.58	0.62
		Classification Accuracy	0.92	0.93	0.98	0.82
Two or More	Classification Consistency (P)	0.91	0.88	0.94	0.74	
	Probability of Chance	0.64	0.51	0.83	0.29	
	Kappa (k)	0.75	0.76	0.67	0.63	
	Classification Accuracy	0.94	0.92	0.96	0.81	
Limited English Proficiency	Yes	Classification Consistency (P)	0.85	0.91	0.99	0.75
		Probability of Chance	0.51	0.74	0.98	0.38
		Kappa (k)	0.69	0.65	0.45	0.59
		Classification Accuracy	0.89	0.94	0.99	0.82
Disability Status	Yes	Classification Consistency (P)	0.87	0.93	0.98	0.78
		Probability of Chance	0.50	0.69	0.94	0.37
		Kappa (k)	0.75	0.76	0.62	0.65
		Classification Accuracy	0.91	0.95	0.98	0.85

Table J-2 Indexes for Classification Consistency and Accuracy, ELA Grade 4 (cont.)

Category	Group	Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Accommodation Use	Yes	Classification Consistency (P)	0.87	0.93	0.99	0.79
		Probability of Chance	0.51	0.79	0.99	0.44
		Kappa (k)	0.73	0.69	0.46	0.63
		Classification Accuracy	0.91	0.95	0.99	0.86
SES Disadvantaged	Yes	Classification Consistency (P)	0.89	0.90	0.97	0.75
		Probability of Chance	0.54	0.58	0.91	0.31
		Kappa (k)	0.75	0.75	0.61	0.63
		Classification Accuracy	0.92	0.93	0.98	0.82

Table J-3 Indexes for Classification Consistency and Accuracy, ELA Grade 5

Category	Group	Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Gender	Female	Classification Consistency (P)	0.91	0.88	0.92	0.72
		Probability of Chance	0.71	0.50	0.78	0.29
		Kappa (k)	0.71	0.76	0.65	0.61
		Classification Accuracy	0.94	0.91	0.95	0.80
	Male	Classification Consistency (P)	0.90	0.88	0.95	0.73
		Probability of Chance	0.63	0.51	0.86	0.30
		Kappa (k)	0.72	0.76	0.61	0.61
		Classification Accuracy	0.93	0.92	0.96	0.81
Race/Ethnicity	White	Classification Consistency (P)	0.92	0.87	0.92	0.72
		Probability of Chance	0.76	0.51	0.78	0.31
		Kappa (k)	0.68	0.74	0.63	0.59
		Classification Accuracy	0.95	0.91	0.95	0.80
	African-American	Classification Consistency (P)	0.86	0.93	0.98	0.77
		Probability of Chance	0.50	0.73	0.96	0.39
		Kappa (k)	0.71	0.74	0.56	0.62
		Classification Accuracy	0.90	0.95	0.99	0.84
	Hispanic	Classification Consistency (P)	0.87	0.89	0.97	0.73
		Probability of Chance	0.57	0.58	0.92	0.31
		Kappa (k)	0.69	0.74	0.57	0.60
		Classification Accuracy	0.91	0.92	0.98	0.80
	Asian	Classification Consistency (P)	0.89	0.89	0.94	0.73
		Probability of Chance	0.64	0.51	0.81	0.28
		Kappa (k)	0.70	0.78	0.69	0.62
		Classification Accuracy	0.93	0.92	0.96	0.81
	American Indian	Classification Consistency (P)	0.86	0.89	0.97	0.73
		Probability of Chance	0.55	0.61	0.95	0.33
		Kappa (k)	0.70	0.73	0.50	0.60
		Classification Accuracy	0.91	0.92	0.98	0.81
	Two or More	Classification Consistency (P)	0.89	0.89	0.95	0.73
		Probability of Chance	0.65	0.51	0.84	0.29
		Kappa (k)	0.70	0.76	0.67	0.62
		Classification Accuracy	0.93	0.91	0.96	0.80
Limited English Proficiency	Yes	Classification Consistency (P)	0.82	0.93	0.99	0.75
		Probability of Chance	0.50	0.82	0.99	0.42
		Kappa (k)	0.64	0.61	0.49	0.56
		Classification Accuracy	0.87	0.95	0.99	0.82
Disability Status	Yes	Classification Consistency (P)	0.86	0.93	0.99	0.78
		Probability of Chance	0.50	0.74	0.97	0.41
		Kappa (k)	0.72	0.74	0.59	0.63
		Classification Accuracy	0.90	0.95	0.99	0.84

Table J-3 Indexes for Classification Consistency and Accuracy, ELA Grade 5 (cont.)

Category	Group	Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Accommodation Use	Yes	Classification Consistency (P)	0.85	0.95	0.99	0.79
		Probability of Chance	0.53	0.84	0.99	0.48
		Kappa (k)	0.68	0.66	0.45	0.60
		Classification Accuracy	0.90	0.96	0.99	0.85
SES Disadvantaged	Yes	Classification Consistency (P)	0.87	0.89	0.97	0.73
		Probability of Chance	0.55	0.59	0.93	0.32
		Kappa (k)	0.71	0.74	0.58	0.61
		Classification Accuracy	0.91	0.92	0.98	0.81

Table J-4 Indexes for Classification Consistency and Accuracy, ELA Grade 6

Category	Group	Indexes	Cut 1	Cut 2	Cut 3	All Cuts	
Gender	Female	Classification Consistency (P)	0.93	0.87	0.90	0.70	
		Probability of Chance	0.75	0.50	0.72	0.28	
		Kappa (k)	0.71	0.74	0.65	0.59	
		Classification Accuracy	0.94	0.91	0.93	0.78	
	Male	Classification Consistency (P)	0.91	0.87	0.93	0.71	
		Probability of Chance	0.65	0.52	0.82	0.29	
		Kappa (k)	0.74	0.73	0.61	0.59	
		Classification Accuracy	0.93	0.91	0.95	0.79	
Race/Ethnicity	White	Classification Consistency (P)	0.94	0.86	0.90	0.70	
		Probability of Chance	0.79	0.50	0.72	0.30	
		Kappa (k)	0.70	0.72	0.63	0.57	
		Classification Accuracy	0.95	0.90	0.93	0.78	
	African-American	Classification Consistency (P)	0.86	0.92	0.98	0.76	
		Probability of Chance	0.50	0.73	0.95	0.38	
		Kappa (k)	0.72	0.70	0.60	0.62	
		Classification Accuracy	0.89	0.94	0.99	0.82	
	Hispanic	Classification Consistency (P)	0.88	0.88	0.96	0.72	
		Probability of Chance	0.58	0.59	0.89	0.32	
		Kappa (k)	0.71	0.71	0.60	0.59	
		Classification Accuracy	0.91	0.91	0.97	0.79	
	Asian	Classification Consistency (P)	0.93	0.87	0.92	0.71	
		Probability of Chance	0.72	0.50	0.74	0.28	
		Kappa (k)	0.74	0.74	0.68	0.60	
		Classification Accuracy	0.94	0.91	0.94	0.79	
	American Indian	Classification Consistency (P)	0.87	0.89	0.97	0.73	
		Probability of Chance	0.56	0.62	0.92	0.33	
		Kappa (k)	0.70	0.70	0.62	0.59	
		Classification Accuracy	0.90	0.92	0.98	0.80	
	Two or More	Classification Consistency (P)	0.91	0.87	0.93	0.71	
		Probability of Chance	0.67	0.51	0.80	0.29	
		Kappa (k)	0.74	0.74	0.64	0.59	
		Classification Accuracy	0.93	0.91	0.95	0.79	
	Limited English Proficiency	Yes	Classification Consistency (P)	0.83	0.94	0.99	0.77
			Probability of Chance	0.50	0.87	0.99	0.45
			Kappa (k)	0.65	0.55	0.43	0.57
			Classification Accuracy	0.87	0.96	0.99	0.82
Disability Status	Yes	Classification Consistency (P)	0.86	0.94	0.99	0.78	
		Probability of Chance	0.50	0.79	0.96	0.42	
		Kappa (k)	0.72	0.69	0.59	0.62	
		Classification Accuracy	0.89	0.95	0.99	0.84	



