

Wisconsin Forward Exam

Technical Report 2018



Submitted to
Wisconsin Department of Public Instruction
November 2018



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 © 2018 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	2
Foreword	3
Part 1: Overview	12
1.1 HISTORICAL BACKGROUND	12
1.2 USES OF TEST SCORES	14
1.2.1 TEST-LEVEL SCORES	15
1.2.2 STANDARD-LEVEL SUBSCORES AND PERFORMANCE LEVELS	16
1.3 TECHNICAL REPORT STRUCTURE.....	17
Part 2: Test Blueprint and Item Development	20
2.1 TEST BLUEPRINTS	21
2.2 READING PASSAGE AND ITEM SELECTION FOR SPRING 2017 FIELD TEST	22
2.3 FIELD TESTING	22
2.4 STATISTICAL ANALYSIS OF SPRING 2017 FIELD TEST DATA	23
2.5 REVIEW OF ITEMS WITH DATA.....	23
2.6 SUMMARY	24
Part 3: Test Form Development	31
3.1 DESIGN OF THE WISCONSIN FORWARD EXAM	31
3.1.1 ENGLISH LANGUAGE ARTS	31
3.1.2 MATHEMATICS	31
3.1.3 SCIENCE.....	32
3.1.4 SOCIAL STUDIES	32
3.2 TEST DEVELOPMENT PROCESS	32
3.2.1 WISCONSIN FORWARD TEST FORM CREATION	33
3.2.2 ITEM SELECTION	33
3.2.3 ITEM AND FORM QUALITY REVIEWS.....	35
3.3 DPI APPROVALS	36
3.4 SUMMARY	36
Part 4: Test Administration.....	40
4.1 ACCESSIBILITY RESOURCES.....	41
4.1.1 UNIVERSAL TOOLS	41
4.1.2 DESIGNATED SUPPORTS.....	42
4.1.3 ACCOMMODATIONS	43
4.1.4 TRANSLATION.....	43
4.1.5 ADDITIONAL ACCESSIBILITY RESOURCES.....	44
4.2 REPORTING RESULTS OF ASSESSMENTS TAKEN WITH ACCOMMODATIONS	44
4.3 TEST SECURITY.....	44
4.3.1 SECURE STUDENT ACCESS.....	45
4.3.2 TEST SECURITY DURING BREAKS.....	46
4.4 TEST ADMINISTRATION TRAINING.....	46
4.5 SUMMARY	49
Part 5: Scoring	60
5.1 MULTIPLE-CHOICE AND MULTI-SELECT ITEM SCORING PROCESS.....	60
5.2 TECHNOLOGY-ENHANCED, SHORT-ANSWER, AND EVIDENCE-BASED SELECTED RESPONSE ITEM SCORING PROCESS	60
5.3 SCORING OF TEXT-DEPENDENT ANALYSIS ITEMS.....	61
5.3.1 DESCRIPTION OF SCORING RUBRICS AND NON-SCORE CODES.....	61
5.3.3 HANDSCORING PROCESS.....	63
5.3.4 HANDSCORING SYSTEM.....	63
5.3.5 ANCHOR PAPERS AND TRAINING PAPERS	63

5.3.6 SCORING PERSONNEL AND QUALIFICATIONS	64
5.3.7 SCORER TRAINING	65
5.3.8 MONITORING THE SCORING PROCESS	65
5.3.9 FINAL SCORES	65
5.4 INTER-RATER RELIABILITY	66
5.4.1 DISTRIBUTION OF TDA ITEM SCORES	66
5.5 SUMMARY	66
Part 6: Calibration, Equating, and Deriving Scale Scores	72
6.1 ITEM CALIBRATION	72
6.1.1 CALIBRATION MODELS	72
6.1.2 CALIBRATION SAMPLE	74
6.1.3 CALIBRATION PROCEDURE	74
6.1.4 CALIBRATION SOFTWARE	74
6.1.5 CALIBRATION RESULTS	75
6.2 TEST EQUATING	77
6.2.1 EVALUATION OF ANCHOR ITEMS	78
6.2.2 REMOVAL OF ANCHOR ITEMS	79
6.2.3 EVALUATION OF EQUATING RESULTS	80
6.2.4 TEST SCALES	80
6.3 DERIVING SCALE SCORES IN THE WISCONSIN FORWARD EXAM	84
6.3.1 CONDITIONAL STANDARD ERROR OF MEASUREMENT	88
6.3.2 LOSS AND HOSS	89
6.4 SUMMARY	90
Part 7: Standard Setting	141
7.1 BACKGROUND INFORMATION	141
7.2 STANDARD SETTING METHODOLOGY	142
7.3 PERFORMANCE LEVEL DESCRIPTORS	142
7.4 CUT SCORES	142
7.5 SUMMARY	143
Part 8: Test Results	145
8.1 CLASSICAL ITEM ANALYSIS: ITEM LEVEL STATISTICS	145
8.1.1 FLAGGING FOR A POSITIVE DISTRACTOR CORRELATION	148
8.1.2 FLAGGING FOR THE ITEM-TOTAL CORRELATION	148
8.1.3 FLAGGING FOR P-VALUE	148
8.1.4 FLAGGING FOR OMIT RATE	148
8.1.5 SPEEDEDNESS	148
8.1.6 SUPPLEMENTAL TABLES ON CLASSICAL ITEM ANALYSIS	149
8.2 RAW SCORE RESULTS	149
8.2.1 SUBGROUP PERFORMANCE PATTERNS IN RAW SCORE RESULTS	151
8.3 SUMMARY STATISTICS FOR SCALE SCORES	153
8.3.1 SUBGROUP PERFORMANCE PATTERNS IN SCALE SCORE RESULTS	154
8.4 CUT SCORES AND PERFORMANCE LEVEL CLASSIFICATIONS	156
8.5 STANDARD PERFORMANCE INDEX FOR CONTENT STANDARDS	158
8.6 LONGITUDINAL COMPARISONS OF TEST SCORES	161
8.7 SUMMARY	162
Part 9: Reliability	240
9.1 MEASURES OF INTERNAL CONSISTENCY AND STANDARD ERROR OF MEASUREMENT	242
9.1.1 CONDITIONAL STANDARD ERROR OF MEASUREMENT	244
9.2 CLASSIFICATION CONSISTENCY AND ACCURACY	245
9.2.1 KOLEN AND KIM'S METHOD FOR PATTERN SCORING	246
9.3 INTER-RATER RELIABILITY FOR TDA ITEMS	249
9.4 SUMMARY	252
Part 10: Validity	267

10.1 DIFFERENTIAL ITEM FUNCTIONING.....	271
10.2 VALIDITY EVIDENCE BASED ON INTERNAL TEST STRUCTURE.....	275
10.2.1 CORRELATIONS BETWEEN CONTENT STANDARDS	275
10.2.2 PRINCIPAL COMPONENT ANALYSIS	276
10.3 VALIDITY EVIDENCE BASED ON RELATIONSHIP WITH OTHER VARIABLES	277
10.3.1 CORRELATIONS BETWEEN CONTENT AREA TEST SCORES	277
10.3.2 COMPARISON OF THE WISCONSIN FORWARD EXAM AND WISCONSIN NAEP IMPACT DATA	278
10.4 TEST INTEGRITY: DATA FORENSIC ANALYSES	280
10.5 STANDARDIZED TEST ADMINISTRATION.....	280
10.6 SUMMARY.....	281
Part 11: Summary Recommendations	299
References	300

APPENDICES

Appendix A: Spring 2016 Item Review Training Slides	304
Appendix B: Spring 2017 Field Test Data Review Training Slides.....	331
Appendix C: Spring 2018 English Language Arts Operational Test Maps.....	354
Appendix D: Spring 2018 Mathematics Operational Test Maps.....	367
Appendix E: Spring 2018 Science Operational Test Maps.....	380
Appendix F: Spring 2018 Social Studies Operational Test Maps.....	385
Appendix G: Classical Item Analysis Results.....	392
Appendix H: Wisconsin Standard Performance Index Score Computation.....	428
Appendix I: Conditional Standard Error of Measurement with Cut Scores.....	436
Appendix J: Classification Consistency and Accuracy Indices by Subgroup	454
Appendix K: Glossary.....	484

LIST OF TABLES

PART 2

Table 2-1 College- and Career-Ready Item Bank Development Activities	25
Table 2-2 Item Type Descriptions for Items on the Wisconsin Forward Exam	26
Table 2-3 English Language Arts Test Blueprints for Grades 3–8	27
Table 2-4 Mathematics Test Blueprints for Grades 3–8	28
Table 2-5 Science Test Blueprints for Grades 4 and 8	29
Table 2-6 Social Studies Test Blueprints for Grades 4, 8, and 10	29
Table 2-7 Items Reviewed at Summer 2016 Item Review	30
Table 2-8 Items Reviewed at Summer 2017 Data Review	30

PART 3

Table 3-1 English Language Arts Test Design	36
Table 3-2 Mathematics Test Design	38
Table 3-3 Science Test Design	39
Table 3-4 Social Studies Test Design	39

PART 4

Table 4-1 Number and Percentage of Students Using Accommodations or Designated Supports: Grade 3	50
Table 4-2 Number and Percentage of Students Using Accommodations or Designated Supports: Grade 4	51
Table 4-3 Number and Percentage of Students Using Accommodations or Designated Supports: Grade 5	52
Table 4-4 Number and Percentage of Students Using Accommodations or Designated Supports: Grade 6	53
Table 4-5 Number and Percentage of Students Using Accommodations or Designated Supports: Grade 7	54
Table 4-6 Number and Percentage of Students Using Accommodations or Designated Supports: Grade 8	55
Table 4-7 Number and Percentage of Students Using Accommodations or Designated Supports: Grade 10	56
Table 4-8 Summary Table of Manual Materials	57

PART 5

Table 5-1 TDA Item Scoring Guidelines, Grades 3–8	67
Table 5-2 TDA Item Non-scorable Codes, Grades 3–8	69
Table 5-3 TDA Item Score Distribution	70
Table 5-4 TDA Item Score Distribution: AI Engine vs. Human Scorer	70
Table 5-5 TDA Item Percentage Score Distribution: AI Engine vs. Human Scorer	71

PART 6

Table 6-A Example of Item Parameters for a Test	85
Table 6-B Example of Item Response Pattern	86
Table 6-1 English Language Arts Calibration Sample Demographics Compared to Population	91
Table 6-2 Mathematics Calibration Sample Demographics Compared to Population	94
Table 6-3 Science Calibration Sample Demographics Compared to Population	97
Table 6-4 Social Studies Calibration Sample Demographics Compared to Population	98
Table 6-5 Item Flagged Based on Yen’s Q1	100
Table 6-6 Equating Evaluation Results, Stocking and Lord Method	101
Table 6-7 Scale Transformation Constants	101
Table 6-8 Scoring Table for English Language Arts Grade 3	102
Table 6-9 Scoring Table for English Language Arts Grade 4	103
Table 6-10 Scoring Table for English Language Arts Grade 5	104
Table 6-11 Scoring Table for English Language Arts Grade 6	105
Table 6-12 Scoring Table for English Language Arts Grade 7	106

Table 6-13 Scoring Table for English Language Arts Grade 8	107
Table 6-14 Scoring Table for Mathematics Grade 3.....	108
Table 6-15 Scoring Table for Mathematics Grade 4.....	109
Table 6-16 Scoring Table for Mathematics Grade 5.....	110
Table 6-17 Scoring Table for Mathematics Grade 6.....	111
Table 6-18 Scoring Table for Mathematics Grade 7.....	112
Table 6-19 Scoring Table for Mathematics Grade 8.....	113
Table 6-20 Scoring Table for Science Grade 4.....	114
Table 6-21 Scoring Table for Science Grade 8.....	115
Table 6-22 Scoring Table for Social Studies Grade 4	116
Table 6-23 Scoring Table for Social Studies Grade 8	117
Table 6-24 Scoring Table for Social Studies Grade 10	118
Table 6-25 The Number and Percentage of Students at LOSS and HOSS	119

PART 7

Table 7-1 Policy Performance Level Descriptors for the Wisconsin Forward Exam	144
Table 7-2 Wisconsin Forward Exam Cut Scores	144

PART 8

Table 8-A Summary of Flagged Operational Items on the Wisconsin Forward Exam.....	163
Table 8-B English Language Arts Items Flagged for Classical Item Analysis Statistics	164
Table 8-C Mathematics Items Flagged for Classical Item Analysis Statistics	165
Table 8-D Science and Social Studies Items Flagged for Classical Item Analysis Statistics	166
Table 8-E Percentage of Students Attempting Last Operational Item in Test	166
Table 8-1 Item Analysis, Grade 3 English Language Arts	167
Table 8-2 Item Analysis, Grade 4 English Language Arts	168
Table 8-4 Item Analysis, Grade 6 English Language Arts	170
Table 8-6 Item Analysis, Grade 8 English Language Arts	172
Table 8-7 Item Analysis, Grade 3 Mathematics	173
Table 8-8 Item Analysis, Grade 4 Mathematics	175
Table 8-9 Item Analysis, Grade 5 Mathematics	177
Table 8-10 Item Analysis, Grade 6 Mathematics	179
Table 8-11 Item Analysis, Grade 7 Mathematics	181
Table 8-12 Item Analysis, Grade 8 Mathematics	183
Table 8-13 Item Analysis, Grade 4 Science.....	185
Table 8-14 Item Analysis, Grade 8 Science.....	187
Table 8-15 Item Analysis, Grade 4 Social Studies	189
Table 8-16 Item Analysis, Grade 8 Social Studies	190
Table 8-17 Item Analysis, Grade 10 Social Studies	191
Table 8-18 Raw Score Descriptive Statistics.....	193
Table 8-19 Raw Score Descriptive Statistics by Gender	194
Table 8-20 Raw Score Descriptive Statistics for English Language Arts by Race/Ethnicity	195
Table 8-21 Raw Score Descriptive Statistics for Mathematics by Race/Ethnicity	196
Table 8-22 Raw Score Descriptive Statistics for Science by Race/Ethnicity	197
Table 8-23 Raw Score Descriptive Statistics for Social Studies by Race/Ethnicity	197
Table 8-24 Raw Score Descriptive Statistics by Socioeconomic Status.....	198
Table 8-25 Raw Score Descriptive Statistics by Disability	199
Table 8-26 Raw Score Descriptive Statistics by English Language Proficiency.....	200
Table 8-27 Raw Score Descriptive Statistics by Accommodation Use	201
Table 8-28 Scale Score Descriptive Statistics	202
Table 8-29 Scale Score Descriptive Statistics by Gender.....	203
Table 8-30 Scale Score Descriptive Statistics for English Language Arts by Race/Ethnicity.....	204
Table 8-31 Scale Score Descriptive Statistics for Mathematics by Race/Ethnicity.....	205
Table 8-32 Scale Score Descriptive Statistics for Science by Race/Ethnicity	206