Table J-4 Indexes for Classification Consistency and Accuracy, ELA Grade 6 (cont.)

Category	Group	Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Accommodation Use	Yes	Classification Consistency (P)	0.87	0.95	0.99	0.82
		Probability of Chance	0.53	0.87	0.99	0.49
		Kappa (k)	0.72	0.63	0.59	0.64
		Classification Accuracy	0.90	0.97	0.99	0.86
SES Disadvantaged	Yes	Classification Consistency (P)	0.88	0.88	0.96	0.72
		Probability of Chance	0.57	0.59	0.90	0.32
		Kappa (k)	0.73	0.71	0.58	0.60
		Classification Accuracy	0.91	0.92	0.97	0.79

Table J-5 Indexes for Classification Consistency and Accuracy, ELA Grade 7

Category	Group	Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Gender	Female	Classification Consistency (P)	0.92	0.88	0.92	0.72
		Probability of Chance	0.70	0.50	0.75	0.28
		Kappa (k)	0.73	0.77	0.67	0.61
		Classification Accuracy	0.94	0.92	0.94	0.81
	Male	Classification Consistency (P)	0.91	0.89	0.94	0.74
		Probability of Chance	0.60	0.52	0.84	0.29
		Kappa (k)	0.76	0.77	0.63	0.63
		Classification Accuracy	0.93	0.93	0.96	0.81
Race/Ethnicity	White	Classification Consistency (P)	0.92	0.88	0.92	0.72
		Probability of Chance	0.73	0.50	0.76	0.29
		Kappa (k)	0.72	0.76	0.65	0.60
		Classification Accuracy	0.94	0.92	0.94	0.80
	African-American	Classification Consistency (P)	0.88	0.93	0.98	0.79
		Probability of Chance	0.50	0.73	0.95	0.39
		Kappa (k)	0.75	0.74	0.62	0.65
		Classification Accuracy	0.91	0.95	0.99	0.85
	Hispanic	Classification Consistency (P)	0.88	0.90	0.96	0.74
		Probability of Chance	0.55	0.59	0.90	0.31
		Kappa (k)	0.74	0.75	0.62	0.63
		Classification Accuracy	0.91	0.93	0.97	0.82
	Asian	Classification Consistency (P)	0.92	0.88	0.92	0.72
		Probability of Chance	0.70	0.50	0.76	0.28
		Kappa (k)	0.72	0.76	0.67	0.61
		Classification Accuracy	0.94	0.92	0.95	0.80
	American Indian	Classification Consistency (P)	0.88	0.91	0.97	0.76
		Probability of Chance	0.53	0.62	0.93	0.32
		Kappa (k)	0.74	0.76	0.60	0.64
		Classification Accuracy	0.91	0.94	0.98	0.83
Two or More	Classification Consistency (P)	0.90	0.89	0.94	0.73	
	Probability of Chance	0.62	0.52	0.82	0.28	
	Kappa (k)	0.74	0.78	0.65	0.62	
	Classification Accuracy	0.93	0.93	0.96	0.81	
Limited English Proficiency	Yes	Classification Consistency (P)	0.84	0.95	0.99	0.79
		Probability of Chance	0.52	0.89	0.99	0.48
		Kappa (k)	0.68	0.55	0.35	0.60
		Classification Accuracy	0.88	0.97	0.99	0.85
Disability Status	Yes	Classification Consistency (P)	0.93	0.97	0.99	0.89
		Probability of Chance	0.56	0.79	0.95	0.50
		Kappa (k)	0.85	0.84	0.77	0.78
		Classification Accuracy	0.95	0.97	0.98	0.90

Table J-5 Indexes for Classification Consistency and Accuracy, ELA Grade 7 (cont.)

Category	Group	Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Accommodation Use	Yes	Classification Consistency (P)	0.87	0.96	0.99	0.83
		Probability of Chance	0.58	0.89	0.99	0.55
		Kappa (k)	0.69	0.66	0.55	0.62
		Classification Accuracy	0.90	0.98	0.99	0.88
SES Disadvantaged	Yes	Classification Consistency (P)	0.88	0.90	0.96	0.75
		Probability of Chance	0.54	0.61	0.91	0.32
		Kappa (k)	0.75	0.75	0.60	0.63
		Classification Accuracy	0.91	0.93	0.98	0.82

Table J-6 Indexes for Classification Consistency and Accuracy, ELA Grade 8

Category	Group	Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Gender	Female	Classification Consistency (P)	0.93	0.88	0.91	0.71
		Probability of Chance	0.72	0.50	0.71	0.28
		Kappa (k)	0.73	0.76	0.68	0.60
		Classification Accuracy	0.95	0.92	0.93	0.80
	Male	Classification Consistency (P)	0.90	0.89	0.93	0.73
		Probability of Chance	0.60	0.54	0.82	0.29
		Kappa (k)	0.76	0.76	0.64	0.62
		Classification Accuracy	0.93	0.92	0.95	0.81
Race/Ethnicity	White	Classification Consistency (P)	0.93	0.87	0.91	0.71
		Probability of Chance	0.73	0.50	0.72	0.28
		Kappa (k)	0.73	0.75	0.66	0.59
		Classification Accuracy	0.95	0.91	0.93	0.80
	African-American	Classification Consistency (P)	0.87	0.93	0.98	0.78
		Probability of Chance	0.50	0.76	0.95	0.39
		Kappa (k)	0.74	0.72	0.58	0.63
		Classification Accuracy	0.91	0.95	0.98	0.84
	Hispanic	Classification Consistency (P)	0.88	0.89	0.96	0.74
		Probability of Chance	0.56	0.62	0.89	0.32
		Kappa (k)	0.74	0.73	0.63	0.61
		Classification Accuracy	0.92	0.93	0.97	0.81
	Asian	Classification Consistency (P)	0.91	0.89	0.92	0.73
		Probability of Chance	0.69	0.51	0.72	0.28
		Kappa (k)	0.71	0.78	0.72	0.62
		Classification Accuracy	0.94	0.92	0.95	0.80
	American Indian	Classification Consistency (P)	0.88	0.90	0.97	0.75
		Probability of Chance	0.54	0.66	0.94	0.34
		Kappa (k)	0.74	0.70	0.52	0.62
		Classification Accuracy	0.92	0.93	0.98	0.82
Two or More	Classification Consistency (P)	0.91	0.90	0.93	0.74	
	Probability of Chance	0.62	0.53	0.79	0.28	
	Kappa (k)	0.75	0.78	0.68	0.63	
	Classification Accuracy	0.94	0.92	0.95	0.81	
Limited English Proficiency	Yes	Classification Consistency (P)	0.83	0.95	0.99	0.79
		Probability of Chance	0.51	0.90	0.99	0.47
		Kappa (k)	0.66	0.54	0.50	0.60
		Classification Accuracy	0.88	0.97	0.99	0.85
Disability Status	Yes	Classification Consistency (P)	0.86	0.95	0.99	0.80
		Probability of Chance	0.52	0.84	0.97	0.47
		Kappa (k)	0.71	0.71	0.58	0.63
		Classification Accuracy	0.90	0.97	0.99	0.86

Table J-6 Indexes for Classification Consistency and Accuracy, ELA Grade 8 (cont.)