Table 8-33 Scale Score Descriptive Statistics for Social Studies by Race/Ethnicity	206
Table 8-34 Scale Score Descriptive Statistics by Socioeconomic Status	207
Table 8-35 Scale Score Descriptive Statistics by Disability	208
Table 8-36 Scale Score Descriptive Statistics by English Language Proficiency	209
Table 8-37 Scale Score Descriptive Statistics by Accommodation Use	210
Table 8-38 Performance Level Cut Scores for All Contents	211
Table 8-39 Cut Scores and Associated Impact Data, English Language Arts	211
Table 8-40 Cut Scores and Associated Impact Data, Mathematics	212
Table 8-41 Cut Scores and Associated Impact Data, Science	212
Table 8-42 Cut Scores and Associated Impact Data, Social Studies	213
Table 8-43 Percentage of Students in Each Performance Level by Subgroup, English Language Arts	214
Table 8-44 Percentage of Students in Each Performance Level by Subgroup, Mathematics	216
Table 8-45 Percentage of Students in Each Performance Level by Subgroup, Science	218
Table 8-46 Percentage of Students in Each Performance Level by Subgroup, Social Studies	219
Table 8-47a Summary Statistics for Content Standards Raw and SPI Scores, English Language Arts	220
Table 8-47b Summary Statistics for Domain Raw and SPI Scores, English Language Arts	223
Table 8-48 Summary Statistics for Content Standards Raw and SPI Scores, Mathematics	224
Table 8-49 Summary Statistics for Content Standards Raw and SPI Scores, Science	226
Table 8-50 Summary Statistics for Content Standards Raw and SPI Scores, Social Studies	227
Table 8-51 SPI Cut Scores, English Language Arts	228
Table 8-52 SPI Cut Scores, Mathematics	230
Table 8-53 SPI Cut Scores, Science	232
Table 8-54 SPI Cut Scores, Social Studies	233
Table 8-55 Longitudinal Comparison of State-Level Scale Score Means: ELA	234
Table 8-56 Longitudinal Comparison of State-Level Scale Score Means: Mathematics	235
Table 8-57 Longitudinal Comparison of State-Level Scale Score Means: Science	236
Table 8-58 Longitudinal Comparison of State-Level Scale Score Means: Social Studies	236
Table 8-59 Longitudinal Comparison of State-Level Impact Data: ELA	237
Table 8-60 Longitudinal Comparison of State-Level Impact Data: Mathematics	238
Table 8-61 Longitudinal Comparison of State-Level Impact Data: Science	239
Table 8-62 Longitudinal Comparison of State-Level Impact Data: Social Studies	239

PART 9

Table 9-1 Reliability for Total Group and Subgroups Using Cronbach’s Alpha	253
Table 9-2 Standard Error of Measurement for Total Group and Subgroups	254
Table 9-3 Cronbach’s Alpha Reliability Coefficients for Content Standard and Domain	255
Table 9-4 Standard Error of Measurement per Content Standard and Domain	256
Table 9-5 Classification Consistency and Classification Accuracy for English Language Arts Grade 3	257
Table 9-6 Classification Consistency and Classification Accuracy for English Language Arts Grade 4	257
Table 9-7 Classification Consistency and Classification Accuracy for English Language Arts Grade 5	258
Table 9-8 Classification Consistency and Classification Accuracy for English Language Arts Grade 6	258
Table 9-9 Classification Consistency and Classification Accuracy for English Language Arts Grade 7	259
Table 9-10 Classification Consistency and Classification Accuracy for English Language Arts Grade 8	259
Table 9-11 Classification Consistency and Classification Accuracy for Mathematics Grade 3	260
Table 9-12 Classification Consistency and Classification Accuracy for Mathematics Grade 4	260
Table 9-13 Classification Consistency and Classification Accuracy for Mathematics Grade 5	261
Table 9-14 Classification Consistency and Classification Accuracy for Mathematics Grade 6	261
Table 9-15 Classification Consistency and Classification Accuracy for Mathematics Grade 7	262
Table 9-16 Classification Consistency and Classification Accuracy for Mathematics Grade 8	262
Table 9-17 Classification Consistency and Classification Accuracy for Science Grade 4	263
Table 9-18 Classification Consistency and Classification Accuracy for Science Grade 8	263
Table 9-19 Classification Consistency and Classification Accuracy for Social Studies Grade 4	264
Table 9-20 Classification Consistency and Classification Accuracy for Social Studies Grade 8	264
Table 9-21 Classification Consistency and Classification Accuracy for Social Studies Grade 10	265
Table 9-22 Inter-Rater Reliability, English Language Arts	266

PART 10

Table 10-1 Items Flagged for DIF by Gender, Focal Group: Female282

Table 10-2 Items Flagged for DIF by Race/Ethnicity, Focal Group: African-American283

Table 10-3 Items Flagged for DIF by Race/Ethnicity, Focal Group: Hispanic284

Table 10-4 Items Flagged for DIF by Race/Ethnicity, Focal Group: Asian285

Table 10-5 Items Flagged for DIF by Race/Ethnicity, Focal Group: American Indian286

Table 10-6 Items Flagged for DIF by English Language Proficiency, Focal Group: Students Not English Language Proficient.....286

Table 10-7 Items Flagged for DIF by Socioeconomic Status, Focal Group: Socioeconomically Disadvantaged Students286

Table 10-8 Items Flagged for DIF by Disability Status, Focal Group: Students with One or More Disabilities287

Table 10-9 Items Flagged for DIF by Accommodation Use, Focal Group: Students using Testing Accommodations287

Table 10-10 Correlations among English Language Arts Test Domains.....288

Table 10-11 Correlations among English Language Arts Standards289

Table 10-12 Correlations among Mathematics Standards290

Table 10-13 Correlations among Science Standards291

Table 10-14 Correlations among Social Studies Standards291

Table 10-15 Principal Components Analysis292

Table 10-16 Correlations between Content Area Scale Scores292

Table 10-17 Correlations between Content Area Scale Scores by Gender293

Table 10-18 Correlations between Content Area Scale Scores by Ethnicity/Race294

Table 10-19 Correlations between Content Area Scale Scores by English Proficiency Status295

Table 10-20 Correlations between Content Area Scale Scores by SES Status296

Table 10-21 Correlations between Content Area Scale Scores by Disability Status297

Table 10-22 Partial Correlations between Content Area Scale Scores297

Table 10-23 Comparison of Most Recent Wisconsin NAEP and Spring 2018 Wisconsin Forward Exam Impact Data298

TABLE OF FIGURES

PART 6

Figure 6-A Examples of Likelihood Functions, or the Probability of Each Ability Level Estimate (or Scale Score).....	87
Figure 6-1 Anchor Set TCCs: ELA Grade 3.....	120
Figure 6-2 Anchor Set TCCs: ELA Grade 4.....	120
Figure 6-3 Anchor Set TCCs: ELA Grade 5.....	121
Figure 6-4 Anchor Set TCCs: ELA Grade 6.....	121
Figure 6-5 Anchor Set TCCs: ELA Grade 7.....	122
Figure 6-6 Anchor Set TCCs: ELA Grade 8.....	122
Figure 6-7 Anchor Set TCCs: Mathematics Grade 3.....	123
Figure 6-8 Anchor Set TCCs: Mathematics Grade 4.....	123
Figure 6-9 Anchor Set TCCs: Mathematics Grade 5.....	124
Figure 6-10 Anchor Set TCCs: Mathematics Grade 6.....	124
Figure 6-11 Anchor Set TCCs: Mathematics Grade 7.....	125
Figure 6-12 Anchor Set TCCs: Mathematics Grade 8.....	125
Figure 6-13 Anchor Set TCCs: Science Grade 4.....	126
Figure 6-14 Anchor Set TCCs: Science Grade 8.....	126
Figure 6-15 Anchor Set TCCs: Social Studies Grade 4.....	127
Figure 6-16 Anchor Set TCCs: Social Studies Grade 8.....	127
Figure 6-17 Anchor Set TCCs: Social Studies Grade 10.....	128
Figure 6-18 English Language Arts Test Characteristic Curves.....	129
Figure 6-19 English Language Arts Standard Error Curves.....	130
Figure 6-20 English Language Arts Growth at Quartiles.....	131
Figure 6-21 Mathematics Test Characteristic Curves.....	132
Figure 6-22 Mathematics Standard Error Curves.....	133
Figure 6-23 Mathematics Growth at Quartiles.....	134
Figure 6-24 Science Test Characteristic Curves.....	135
Figure 6-25 Science Standard Error Curves.....	136
Figure 6-26 Science Growth at Quartiles.....	137
Figure 6-27 Social Studies Test Characteristic Curves.....	138
Figure 6-28 Social Studies Standard Error Curves.....	139
Figure 6-29 Social Studies Growth at Quartiles.....	140

Part 1: Overview

The *Wisconsin Forward Exam Spring 2018 Technical Report* documents the processes and procedures applied in the Spring 2018 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 2018 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. At that time, *TerraNova*TM tests developed by CTB/McGraw-Hill (1997, 2000, 2009) were used. 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 2007–08 school year, 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 English Language Arts (ELA), 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. 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, ELA, 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 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 3–8 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 years' assessments included CTB's *TerraNova* items that yielded 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 3–8 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 Summer 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 the Science and Social Studies tests were mainly drawn from the pool of previously administered items and also included 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 Summer 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 2018 was the third administration year for the Wisconsin Forward Exam in ELA, Mathematics, Science, and Social Studies. The ELA, Mathematics, and Social Studies 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 and Spring 2017 as part of the operational test or a field test. Science test forms were reused from the Spring 2017 test administration. Previously administered operational test items served as linking items between the Spring 2017 and Spring 2018 administrations, allowing the Spring 2018 assessments to be placed on the Wisconsin Forward Exam scales 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 2017–18 testing cycle. The structure of this 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. (p. 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
- 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 contribute to 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 were the basis of the test blueprints and item specifications for their respective content areas. Wisconsin educators were involved in reviewing the items in all content areas to ensure the appropriateness of the test to the standards. The first item review for grades 3–8 in ELA and Mathematics, grades 4 and 8 in Science, and grades 4, 8, and 10 in Social Studies occurred in December 2015. 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 subsequent item reviews, supported by the item data acquired after the Spring 2016 and Spring 2017 test administrations, occurred in Summer 2016 and Summer 2017 and were conducted by DPI content experts. The purpose of these reviews was to refine the pool of items from which the subsequent operational test forms were selected.

Part 3: Test Form Development

Part 3 presents the Wisconsin Forward Exam design and discusses key development tasks related to creating the Spring 2018 Wisconsin Forward Exam forms. The Spring 2018 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 forms were supplemented through the use of *TerraNova* items (CTB/McGraw-Hill, 2009). Social Studies test forms consisted of Wisconsin-owned items supplemented by *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 2018 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), which is 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 2017–18 school year, the Wisconsin Forward Exam was administered to Wisconsin students for the third time.

Test administration was conducted during a seven-week window: March 19–May 4, 2018. 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) items and multi-select (MS) items; autoscoring of technology-enhanced (TE) items, short-answer (SA) items, and evidence-based selected response (EBSR) 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 the handscoring of the TDA writing items is presented in detail.

Part 6: Calibration, Equating, and Deriving Scale Scores

The Spring 2018 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 2018 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 includes model-to-data fit and the standard error of measurement (SEM). The equating of Spring 2018 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 2018 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 the results of item analysis as well as the 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, accommodation use, 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, accommodation use, 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 a 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 2018 Wisconsin Forward Exam administration are presented in the body of the report. However, some issues 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 2018 Wisconsin Forward Exam administration cover 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 2018 administration.

This part of the Technical Report is particularly relevant to American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (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. (p. 85)

English Language Arts (ELA), Mathematics, and Science test items included in the Spring 2018 Wisconsin Forward Exam were selected from DRC's College- and Career-Ready (CCR) item bank. DRC's CCR item bank contains nationally field tested CCR items that support the next generation of standards and assessments. It is aligned to the College and Career Readiness standards in ELA and Mathematics grades 3–8. Science items are 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 process 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 DRC’s ELA, Mathematics, and Science items, appearing on the Wisconsin Forward Exam, 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.

Wisconsin-owned Social Studies items were developed by DRC content specialists. These items are aligned to Wisconsin's Model Academic Standards for Social Studies. Social Studies items underwent reviews by DRC content experts as well as DRC bias and sensitivity experts. All Social Studies items were also reviewed and approved by committees of Wisconsin educators.

Various item types were developed and included in the Wisconsin Forward Exam in order to best assess students’ understandings of the standards. Descriptions of each item type used in the Wisconsin Forward Exam are provided 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. (p. 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. (p. 64)

As stated earlier, the State of Wisconsin licensed ELA, Mathematics, and Science items from DRC’s CCR item bank for the Spring 2018 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*, 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 2017 Field Test

The test items typically begin their life cycle two years prior to their operational administration. New ELA, Mathematics, Science, and Social Studies passages and items were first reviewed and approved for placement on the Wisconsin Forward Exam 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 students during the actual administration. ELA item reviews were held August 22–25, 2016, Mathematics item reviews were held August 22–24, 2016, and the Social Studies item review was held August 22, 2016. The Science item review was conducted October 18–19, 2016. An example of the training PowerPoint presentation used at the reviews can be found in Appendix A of this report.