Category	Group	Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Accommodation Use	Yes	Classification Consistency (P)	0.85	0.96	0.99	0.81
		Probability of Chance	0.54	0.89	0.98	0.51
		Kappa (k)	0.68	0.65	0.68	0.61
		Classification Accuracy	0.90	0.97	0.99	0.87
SES Disadvantaged	Yes	Classification Consistency (P)	0.88	0.90	0.96	0.74
		Probability of Chance	0.54	0.63	0.90	0.32
		Kappa (k)	0.74	0.73	0.62	0.62
		Classification Accuracy	0.92	0.93	0.97	0.82

Table J-7 Indexes for Classification Consistency and Accuracy, Mathematics Grade 3

Category	Group	Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Gender	Female	Classification Consistency (P)	0.94	0.92	0.93	0.79
		Probability of Chance	0.65	0.50	0.74	0.27
		Kappa (k)	0.82	0.85	0.74	0.72
		Classification Accuracy	0.95	0.94	0.93	0.82
	Male	Classification Consistency (P)	0.93	0.89	0.93	0.75
		Probability of Chance	0.68	0.50	0.78	0.29
		Kappa (k)	0.76	0.79	0.68	0.65
		Classification Accuracy	0.95	0.92	0.95	0.82
Race/Ethnicity	White	Classification Consistency (P)	0.94	0.88	0.92	0.74
		Probability of Chance	0.78	0.51	0.75	0.31
		Kappa (k)	0.72	0.76	0.66	0.61
		Classification Accuracy	0.96	0.92	0.94	0.82
	African-American	Classification Consistency (P)	0.86	0.93	0.99	0.78
		Probability of Chance	0.50	0.72	0.97	0.38
		Kappa (k)	0.73	0.74	0.56	0.64
		Classification Accuracy	0.91	0.95	0.99	0.85
	Hispanic	Classification Consistency (P)	0.88	0.89	0.97	0.75
		Probability of Chance	0.57	0.58	0.92	0.31
		Kappa (k)	0.73	0.75	0.63	0.63
		Classification Accuracy	0.92	0.93	0.98	0.82
	Asian	Classification Consistency (P)	0.91	0.90	0.94	0.74
		Probability of Chance	0.68	0.50	0.75	0.27
		Kappa (k)	0.71	0.79	0.75	0.65
		Classification Accuracy	0.94	0.93	0.96	0.82
	American Indian	Classification Consistency (P)	0.88	0.90	0.97	0.75
		Probability of Chance	0.56	0.59	0.92	0.32
		Kappa (k)	0.73	0.76	0.57	0.63
		Classification Accuracy	0.92	0.93	0.98	0.83
Two or More	Classification Consistency (P)	0.91	0.89	0.94	0.74	
	Probability of Chance	0.65	0.51	0.81	0.29	
	Kappa (k)	0.75	0.78	0.68	0.64	
	Classification Accuracy	0.94	0.92	0.96	0.82	
Limited English Proficiency	Yes	Classification Consistency (P)	0.87	0.90	0.98	0.75
		Probability of Chance	0.54	0.63	0.94	0.33
		Kappa (k)	0.71	0.74	0.64	0.62
		Classification Accuracy	0.91	0.93	0.99	0.83
Disability Status	Yes	Classification Consistency (P)	0.89	0.92	0.97	0.78
		Probability of Chance	0.51	0.62	0.91	0.33
		Kappa (k)	0.77	0.79	0.68	0.67
		Classification Accuracy	0.92	0.95	0.98	0.85

Table J-7 Indexes for Classification Consistency and Accuracy, Mathematics Grade 3 (cont.)

Category	Group	Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Accommodation Use	Yes	Classification Consistency (P)	0.86	0.95	0.99	0.81
		Probability of Chance	0.53	0.83	0.99	0.47
		Kappa (k)	0.71	0.69	0.55	0.63
		Classification Accuracy	0.90	0.96	0.99	0.87
SES Disadvantaged	Yes	Classification Consistency (P)	0.89	0.90	0.97	0.75
		Probability of Chance	0.57	0.57	0.91	0.31
		Kappa (k)	0.74	0.76	0.61	0.64
		Classification Accuracy	0.92	0.93	0.98	0.83

Table J-8 Indexes for Classification Consistency and Accuracy, Mathematics Grade 4

Category	Group	Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Gender	Female	Classification Consistency (P)	0.89	0.90	0.95	0.74
		Probability of Chance	0.67	0.52	0.83	0.30
		Kappa (k)	0.68	0.79	0.70	0.64
		Classification Accuracy	0.93	0.93	0.97	0.82
	Male	Classification Consistency (P)	0.91	0.91	0.94	0.75
		Probability of Chance	0.68	0.50	0.77	0.28
		Kappa (k)	0.72	0.81	0.72	0.66
		Classification Accuracy	0.94	0.94	0.96	0.83
Race/Ethnicity	White	Classification Consistency (P)	0.93	0.89	0.93	0.75
		Probability of Chance	0.79	0.50	0.75	0.31
		Kappa (k)	0.67	0.78	0.70	0.63
		Classification Accuracy	0.95	0.92	0.95	0.83
	African-American	Classification Consistency (P)	0.83	0.95	0.99	0.77
		Probability of Chance	0.50	0.79	0.97	0.41
		Kappa (k)	0.65	0.76	0.70	0.61
		Classification Accuracy	0.88	0.97	0.99	0.84
	Hispanic	Classification Consistency (P)	0.85	0.92	0.98	0.75
		Probability of Chance	0.55	0.64	0.92	0.33
		Kappa (k)	0.67	0.77	0.70	0.62
		Classification Accuracy	0.90	0.94	0.98	0.82
	Asian	Classification Consistency (P)	0.89	0.92	0.95	0.76
		Probability of Chance	0.67	0.51	0.74	0.28
		Kappa (k)	0.67	0.83	0.82	0.67
		Classification Accuracy	0.92	0.94	0.97	0.83
	American Indian	Classification Consistency (P)	0.84	0.93	0.98	0.75
		Probability of Chance	0.55	0.64	0.92	0.33
		Kappa (k)	0.64	0.81	0.70	0.62
		Classification Accuracy	0.89	0.95	0.98	0.83
Two or More	Classification Consistency (P)	0.88	0.90	0.95	0.74	
	Probability of Chance	0.65	0.53	0.82	0.29	
	Kappa (k)	0.67	0.80	0.72	0.63	
	Classification Accuracy	0.92	0.93	0.97	0.82	
Limited English Proficiency	Yes	Classification Consistency (P)	0.83	0.94	0.99	0.75
		Probability of Chance	0.51	0.77	0.97	0.39
		Kappa (k)	0.65	0.72	0.65	0.60
		Classification Accuracy	0.88	0.96	0.99	0.83
Disability Status	Yes	Classification Consistency (P)	0.85	0.94	0.98	0.77
		Probability of Chance	0.50	0.69	0.93	0.36
		Kappa (k)	0.70	0.81	0.73	0.65
		Classification Accuracy	0.90	0.96	0.99	0.84



Table J-8 Indexes for Classification Consistency and Accuracy, Mathematics Grade 4 (cont.)