Table 2-7 shows the number of items taken to the item review by grade and content area. 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.

Items and passages that were approved by the Wisconsin educators were then included in the next field test administration in Spring 2017. The purpose of the Spring 2017 field test was to expand the pool of items eligible for inclusion in the subsequent operational test forms, such as the Spring 2018 Forward Exam.

2.3 Field Testing

Items approved for the field test administration during the Spring 2016 item review were field tested in Spring 2017 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 352 new items were field tested for ELA. A total of 184 items (24 for grade 3 and 32 per grade for grades 4–8) were field tested for Mathematics. A total of 179 items were field tested for Science, and a total of 104 items (32 to 40 per grade) were field tested for Social Studies in the Spring 2017 test administration.

2.4 Statistical Analysis of Spring 2017 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 analyses, 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).

The field test 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 $>5\%$
- 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 2017 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, who were broken out into content area and/or grade-level groups. The data review took place in Madison, Wisconsin, on August 15 and 16, 2017. 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 test 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 B. A summary of the data review results, including the number of items that were 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 Table 2-8. Items accepted for subsequent use in the Wisconsin Forward Exam were included in the pool of items for Spring 2018 operational test form selection.

2.6 Summary

In summary, the items included in the Spring 2018 Wisconsin Forward Exam were reviewed by DRC, DPI, and Wisconsin educators for issues regarding 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. Following Spring 2017 field testing, items flagged for accessibility, bias and sensitivity, and/or other content concerns were further reviewed by DRC and DPI to determine if these flagged items should be removed from the Wisconsin item pool prior to the form construction of the Wisconsin Forward Exam. In addition, item statistics from the Spring 2017 operational and field test administration were used to refine the item pool used in the selection of Spring 2018 Wisconsin Forward Exam forms.

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

DRC College- and Career-Ready Item Bank Development Activities
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 Item Type Descriptions for Items on the Wisconsin Forward Exam

Item Type	Name	Description
EBSR	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 a given content area. Note that there were no operational MS items on Spring 2018 Forward Exam. All MS items were field test items.
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.
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. A weight of 2 is applied to the item scores in computation of the student total test raw scores and scale scores. That is, the TDA prompts contribute up to 8 raw score points towards the student total test raw score. 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)	Depth of Knowledge	Total Points by Grade					
		3	4	5	6	7	8
Reading		22	24	24	24	24	24
Key Ideas and Details	grade 3: 1-3 grades 4-8: 2-3	6–12	6–12	6–12	6–12	6–12	6–12
Craft and Structure/Integration of Knowledge and Ideas	all grades: 2-3	4–10	4–10	4–10	4–10	4–10	4–10
Vocabulary Use - Includes Language Standards 4 and 5	grades 3-5: 1-3 grades 6-8: 2-3	4–6	4–6	4–6	4–6	4–6	4–6
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%
Writing/Language		24	24	24	24	24	24
Text Types and Purposes/ Text-Dependent Analysis	all grades: 2-3	10-14	10-14	10-14	10-14	10-14	10-14
Research	all grades: 2-3	6–8	6–8	6–8	6–8	6–8	6–8
Language Conventions	all grades: 1-3	6–8	6–8	6–8	6–8	6–8	6–8
Listening	all grades: 2-3	7	8	8	8	8	8
ELA Points Total		53	56	56	56	56	56

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

Reporting Category	Depth of Knowledge	Total Points by Grade					
		3	4	5	6	7	8
Operations and Algebraic Thinking	grade 3: 1-3 grades 4-5: 1-2	8–10	9–11	8–10			
Number and Operations in Base Ten	grades 3-5: 1-3	7–9	8–10	8–10			
Number and Operations–Fractions	grades 3-5: 1-3	7–9	9–11	8–10			
Measurement and Data	grades 3-5: 1-3	9–11	9–11	9–11			
Geometry	grades 3-4: 1-2 grades 5-8: 1-3	6–8	6–8	8–10	6–8	9–11	9–11
Ratios and Proportional Relationships	grades 6-7: 1-3				6–8	7–9	
The Number System	grades 6-7: 1-3 grade 8: 1-2				10–12	6–8	7–9
Expressions and Equations	grades 6,8: 1-3 grade 7: 1-2				10–12	9–11	9–11
Statistics and Probability	grade 6: 1-2 grades 7-8: 1-3				9–11	10–12	7–9
Functions	grade 8: 1-3						9–11
Mathematics Points Total		42	46	46	46	46	46

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

Reporting Category	Depth of Knowledge	Total Points by Grade	
		4	8
Science Connections & Nature of Science	grade 4: 2-3 grade 8: 1-3	7–10	6–9
Science Inquiry	grades 4, 8: 2-3	6–9	7–10
Physical Science	grades 4, 8: 1-3	5–7	5–7
Earth and Space Science	grades 4, 8: 1-3	5–7	5–7
Life and Environmental Science	grades 4-8: 1-3	5–7	5–7
Science Applications and Science in Social and Personal Perspectives	grade 4: 1-3 grade 8: 2-3	6–9	6–9
Science Total Points		40	40

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

Reporting Categories	Depth of Knowledge	Total Points by Grade		
		4	8	10
Geography: People, Places, and Environments	all grades: 1-3	7–11	8–12	9–11
History: Time, Continuity, and Change	all grades: 1-3	6–10	10–15	11–14
Political Science and Citizenship: Power, Authority, Governance, and Responsibility	grade 4: 2-3 grades 8, 10: 1-3	5–9	5–7	11–14
Economics: Production, Distribution, Exchange, and Consumption	all grades: 1-3	5–9	5–7	7–10
The Behavioral Sciences: Individuals, Institutions, and Cultures	all grades: 2-3	5–9	4–6	7–10
Social Studies Total Points		38	40	50

Table 2-7 Items Reviewed at Summer 2016 Item Review

Grade	Number of Items			
	English Language Arts	Mathematics	Science	Social Studies
3	88	37		
4	96	37	153	51
5	89	35		
6	92	37		
7	87	37		
8	85	37	140	49
10				68
TOTAL	537	220	293	168

Table 2-8 Items Reviewed at Summer 2017 Data Review

Content	Number of FT Items in Spring 2017	Flagged Items in Spring 2017 FT Examined at Data Review		Flagged Items in Spring 2017 FT Rejected at Data Review	
		Number of Items	% of Field Test	Number of Items	% of Field Test
English Language Arts	352	90	26%	63	18%
Mathematics	184	80	43%	40	22%
Social Studies	104	30	29%	12	12%
Science	179	78	44%	68	38%
TOTAL	819	278		183	

Part 3: Test Form Development

Part 3 of this report focuses on key development tasks related to creating the Spring 2018 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 information about the test design of the content areas assessed on the Spring 2018 Wisconsin Forward Exam assessments.

3.1.1 English Language Arts

Table 3-1 shows the ELA test design, including the number of passages, items, and points at each grade level that were used in the core and embedded field test positions. There was one common set of core operational 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 Table 2-2 of this report.

The ELA section of the 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 Mathematics test design, including the number of items and points at each grade level that were used in the core and embedded field test positions. There was one common set of core operational items in each of the four field test forms at each grade level.

The Mathematics section of the exam was divided into two testing sessions, with students able to take the sessions in either order. In grades 3-5, no calculator was allowed for any of the Mathematics items. In grades 6–8, no calculator was allowed for the first session, and the second session allowed students to use an embedded calculator. 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 shows the Science test design, including the number of items and points at each grade level that were used in the core and embedded field test positions. There was one common set of core operational items in each of the twelve field test forms at each grade level. The operational portion of the Science tests was re-used from the Spring 2017 test administration. Reporting for the operational test remained aligned to the WMASS standards in Spring 2018. The Science exam included two sessions which could be administered in either order. Recommended testing times for both sessions were included in the test design document as well as in the test administration manual.

3.1.4 Social Studies

Table 3-4 shows the Social Studies test design, including the number of items and points at each grade level that were used in the core and embedded field test positions. There was one common set of core operational items in each of the four field test forms at each grade level. The Social Studies exam included two test sessions which could be administered in either order. The Social Studies exam at grades 4, 8, and 10 included custom items developed specifically for the Wisconsin Forward Exam and *TerraNova* items (CTB/McGraw-Hill, 2009). Recommended testing times for both sessions were included in the test design document as well as 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 process 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). (p. 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. (p. 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. (p. 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, which included content area specialists, psychometricians, and scoring specialists working as a unit in collaboration with DPI content experts.

3.2.2 Item Selection

New operational test forms were developed for ELA, Mathematics, and Social Studies for the Spring 2018 test administration. The operational portions of the Science tests were reused from the Spring 2017 administration, with new field test items added to create the 2018 test forms. 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, which is 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. The following process was used for item selection:

- Using the pool of vendor-owned items for ELA, Mathematics, and Science, and Wisconsin-owned Social Studies 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 the item was 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 ≤ 0.20 or ≥ 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 2017 test forms were used as targets for selection of the Spring 2018 test forms. The item 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 2018 test forms were selected from the Spring 2016 and 2017 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 Spring 2016 and 2017 Wisconsin Forward Exam operational and field test item pool. The TDA items selected for the Spring 2018 test administration were 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 each grade and content area. In constructing the final 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.
- 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 2018 Wisconsin Forward Exam per grade and content area. The test maps include the session number, item sequence, item type, item usage, item maximum score, depth-of-knowledge level, 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 Item and Form Quality Reviews

In all phases of the item and form development process, content area test development specialists and 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 multitiered internal review of all passages and items prior to submission for review by DPI and then, with approval by DPI prior to submission, for review by Wisconsin educators 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 and test forms, 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 multitiered quality assurance process made certain that 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 the multiple electronic platforms on which students took the test to ensure a smooth testing experience.

3.3 DPI Approvals

DPI had the opportunity to review passages and items to be placed on the Spring 2018 Wisconsin Forward Exam during the following phases:

- prior to item content review in Spring 2016
- at item content review in Summer 2016
- during review of flagged field test data in Summer 2017
- during the Spring 2018 form 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 2018 Wisconsin Forward Exam balanced the content and psychometric requirements for the form development. The content of the Spring 2018 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 2017 forms. Overall, the process implemented in the Spring 2018 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	3	2	2	2
	Informational	2	2	2	2	2	2
	Listening	2	2	2	2	3	2
Number of Core (OP) Items	Item Type: SR/TE (1 pt)	27	28	32	23	28	30
	Item Type: SR/TE/EBSR (2 pts)	9	10	8	13	10	9
	Item Type: TDA (4 pts x 2)	1	1	1	1	1	1
	Total Core Items	37	39	41	37	39	40
Total Core Points		53	56	56	57	56	56
Embedded Field Test (FT)	Number of Forms	8	8	8	8	8	8
	Passages (Reading + Listening)	2	2	2	2	2	2
	FT Items per Form	8	8	8	8	8	8
	Total Items Field Tested	63	63	63	63	63	63
Total Items (Core + FT) per Form		45	47	49	45	47	48
Total Estimated Testing Time (minutes)		125	125	125	125	125	125

Note: TDA items are scored using a 1–4-point scoring rubric. A weight of 2 is applied to item scores in computation of the student total test raw scores and scale scores.

Table 3-2 Mathematics Test Design

Test Design		Grade					
		3	4	5	6	7	8
Number of Core (OP) Items	Item Type: MC/EBSR/SA (1 pt)	37	41	39	40	41	40
	Item Type: TE (1 pt)	5	5	7	6	5	6
	Total Core Items	42	46	46	46	46	46
Total Core Points		42	42	46	46	46	46
Embedded Field Test (FT)	Number of Forms	4	4	4	4	4	4
	FT Items per Form	8	8	8	8	8	8
	Total Items Field Tested	32	32	32	32	32	32
Total Items (Core + FT) per Form		50	50	54	54	54	54
Total Estimated Testing Time (minutes)		90	90	90	90	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	10	10
	Total Items Field Tested	120	120
Total Items (Core + FT) per Form		48	50
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 2018, Wisconsin administered assessments in ELA and Mathematics for grades 3–8. Science was administered in grades 4 and 8, and Social Studies was administered in grades 4, 8, and 10. The test administration window was March 19–May 4, 2018. 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 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 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 or supports. Accommodations and supports are practices and procedures that provide equitable access to grade-level content. They are intended to reduce or eliminate the effects of a student’s disability or level of language acquisition; they do not reduce learning expectations. DPI guidance makes it clear that the accommodations or supports provided to a student must be consistent with 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. (p. 115)

An overview of the types of accommodations and supports, 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 and supports were to be available for use by each student within the student learning profile in DRC's eDIRECT system. All student accommodations and supports 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 students' needs. Those supports 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 accessibility features of the assessment are either provided as digitally-delivered components of the test administration system (embedded) or separate from it (non-embedded).

Embedded Universal Tools (“Online”)

- Calculators
- Click to Enlarge
- Cross-off Tools
- Flag/Mark for Review
- Help/What's This?
- Highlighter
- Go to Question
- Keyboard Navigation
- Line Guide

- Magnifier Tool (Zoom)
- Measuring Tools
- Pause (Breaks)
- Review Page
- Sticky Notes (Digital Notepad)
- Test Directions
- Tool Tips

Non-Embedded Universal Tools (“Standard”)

- Scratch Paper

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 student’s classroom instruction. They are either provided as part of the online test administration system or separate from it (i.e., embedded or non-embedded). All embedded and non-embedded designated supports must be entered into eDIRECT prior to test administration. Embedded and non-embedded supports will appear on student test tickets.