Category	Group	Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Accommodation Use	Yes	Classification Consistency (P)	0.83	0.96	0.99	0.79
		Probability of Chance	0.52	0.88	0.99	0.48
		Kappa (k)	0.64	0.72	0.61	0.60
		Classification Accuracy	0.88	0.98	0.99	0.85
SES Disadvantaged	Yes	Classification Consistency (P)	0.86	0.91	0.97	0.75
		Probability of Chance	0.56	0.62	0.92	0.33
		Kappa (k)	0.68	0.78	0.68	0.63
		Classification Accuracy	0.90	0.94	0.98	0.83

Table J-9 Indexes for Classification Consistency and Accuracy, Mathematics Grade 5

Category	Group	Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Gender	Female	Classification Consistency (P)	0.89	0.89	0.96	0.75
		Probability of Chance	0.62	0.51	0.83	0.29
		Kappa (k)	0.72	0.78	0.75	0.64
		Classification Accuracy	0.93	0.93	0.97	0.82
	Male	Classification Consistency (P)	0.90	0.91	0.95	0.77
		Probability of Chance	0.61	0.50	0.79	0.28
		Kappa (k)	0.74	0.82	0.77	0.68
		Classification Accuracy	0.93	0.94	0.97	0.83
Race/Ethnicity	White	Classification Consistency (P)	0.91	0.89	0.94	0.75
		Probability of Chance	0.72	0.50	0.77	0.30
		Kappa (k)	0.70	0.77	0.76	0.64
		Classification Accuracy	0.94	0.92	0.96	0.82
	African-American	Classification Consistency (P)	0.85	0.94	0.99	0.79
		Probability of Chance	0.52	0.78	0.98	0.45
		Kappa (k)	0.69	0.74	0.69	0.62
		Classification Accuracy	0.90	0.96	0.99	0.85
	Hispanic	Classification Consistency (P)	0.86	0.91	0.98	0.75
		Probability of Chance	0.52	0.61	0.93	0.33
		Kappa (k)	0.70	0.78	0.69	0.63
		Classification Accuracy	0.90	0.94	0.99	0.82
	Asian	Classification Consistency (P)	0.90	0.90	0.95	0.75
		Probability of Chance	0.64	0.50	0.77	0.27
		Kappa (k)	0.72	0.79	0.80	0.66
		Classification Accuracy	0.93	0.93	0.97	0.82
	American Indian	Classification Consistency (P)	0.85	0.92	0.98	0.75
		Probability of Chance	0.51	0.66	0.95	0.35
		Kappa (k)	0.69	0.77	0.63	0.62
		Classification Accuracy	0.89	0.94	0.99	0.81
Two or More	Classification Consistency (P)	0.88	0.90	0.96	0.75	
	Probability of Chance	0.57	0.53	0.85	0.29	
	Kappa (k)	0.73	0.79	0.77	0.65	
	Classification Accuracy	0.91	0.93	0.97	0.82	
Limited English Proficiency	Yes	Classification Consistency (P)	0.83	0.93	0.99	0.76
		Probability of Chance	0.50	0.79	0.99	0.42
		Kappa (k)	0.67	0.68	0.71	0.59
		Classification Accuracy	0.88	0.95	0.99	0.84
Disability Status	Yes	Classification Consistency (P)	0.86	0.94	0.99	0.79
		Probability of Chance	0.51	0.72	0.95	0.41
		Kappa (k)	0.72	0.78	0.76	0.64
		Classification Accuracy	0.90	0.96	0.99	0.85

Table J-9 Indexes for Classification Consistency and Accuracy, Mathematics Grade 5 (cont.)

Category	Group	Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Accommodation Use	Yes	Classification Consistency (P)	0.86	0.96	0.99	0.82
		Probability of Chance	0.58	0.88	0.99	0.55
		Kappa (k)	0.67	0.66	0.54	0.60
		Classification Accuracy	0.90	0.97	0.99	0.87
SES Disadvantaged	Yes	Classification Consistency (P)	0.86	0.92	0.98	0.76
		Probability of Chance	0.52	0.61	0.93	0.33
		Kappa (k)	0.71	0.79	0.71	0.64
		Classification Accuracy	0.90	0.94	0.99	0.83

Table J-10 Indexes for Classification Consistency and Accuracy, Mathematics Grade 6

Category	Group	Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Gender	Female	Classification Consistency (P)	0.90	0.89	0.96	0.75
		Probability of Chance	0.63	0.51	0.88	0.31
		Kappa (k)	0.73	0.78	0.67	0.65
		Classification Accuracy	0.93	0.92	0.98	0.83
	Male	Classification Consistency (P)	0.90	0.90	0.97	0.77
		Probability of Chance	0.61	0.51	0.87	0.30
		Kappa (k)	0.76	0.79	0.74	0.67
		Classification Accuracy	0.93	0.92	0.97	0.82
Race/Ethnicity	White	Classification Consistency (P)	0.91	0.89	0.96	0.76
		Probability of Chance	0.72	0.50	0.85	0.32
		Kappa (k)	0.70	0.78	0.72	0.64
		Classification Accuracy	0.94	0.92	0.97	0.83
	African-American	Classification Consistency (P)	0.86	0.95	0.99	0.80
		Probability of Chance	0.52	0.79	0.99	0.45
		Kappa (k)	0.71	0.75	0.66	0.64
		Classification Accuracy	0.90	0.96	0.99	0.86
	Hispanic	Classification Consistency (P)	0.86	0.90	0.99	0.75
		Probability of Chance	0.52	0.63	0.96	0.34
		Kappa (k)	0.71	0.74	0.67	0.62
		Classification Accuracy	0.90	0.93	0.99	0.83
	Asian	Classification Consistency (P)	0.91	0.89	0.96	0.76
		Probability of Chance	0.65	0.50	0.80	0.29
		Kappa (k)	0.73	0.78	0.79	0.66
		Classification Accuracy	0.94	0.92	0.97	0.83
	American Indian	Classification Consistency (P)	0.85	0.92	0.99	0.76
		Probability of Chance	0.51	0.65	0.98	0.35
		Kappa (k)	0.70	0.77	0.40	0.63
		Classification Accuracy	0.89	0.94	0.99	0.82
Two or More	Classification Consistency (P)	0.88	0.90	0.97	0.75	
	Probability of Chance	0.58	0.53	0.89	0.30	
	Kappa (k)	0.72	0.78	0.74	0.65	
	Classification Accuracy	0.92	0.93	0.98	0.83	
Limited English Proficiency	Yes	Classification Consistency (P)	0.84	0.95	0.99	0.79
		Probability of Chance	0.53	0.86	0.99	0.48
		Kappa (k)	0.66	0.64	0.60	0.59
		Classification Accuracy	0.89	0.97	0.99	0.85
Disability Status	Yes	Classification Consistency (P)	0.88	0.95	0.99	0.83
		Probability of Chance	0.54	0.78	0.98	0.47
		Kappa (k)	0.75	0.78	0.72	0.67
		Classification Accuracy	0.91	0.96	0.99	0.87

Table J-10 Indexes for Classification Consistency and Accuracy, Mathematics Grade 6 (cont.)

Category	Group	Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Accommodation Use	Yes	Classification Consistency (P)	0.88	0.97	0.99	0.86
		Probability of Chance	0.65	0.91	0.99	0.63
		Kappa (k)	0.67	0.69	0.53	0.61
		Classification Accuracy	0.92	0.98	0.99	0.90
SES Disadvantaged	Yes	Classification Consistency (P)	0.87	0.91	0.99	0.77
		Probability of Chance	0.51	0.63	0.96	0.34
		Kappa (k)	0.73	0.75	0.68	0.65
		Classification Accuracy	0.90	0.93	0.99	0.82

Table J-11 Indexes for Classification Consistency and Accuracy, Mathematics Grade 7

Category	Group	Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Gender	Female	Classification Consistency (P)	0.88	0.90	0.97	0.75
		Probability of Chance	0.57	0.53	0.91	0.30
		Kappa (k)	0.72	0.79	0.69	0.64
		Classification Accuracy	0.92	0.93	0.98	0.83
	Male	Classification Consistency (P)	0.89	0.90	0.97	0.76
		Probability of Chance	0.57	0.52	0.89	0.30
		Kappa (k)	0.75	0.80	0.71	0.66
		Classification Accuracy	0.93	0.94	0.98	0.84
Race/Ethnicity	White	Classification Consistency (P)	0.90	0.89	0.96	0.75
		Probability of Chance	0.66	0.50	0.87	0.31
		Kappa (k)	0.71	0.77	0.69	0.63
		Classification Accuracy	0.93	0.93	0.97	0.83
	African-American	Classification Consistency (P)	0.86	0.96	0.99	0.81
		Probability of Chance	0.57	0.84	0.99	0.54
		Kappa (k)	0.67	0.73	0.63	0.60
		Classification Accuracy	0.90	0.97	0.99	0.87
	Hispanic	Classification Consistency (P)	0.85	0.92	0.99	0.76
		Probability of Chance	0.50	0.68	0.97	0.37
		Kappa (k)	0.70	0.76	0.69	0.62
		Classification Accuracy	0.90	0.95	0.99	0.84
	Asian	Classification Consistency (P)	0.88	0.91	0.96	0.75
		Probability of Chance	0.59	0.51	0.83	0.28
		Kappa (k)	0.70	0.81	0.78	0.65
		Classification Accuracy	0.92	0.94	0.97	0.83
	American Indian	Classification Consistency (P)	0.86	0.93	0.99	0.78
		Probability of Chance	0.50	0.69	0.98	0.40
		Kappa (k)	0.72	0.77	0.58	0.63
		Classification Accuracy	0.90	0.95	0.99	0.85
Two or More	Classification Consistency (P)	0.87	0.91	0.98	0.76	
	Probability of Chance	0.53	0.56	0.92	0.31	
	Kappa (k)	0.73	0.79	0.72	0.65	
	Classification Accuracy	0.91	0.94	0.98	0.84	
Limited English Proficiency	Yes	Classification Consistency (P)	0.84	0.97	0.99	0.81
		Probability of Chance	0.60	0.91	0.99	0.58
		Kappa (k)	0.60	0.66	0.75	0.55
		Classification Accuracy	0.89	0.98	0.99	0.87
Disability Status	Yes	Classification Consistency (P)	0.87	0.96	0.99	0.83
		Probability of Chance	0.58	0.82	0.98	0.54
		Kappa (k)	0.69	0.78	0.64	0.62
		Classification Accuracy	0.91	0.97	0.99	0.88

Table J-11 Indexes for Classification Consistency and Accuracy, Mathematics Grade 7

Category	Group	Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Accommodation Use	Yes	Classification Consistency (P)	0.87	0.98	0.99	0.86
		Probability of Chance	0.71	0.94	0.99	0.70
		Kappa (k)	0.57	0.69	0.74	0.53
		Classification Accuracy	0.92	0.99	0.99	0.90
SES Disadvantaged	Yes	Classification Consistency (P)	0.86	0.92	0.99	0.77
		Probability of Chance	0.50	0.67	0.97	0.37
		Kappa (k)	0.71	0.76	0.64	0.63
		Classification Accuracy	0.90	0.95	0.99	0.84

Table J-12 Indexes for Classification Consistency and Accuracy, Mathematics Grade 8

Category	Group	Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Gender	Female	Classification Consistency (P)	0.89	0.90	0.96	0.75
		Probability of Chance	0.61	0.54	0.87	0.30
		Kappa (k)	0.73	0.78	0.71	0.65
		Classification Accuracy	0.93	0.93	0.97	0.83
	Male	Classification Consistency (P)	0.89	0.91	0.96	0.77
		Probability of Chance	0.57	0.56	0.87	0.30
		Kappa (k)	0.75	0.80	0.74	0.67
		Classification Accuracy	0.93	0.94	0.97	0.84
Race/Ethnicity	White	Classification Consistency (P)	0.90	0.89	0.95	0.75
		Probability of Chance	0.67	0.51	0.84	0.30
		Kappa (k)	0.70	0.77	0.71	0.64
		Classification Accuracy	0.94	0.92	0.97	0.82
	African-American	Classification Consistency (P)	0.86	0.96	0.99	0.82
		Probability of Chance	0.55	0.86	0.99	0.51
		Kappa (k)	0.70	0.74	0.69	0.64
		Classification Accuracy	0.91	0.97	0.99	0.88
	Hispanic	Classification Consistency (P)	0.85	0.93	0.99	0.77
		Probability of Chance	0.50	0.72	0.96	0.37
		Kappa (k)	0.71	0.76	0.68	0.64
		Classification Accuracy	0.90	0.95	0.99	0.84
	Asian	Classification Consistency (P)	0.89	0.90	0.95	0.74
		Probability of Chance	0.67	0.52	0.83	0.30
		Kappa (k)	0.67	0.78	0.73	0.63
		Classification Accuracy	0.92	0.92	0.96	0.80
	American Indian	Classification Consistency (P)	0.85	0.94	0.99	0.78
		Probability of Chance	0.50	0.74	0.97	0.39
		Kappa (k)	0.70	0.76	0.68	0.64
		Classification Accuracy	0.90	0.96	0.99	0.85
Two or More	Classification Consistency (P)	0.88	0.92	0.97	0.76	
	Probability of Chance	0.56	0.58	0.88	0.31	
	Kappa (k)	0.72	0.80	0.72	0.65	
	Classification Accuracy	0.92	0.94	0.98	0.83	
Limited English Proficiency	Yes	Classification Consistency (P)	0.84	0.97	0.99	0.81
		Probability of Chance	0.56	0.91	0.99	0.53
		Kappa (k)	0.64	0.66	0.59	0.59
		Classification Accuracy	0.89	0.98	0.99	0.87
Disability Status	Yes	Classification Consistency (P)	0.86	0.97	0.99	0.83
		Probability of Chance	0.57	0.88	0.98	0.54
		Kappa (k)	0.68	0.76	0.77	0.63
		Classification Accuracy	0.91	0.98	0.99	0.88



Table J-12 Indexes for Classification Consistency and Accuracy, Mathematics Grade 8 (cont.)