Embedded Designated Supports (“Online”)

- Color Choices (CC)
- Contrasting Color (CTC)
- Reverse Contrast (RC)
- Masking (MSK)
- Text-to-Speech (TTS)
- Translation (Stacked)

Non-Embedded Designated Supports (“Standard”)

- Word-to-Word Bilingual Dictionary
- Color Overlay
- Magnification
- Noise Buffers
- Read Aloud
- Scribe
- Separate Setting
- Small Group Translation (new in 2018)
- Translation
- Translator/Interpreter (new in 2018)
- Amplification Device (new in 2018)

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. They are available for students for whom there is a documented need in the 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 online administration system or separate from it (i.e., embedded or non-embedded). All embedded and non-embedded accommodations must be entered into eDIRECT prior to test administration. Embedded and non-embedded accommodations will appear on student test tickets.

Embedded Accommodations (“Online”)

- Video Sign Language (VSL)
- Closed Captioning (C CAP)
- Text-to-Speech (Reading Passages) (TTS PSGS)

Non-Embedded Accommodations (“Standard”)

- Abacus
- Alternate Response Options
- Braille (Unified English Braille or English Braille American Edition) (BRL)
- Calculator
- Listening Scripts (LS)
- Multiplication Table
- Print on Demand (POD)
- Read Aloud (Reading Passages)

4.1.4 Translation

For the Spring 2018 Wisconsin Forward Exam administration, the State of Wisconsin used an embedded stacked Spanish translation for Mathematics, Science, and Social Studies items. For ELA assessments, only the test directions are available in stacked translation. The stacked Spanish translation is a designated support for students who are native Spanish speakers and are limited English proficient, to demonstrate their knowledge on the Wisconsin Forward Exam. In addition to the embedded stacked translation, bilingual word lists and translation of the test directions are allowed designated supports

DPI recognizes that approximately five 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. DPI recommended that educators consult the list of allowable accommodations and supports (referenced above) to create the most appropriate testing situation for their students.

4.1.5 Additional Accessibility Resources

Additional accessibility resources and guidance 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 2018 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. (p. 116)

Standard 6.7 Test users have the responsibility of protecting the security of test materials at all times. (p. 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 (i.e., Braille and Print-on-Demand), so the assessment exposure is limited to those educators who require 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 software quality assurance 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. Therefore, it is recommended that test tickets be printed as close to the date of testing as possible, and sites are instructed to keep test tickets and rosters in a secure location until the session is scheduled to begin. Test tickets are distributed just prior to students logging in 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 2018 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 assessments, 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. (p. 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. (p. 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. (p. 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. (p. 115)

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

Standard 6.4 The testing environment should furnish reasonable comfort with minimal distractions to avoid construct-irrelevant variance. (p. 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 (TAs)/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
- Checklists for Responsible Parties (DACs, SACs, TAs)

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

- 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
- 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, schools and districts were encouraged to provide students with time to complete both a tutorial video series and an online tools training. Sample test items were also provided for each grade and content area.

Student and Administrator Tutorial Videos

Student and administrator tutorial videos were available for students and test administrators to become familiar with the online testing environment. 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 (OTT) was provided for students to allow them a hands-on opportunity to practice the types of items and tools available in the online testing system. The OTTs were 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. The 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 OTTs were not scored and were not intended for content practice.

Item Samplers

Item samplers were developed for use by both educators and students to gain familiarity with the types of items and their functionality. The format appears as a ‘guided practice test’ in the online, PDF and braille versions of the tests.

Accommodation versions of the item samplers, reflecting the Forward Exam were produced including TTS, stacked Spanish translation (in Mathematics, Science and Social Studies), VSL with CC, and HVA for listening passages. All tools and supports available in the test engine were applied to this student online experience.

Access to the item samplers was granted through the OTT menu page. A user name and password was displayed on the login screen. The “click to enlarge” item displayed the answer key and scoring guide for each item online. In addition, a paper answer key and scoring guide were provided as a document for posting.

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 sign-off 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
Braille [BRL]	3	0.00	2	0.00
Bilingual Dictionary			230	0.36
Magnification	98	0.16	101	0.16
Noise Buffers	693	1.10	684	1.08
Read Aloud	1386	2.19	1502	2.37
Scribe	603	0.95	559	0.88
Separate Setting	7704	12.19	7719	12.19
Alternate Response Options	10	0.02	9	0.01
Read Aloud (Reading Passages)	5	0.01		
Color Choices [CC]	69	0.11	69	0.11
Contrasting Color [CTC]	47	0.07	47	0.07
Reverse Contrast [RC]	25	0.04	24	0.04
Masking [MSK]	796	1.26	794	1.25
Text-to-Speech [TTS]	11057	17.50	11563	18.26
Spanish Translation [ST]	613	0.97	907	1.43
Video Sign Language [VSL (ASL)]	12	0.02	12	0.02
Color Overlay	46	0.07	45	0.07
Amplification Device	38	0.06	38	0.06
Small Group Translation	118	0.19	149	0.24
Translator/Interpreter	33	0.05	57	0.09
Closed Captioning [C CAP] ELA	45	0.07		
Listening Scripts [LS] ELA	4	0.01		
Text-to-Speech for Reading Passages [TTS (PSGS)] ELA	5	0.01		
Abacus Math			53	0.08
Non-embedded Calculator Math			182	0.29
Multiplication Table Math			814	1.29

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
Braille [BRL]	5	0.01	5	0.01	5	0.01	5	0.01
Bilingual Dictionary			254	0.39	255	0.40	255	0.40
Magnification	121	0.19	122	0.19	118	0.18	119	0.18
Noise Buffers	811	1.26	815	1.26	784	1.22	783	1.21
Read Aloud	1213	1.88	1327	2.06	1269	1.97	1259	1.95
Scribe	636	0.99	576	0.89	578	0.90	579	0.90
Separate Setting	8296	12.89	8345	12.95	8199	12.72	8191	12.71
Alternate Response Options	10	0.02	10	0.02	10	0.02	10	0.02
Read Aloud (Reading Passages)	10	0.02						
Color Choices [CC]	124	0.19	124	0.19	121	0.19	120	0.19
Contrasting Color [CTC]	103	0.16	104	0.16	103	0.16	104	0.16
Reverse Contrast [RC]	70	0.11	70	0.11	69	0.11	69	0.11
Masking [MSK]	841	1.31	854	1.32	835	1.30	834	1.29
Text-to-Speech [TTS]	11242	17.47	11660	18.09	11464	17.79	11471	17.80
Spanish Translation [ST]	687	1.07	943	1.46	922	1.43	885	1.37
Video Sign Language [VSL (ASL)]	20	0.03	18	0.03	18	0.03	18	0.03
Color Overlay	77	0.12	78	0.12	78	0.12	78	0.12
Amplification Device	41	0.06	40	0.06	40	0.06	40	0.06
Small Group Translation	115	0.18	145	0.22	138	0.21	135	0.21
Translator/Interpreter	52	0.08	73	0.11	65	0.10	63	0.10
Closed Captioning [C CAP] ELA	48	0.07						
Listening Scripts [LS] ELA	5	0.01						
Abacus Math			29	0.04				
Non-embedded Calculator Math			257	0.40				
Multiplication Table Math			2119	3.29				