Category	Group	Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Accommodation Use	Yes	Classification Consistency (P)	0.87	0.99	0.99	0.86
		Probability of Chance	0.68	0.96	0.99	0.67
		Kappa (k)	0.59	0.66	0.12	0.56
		Classification Accuracy	0.91	0.99	0.99	0.90
SES Disadvantaged	Yes	Classification Consistency (P)	0.86	0.93	0.99	0.78
		Probability of Chance	0.50	0.72	0.96	0.37
		Kappa (k)	0.72	0.75	0.66	0.65
		Classification Accuracy	0.91	0.95	0.99	0.85

Table J-13 Indexes for Classification Consistency and Accuracy, Science Grade 4

Category	Group	Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Gender	Female	Classification Consistency (P)	0.93	0.88	0.88	0.69
		Probability of Chance	0.74	0.50	0.71	0.28
		Kappa (k)	0.72	0.75	0.60	0.56
		Classification Accuracy	0.95	0.91	0.92	0.78
	Male	Classification Consistency (P)	0.93	0.88	0.89	0.70
		Probability of Chance	0.73	0.50	0.69	0.27
		Kappa (k)	0.73	0.76	0.64	0.58
		Classification Accuracy	0.95	0.91	0.92	0.78
Race/Ethnicity	White	Classification Consistency (P)	0.98	0.94	0.93	0.85
		Probability of Chance	0.84	0.61	0.50	0.35
		Kappa (k)	0.85	0.86	0.87	0.78
		Classification Accuracy	0.98	0.88	0.70	0.57
	African-American	Classification Consistency (P)	0.86	0.92	0.97	0.75
		Probability of Chance	0.50	0.72	0.93	0.37
		Kappa (k)	0.71	0.72	0.56	0.61
		Classification Accuracy	0.90	0.94	0.98	0.82
	Hispanic	Classification Consistency (P)	0.89	0.88	0.94	0.71
		Probability of Chance	0.61	0.56	0.85	0.31
		Kappa (k)	0.70	0.73	0.58	0.58
		Classification Accuracy	0.92	0.92	0.96	0.79
	Asian	Classification Consistency (P)	0.91	0.88	0.91	0.70
		Probability of Chance	0.70	0.51	0.73	0.28
		Kappa (k)	0.69	0.75	0.68	0.58
		Classification Accuracy	0.93	0.91	0.94	0.79
	American Indian	Classification Consistency (P)	0.88	0.89	0.93	0.70
		Probability of Chance	0.63	0.57	0.84	0.31
		Kappa (k)	0.67	0.74	0.58	0.56
		Classification Accuracy	0.91	0.92	0.95	0.78
Two or More	Classification Consistency (P)	0.92	0.88	0.89	0.69	
	Probability of Chance	0.71	0.50	0.73	0.28	
	Kappa (k)	0.72	0.75	0.61	0.57	
	Classification Accuracy	0.94	0.91	0.92	0.78	
Limited English Proficiency	Yes	Classification Consistency (P)	0.85	0.90	0.98	0.73
		Probability of Chance	0.54	0.71	0.96	0.37
		Kappa (k)	0.68	0.64	0.46	0.57
		Classification Accuracy	0.90	0.93	0.98	0.81
Disability Status	Yes	Classification Consistency (P)	0.87	0.91	0.95	0.73
		Probability of Chance	0.55	0.61	0.86	0.31
		Kappa (k)	0.72	0.76	0.63	0.60
		Classification Accuracy	0.91	0.93	0.96	0.81
SES Disadvantaged	Yes	Classification Consistency (P)	0.89	0.88	0.93	0.71
		Probability of Chance	0.61	0.55	0.83	0.30
		Kappa (k)	0.72	0.74	0.59	0.58
		Classification Accuracy	0.92	0.92	0.95	0.79

Note: Classification consistency and accuracy not computed for students with accommodations due to N-count < 50.

Table J-14 Indexes for Classification Consistency and Accuracy, Science Grade 8

Category	Group	Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Gender	Female	Classification Consistency (P)	0.93	0.86	0.89	0.69
		Probability of Chance	0.73	0.50	0.74	0.28
		Kappa (k)	0.75	0.72	0.58	0.56
		Classification Accuracy	0.95	0.90	0.92	0.77
	Male	Classification Consistency (P)	0.93	0.87	0.90	0.70
		Probability of Chance	0.68	0.50	0.73	0.27
		Kappa (k)	0.78	0.75	0.61	0.59
		Classification Accuracy	0.95	0.91	0.93	0.78
Race/Ethnicity	White	Classification Consistency (P)	0.95	0.86	0.87	0.68
		Probability of Chance	0.80	0.51	0.69	0.29
		Kappa (k)	0.73	0.71	0.59	0.54
		Classification Accuracy	0.96	0.89	0.91	0.76
	African-American	Classification Consistency (P)	0.87	0.92	0.98	0.78
		Probability of Chance	0.50	0.75	0.95	0.40
		Kappa (k)	0.75	0.69	0.54	0.63
		Classification Accuracy	0.91	0.94	0.98	0.84
	Hispanic	Classification Consistency (P)	0.90	0.88	0.95	0.72
		Probability of Chance	0.58	0.59	0.88	0.32
		Kappa (k)	0.75	0.70	0.55	0.59
		Classification Accuracy	0.93	0.91	0.96	0.80
	Asian	Classification Consistency (P)	0.93	0.87	0.90	0.70
		Probability of Chance	0.70	0.50	0.74	0.28
		Kappa (k)	0.76	0.74	0.62	0.59
		Classification Accuracy	0.95	0.90	0.93	0.78
	American Indian	Classification Consistency (P)	0.93	0.89	0.93	0.75
		Probability of Chance	0.59	0.52	0.79	0.27
		Kappa (k)	0.84	0.77	0.67	0.66
		Classification Accuracy	0.95	0.86	0.92	0.73
Two or More	Classification Consistency (P)	0.92	0.87	0.91	0.70	
	Probability of Chance	0.67	0.51	0.76	0.28	
	Kappa (k)	0.76	0.73	0.62	0.58	
	Classification Accuracy	0.94	0.90	0.93	0.77	
Limited English Proficiency	Yes	Classification Consistency (P)	0.86	0.94	0.99	0.79
		Probability of Chance	0.51	0.86	0.99	0.45
		Kappa (k)	0.71	0.57	0.37	0.62
		Classification Accuracy	0.90	0.96	0.99	0.85
Disability Status	Yes	Classification Consistency (P)	0.88	0.93	0.97	0.78
		Probability of Chance	0.50	0.74	0.93	0.39
		Kappa (k)	0.75	0.74	0.62	0.64
		Classification Accuracy	0.91	0.95	0.98	0.84
SES Disadvantaged	Yes	Classification Consistency (P)	0.90	0.88	0.94	0.72
		Probability of Chance	0.57	0.58	0.87	0.31
		Kappa (k)	0.76	0.72	0.57	0.60
		Classification Accuracy	0.93	0.91	0.96	0.80

Note: Classification consistency and accuracy not computed for students with accommodations due to N-count < 50.

Table J-15 Indexes for Classification Consistency and Accuracy, Social Studies Grade 4

Category	Group	Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Gender	Female	Classification Consistency (P)	0.91	0.87	0.88	0.67
		Probability of Chance	0.65	0.50	0.65	0.25
		Kappa (k)	0.75	0.74	0.66	0.56
		Classification Accuracy	0.94	0.91	0.91	0.76
	Male	Classification Consistency (P)	0.91	0.88	0.89	0.68
		Probability of Chance	0.63	0.50	0.65	0.25
		Kappa (k)	0.76	0.75	0.68	0.57
		Classification Accuracy	0.93	0.91	0.92	0.76
Race/Ethnicity	White	Classification Consistency (P)	0.93	0.87	0.86	0.66
		Probability of Chance	0.75	0.53	0.59	0.27
		Kappa (k)	0.70	0.72	0.65	0.54
		Classification Accuracy	0.95	0.90	0.90	0.75
	African-American	Classification Consistency (P)	0.88	0.91	0.96	0.76
		Probability of Chance	0.51	0.69	0.90	0.41
		Kappa (k)	0.75	0.72	0.63	0.60
		Classification Accuracy	0.91	0.94	0.97	0.82
	Hispanic	Classification Consistency (P)	0.87	0.87	0.93	0.68
		Probability of Chance	0.53	0.55	0.81	0.29
		Kappa (k)	0.73	0.71	0.64	0.56
		Classification Accuracy	0.91	0.91	0.95	0.76
	Asian	Classification Consistency (P)	0.89	0.87	0.91	0.68
		Probability of Chance	0.60	0.51	0.69	0.26
		Kappa (k)	0.73	0.74	0.70	0.57
		Classification Accuracy	0.92	0.91	0.93	0.76
	American Indian	Classification Consistency (P)	0.85	0.88	0.94	0.68
		Probability of Chance	0.52	0.58	0.81	0.30
		Kappa (k)	0.68	0.72	0.68	0.54
		Classification Accuracy	0.90	0.92	0.96	0.77
	Two or More	Classification Consistency (P)	0.89	0.87	0.90	0.67
		Probability of Chance	0.62	0.50	0.68	0.25
		Kappa (k)	0.72	0.74	0.68	0.56
		Classification Accuracy	0.93	0.91	0.93	0.76
Limited English Proficiency	Yes	Classification Consistency (P)	0.84	0.89	0.97	0.71
		Probability of Chance	0.50	0.70	0.94	0.37
		Kappa (k)	0.69	0.62	0.51	0.53
		Classification Accuracy	0.88	0.92	0.98	0.78
Disability Status	Yes	Classification Consistency (P)	0.88	0.90	0.95	0.74
		Probability of Chance	0.50	0.61	0.84	0.34
		Kappa (k)	0.76	0.75	0.67	0.60
		Classification Accuracy	0.91	0.93	0.96	0.80
SES Disadvantaged	Yes	Classification Consistency (P)	0.88	0.88	0.93	0.69
		Probability of Chance	0.53	0.55	0.80	0.28
		Kappa (k)	0.75	0.73	0.64	0.57
		Classification Accuracy	0.91	0.91	0.95	0.77

Note: Classification consistency and accuracy not computed for students with accommodations due to N-count < 50.

Table J-16 Indexes for Classification Consistency and Accuracy, Social Studies Grade 8

Category	Group	Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Gender	Female	Classification Consistency (P)	0.93	0.88	0.89	0.71
		Probability of Chance	0.67	0.50	0.67	0.26
		Kappa (k)	0.79	0.77	0.65	0.60
		Classification Accuracy	0.94	0.92	0.92	0.78
	Male	Classification Consistency (P)	0.93	0.90	0.90	0.72
		Probability of Chance	0.62	0.50	0.67	0.25
		Kappa (k)	0.81	0.79	0.68	0.63
		Classification Accuracy	0.94	0.93	0.93	0.79
Race/Ethnicity	White	Classification Consistency (P)	0.94	0.88	0.87	0.70
		Probability of Chance	0.73	0.51	0.62	0.26
		Kappa (k)	0.77	0.76	0.66	0.59
		Classification Accuracy	0.95	0.92	0.91	0.78
	African-American	Classification Consistency (P)	0.89	0.93	0.97	0.80
		Probability of Chance	0.51	0.72	0.92	0.42
		Kappa (k)	0.78	0.76	0.58	0.65
		Classification Accuracy	0.91	0.95	0.98	0.84
	Hispanic	Classification Consistency (P)	0.90	0.89	0.94	0.73
		Probability of Chance	0.54	0.57	0.84	0.29
		Kappa (k)	0.78	0.74	0.61	0.62
		Classification Accuracy	0.92	0.92	0.96	0.80
	Asian	Classification Consistency (P)	0.92	0.88	0.90	0.70
		Probability of Chance	0.65	0.50	0.67	0.25
		Kappa (k)	0.77	0.76	0.70	0.60
		Classification Accuracy	0.94	0.91	0.93	0.78
	American Indian	Classification Consistency (P)	0.88	0.89	0.94	0.71
		Probability of Chance	0.53	0.59	0.85	0.30
		Kappa (k)	0.74	0.74	0.61	0.59
		Classification Accuracy	0.91	0.93	0.96	0.79
	Two or More	Classification Consistency (P)	0.90	0.89	0.90	0.70
		Probability of Chance	0.60	0.51	0.71	0.26
		Kappa (k)	0.76	0.78	0.66	0.60
		Classification Accuracy	0.93	0.93	0.93	0.79
Limited English Proficiency	Yes	Classification Consistency (P)	0.87	0.94	0.99	0.81
		Probability of Chance	0.53	0.84	0.98	0.48
		Kappa (k)	0.73	0.60	0.44	0.63
		Classification Accuracy	0.90	0.96	0.99	0.85
Disability Status	Yes	Classification Consistency (P)	0.89	0.94	0.97	0.81
		Probability of Chance	0.52	0.74	0.92	0.44
		Kappa (k)	0.78	0.77	0.64	0.66
		Classification Accuracy	0.91	0.96	0.98	0.85
SES Disadvantaged	Yes	Classification Consistency (P)	0.90	0.90	0.94	0.74
		Probability of Chance	0.52	0.57	0.83	0.30
		Kappa (k)	0.79	0.76	0.62	0.63
		Classification Accuracy	0.92	0.93	0.96	0.80

Note: Classification consistency and accuracy not computed for students with accommodations due to N-count < 50.

Table J-17 Indexes for Classification Consistency and Accuracy, Social Studies Grade 10

Category	Group	Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Gender	Female	Classification Consistency (P)	0.91	0.88	0.90	0.70
		Probability of Chance	0.61	0.50	0.68	0.25
		Kappa (k)	0.77	0.76	0.70	0.60
		Classification Accuracy	0.93	0.91	0.93	0.78
	Male	Classification Consistency (P)	0.92	0.90	0.90	0.72
		Probability of Chance	0.60	0.50	0.63	0.25
		Kappa (k)	0.80	0.80	0.74	0.63
		Classification Accuracy	0.94	0.93	0.93	0.80
Race/Ethnicity	White	Classification Consistency (P)	0.92	0.88	0.89	0.70
		Probability of Chance	0.67	0.51	0.61	0.25
		Kappa (k)	0.76	0.76	0.71	0.60
		Classification Accuracy	0.94	0.92	0.92	0.78
	African-American	Classification Consistency (P)	0.89	0.93	0.97	0.80
		Probability of Chance	0.53	0.73	0.92	0.46
		Kappa (k)	0.76	0.75	0.64	0.63
		Classification Accuracy	0.92	0.95	0.98	0.85
	Hispanic	Classification Consistency (P)	0.88	0.89	0.95	0.73
		Probability of Chance	0.51	0.58	0.82	0.31
		Kappa (k)	0.76	0.74	0.70	0.61
		Classification Accuracy	0.92	0.92	0.96	0.80
	Asian	Classification Consistency (P)	0.91	0.89	0.92	0.72
		Probability of Chance	0.61	0.50	0.65	0.25
		Kappa (k)	0.76	0.77	0.76	0.62
		Classification Accuracy	0.94	0.91	0.94	0.79
	American Indian	Classification Consistency (P)	0.88	0.91	0.95	0.74
		Probability of Chance	0.51	0.58	0.84	0.32
		Kappa (k)	0.76	0.78	0.67	0.61
		Classification Accuracy	0.92	0.93	0.96	0.82
Two or More	Classification Consistency (P)	0.91	0.89	0.92	0.72	
	Probability of Chance	0.59	0.50	0.67	0.25	
	Kappa (k)	0.78	0.77	0.75	0.62	
	Classification Accuracy	0.94	0.92	0.94	0.79	
Limited English Proficiency	Yes	Classification Consistency (P)	0.88	0.96	0.99	0.84
		Probability of Chance	0.64	0.89	0.98	0.62
		Kappa (k)	0.67	0.66	0.54	0.58
		Classification Accuracy	0.92	0.97	0.99	0.88
Disability Status	Yes	Classification Consistency (P)	0.89	0.94	0.97	0.81
		Probability of Chance	0.54	0.73	0.90	0.47
		Kappa (k)	0.76	0.79	0.73	0.64
		Classification Accuracy	0.92	0.96	0.98	0.86
SES Disadvantaged	Yes	Classification Consistency (P)	0.89	0.90	0.94	0.74
		Probability of Chance	0.51	0.58	0.82	0.31
		Kappa (k)	0.77	0.76	0.69	0.62
		Classification Accuracy	0.92	0.93	0.96	0.81

Note: Classification consistency and accuracy not computed for students with accommodations due to N-count < 50.

**Appendix K**  
**Glossary**

## **Glossary: Abbreviations most commonly used in the Wisconsin Forward Exam Technical Report**

**2PPC:** Two-parameter partial-credit item response theory model. A mathematical model that shows the relationship between student achievement on a test and the discrimination and difficulty of score points for a constructed-response item.

**3PL:** Three-parameter logistic item response theory model. A mathematical model that shows the relationship between student achievement on a test and a single multiple-choice item by decomposing the item into three components: difficulty, discrimination, and guessing.

**AERA:** American Education Research Association. A professional organization whose purpose is to advance the science of educational research and its application.

**APA:** American Psychological Association. A professional organization centered in psychology.

**CCR:** College- and Career Ready item bank. Items measuring knowledge and skills in English Language Arts and Mathematics necessary to prepare students for college and the workplace.

**CR:** Constructed-response item. A type of question, designed to elicit student knowledge of content, that typically comprises a question for which students create (write) a response.

**DIF:** Differential item functioning. The degree to which an item performs differently for one group of examinees than it performs for another group of equally able examinees. Refers to differential statistical properties of an item in two equally able groups.

**DOK:** Depth of knowledge. A system of describing the cognitive level a test item elicits from a student. Items are coded such that level 1 indicates students use lower cognitive levels, such as recall, to answer the item correctly; level 4 indicates students use higher cognitive levels, such as analysis skills, to answer the item correctly.

**DPI:** Wisconsin Department of Public Instruction. The state agency overseeing the implementation of federal and state laws related to public education in Wisconsin.

**DRC:** Data Recognition Corporation. A testing company partnering with DPI for delivery, scoring, and reporting of Wisconsin Forward Exam assessments.

**ELA:** English Language Arts. A content area in the Wisconsin Forward Exam.

**ELP:** English language proficiency. A student population subgroup category describing students for whom English is a second language. Students are described as fully English proficient or limited English proficient.

**HOSS:** Highest obtainable scale score. The highest possible scale score on a test.

**IRT:** Item response theory. A mathematic model that shows the relationship between



student achievement on a test and the performance on a test item.

LOSS: Lowest obtainable scale score. The lowest possible scale score on a test.

MA: Mathematics. A content area in the Wisconsin Forward Exam.

MC: Multiple-choice item. A type of question, designed to elicit student knowledge of content, that typically comprises a stem and four options. Students must select the correct option.

MH: Mantel-Haenszel ( $MH_{2MH}\chi$ ) statistic. A commonly used DIF statistic for multiple-choice items.

NCME: National Council on Measurement in Education. A professional organization centered in assessment, evaluation, testing, and educational measurement.

OP: Operational item. An item that has previously undergone field testing and contributes to a student's score in a specific content area on the Wisconsin Forward Exam.

OTTs: Online Training Tools. Provided for students to allow them a hands-on opportunity to practice answering the types of items and using the tools available in the online testing system.

SC: Science. A content area in the Wisconsin Forward Exam.

SD: Standard deviation. A measure of the variability of observations from the mean.

SEM: Standard error of measurement. An estimate of how repeated measures of a person on the same test tend to be distributed around his or her "true" score.

SES: Socioeconomic status. A student population subgroup category describing students as economically disadvantaged or not economically disadvantaged.

SMD: Standardized mean difference. A commonly used DIF statistic for constructed-response items.

SPI: Standard performance index. A content category reporting score based on items from a single content standard or domain within a given content area.

SS: Social Studies. A content area in the Wisconsin Forward Exam.

TDA: Text-dependent analysis. An item based on a passage or a multiple-passage set that each student has read during the assessment. Students must draw on basic writing skills while inferring and synthesizing information from the passage in order to develop a comprehensive, holistic essay response.

TCC: Test characteristic curve. Shows the mathematical relationship between students with varying degrees of achievement and their estimated overall test performance.

WKCE: Wisconsin Knowledge and Concepts Examination. Previous Wisconsin assessment program.