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
Braille [BRL]	5	0.01	5	0.01
Print on Demand [POD]	2	0.00	2	0.00
Bilingual Dictionary			277	0.43
Magnification	91	0.14	92	0.14
Noise Buffers	749	1.15	748	1.15
Read Aloud	1255	1.93	1355	2.08
Scribe	629	0.97	585	0.90
Separate Setting	8387	12.92	8435	12.97
Alternate Response Options	11	0.02	10	0.02
Read Aloud (Reading Passages)	3	0.00		
Color Choices [CC]	102	0.16	102	0.16
Contrasting Color [CTC]	77	0.12	77	0.12
Reverse Contrast [RC]	43	0.07	43	0.07
Masking [MSK]	732	1.13	730	1.12
Text-to-Speech [TTS]	10493	16.17	10846	16.68
Spanish Translation [ST]	488	0.75	692	1.06
Video Sign Language [VSL (ASL)]	24	0.04	24	0.04
Color Overlay	37	0.06	37	0.06
Amplification Device	34	0.05	35	0.05
Small Group Translation	86	0.13	104	0.16
Translator/Interpreter	33	0.05	49	0.08
Closed Captioning [C CAP] ELA	68	0.10		
Listening Scripts [LS] ELA	8	0.01		
Text-to-Speech for Reading Passages [TTS (PSGS)] ELA	6	0.01		
Abacus Math			28	0.04
Non-embedded Calculator Math			294	0.45
Multiplication Table Math			2561	3.94

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
Braille [BRL]	4	0.01	5	0.01
Print on Demand [POD]	1	0.00	1	0.00
Bilingual Dictionary			202	0.32
Magnification	87	0.14	86	0.14
Noise Buffers	609	0.96	607	0.95
Read Aloud	827	1.30	920	1.44
Scribe	388	0.61	371	0.58
Separate Setting	6956	10.94	6976	10.96
Alternate Response Options	7	0.01	8	0.01
Read Aloud (Reading Passages)	5	0.01		
Color Choices [CC]	126	0.20	126	0.20
Contrasting Color [CTC]	87	0.14	87	0.14
Reverse Contrast [RC]	41	0.06	41	0.06
Masking [MSK]	594	0.93	587	0.92
Text-to-Speech [TTS]	8765	13.78	9060	14.23
Spanish Translation [ST]	254	0.40	326	0.51
Video Sign Language [VSL (ASL)]	18	0.03	17	0.03
Color Overlay	36	0.06	37	0.06
Amplification Device	23	0.04	24	0.04
Small Group Translation	90	0.14	100	0.16
Translator/Interpreter	20	0.03	37	0.06
Closed Captioning [C CAP] ELA	68	0.11		
Text-to-Speech for Reading Passages [TTS (PSGS)] ELA	5	0.01		
Abacus Math			15	0.02
Non-embedded Calculator Math			477	0.75
Multiplication Table Math			2620	4.12

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
Braille [BRL]	4	0.01	4	0.01
Print on Demand [POD]	3	0.00	3	0.00
Bilingual Dictionary			265	0.42
Magnification	72	0.11	73	0.12
Noise Buffers	479	0.76	472	0.75
Read Aloud	719	1.14	791	1.25
Scribe	277	0.44	259	0.41
Separate Setting	6910	10.94	6943	10.98
Alternate Response Options	9	0.01	11	0.02
Read Aloud (Reading Passages)	4	0.01		
Color Choices [CC]	123	0.19	123	0.19
Contrasting Color [CTC]	98	0.16	98	0.16
Reverse Contrast [RC]	59	0.09	59	0.09
Masking [MSK]	763	1.21	758	1.20
Text-to-Speech [TTS]	7889	12.49	8079	12.78
Spanish Translation [ST]	296	0.47	381	0.60
Video Sign Language [VSL (ASL)]	22	0.03	23	0.04
Color Overlay	31	0.05	31	0.05
Amplification Device	28	0.04	28	0.04
Small Group Translation	57	0.09	65	0.10
Translator/Interpreter	19	0.03	25	0.04
Closed Captioning [C CAP] ELA	56	0.09		
Listening Scripts [LS] ELA	4	0.01		
Text-to-Speech for Reading Passages [TTS (PSGS)] ELA	3	0.00		
Abacus Math			9	0.01
Non-embedded Calculator Math			583	0.92
Multiplication Table Math			2558	4.05

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
Braille [BRL]	5	0.01	5	0.01	5	0.01	5	0.01
Print on Demand [POD]	3	0.00	3	0.00	3	0.00	3	0.00
Bilingual Dictionary			241	0.38	236	0.37	235	0.37
Magnification	80	0.13	81	0.13	79	0.12	79	0.12
Noise Buffers	436	0.69	434	0.69	416	0.66	413	0.65
Read Aloud	563	0.89	596	0.94	594	0.94	579	0.92
Scribe	242	0.38	204	0.32	222	0.35	223	0.35
Separate Setting	6741	10.66	6800	10.74	6720	10.62	6720	10.63
Alternate Response Options	10	0.02	10	0.02	8	0.01	8	0.01
Read Aloud (Reading Passages)	4	0.01						
Color Choices [CC]	209	0.33	211	0.33	210	0.33	211	0.33
Contrasting Color [CTC]	86	0.14	86	0.14	81	0.13	81	0.13
Reverse Contrast [RC]	56	0.09	56	0.09	56	0.09	56	0.09
Masking [MSK]	648	1.02	643	1.02	644	1.02	643	1.02
Text-to-Speech [TTS]	7615	12.04	7845	12.39	7988	12.62	7730	12.23
Spanish Translation [ST]	255	0.40	350	0.55	348	0.55	347	0.55
Video Sign Language [VSL (ASL)]	18	0.03	18	0.03	18	0.03	18	0.03
Color Overlay	24	0.04	24	0.04	24	0.04	24	0.04
Amplification Device	26	0.04	26	0.04	26	0.04	26	0.04
Small Group Translation	56	0.09	63	0.10	62	0.10	62	0.10
Translator/Interpreter	15	0.02	23	0.04	23	0.04	23	0.04
Closed Captioning [C CAP] ELA	58	0.09						
Listening Scripts [LS] ELA	4	0.01						
Text-to-Speech for Reading Passages [TTS (PSGS)] ELA	4	0.01						
Abacus Math			3	0.00				
Non-embedded Calculator Math			607	0.96				
Multiplication Table Math			2290	3.62				

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

Grade 10	Social Studies	
	N Count	Percent
Braille [BRL]	5	0.01
Print on Demand [POD]	4	0.01
Bilingual Dictionary	208	0.33
Magnification	46	0.07
Noise Buffers	81	0.13
Read Aloud	382	0.61
Scribe	68	0.11
Separate Setting	4153	6.63
Alternate Response Options	2	0.00
Color Choices [CC]	24	0.04
Contrasting Color [CTC]	16	0.03
Reverse Contrast [RC]	14	0.02
Masking [MSK]	235	0.38
Text-to-Speech [TTS]	3119	4.98
Spanish Translation [ST]	170	0.27
Video Sign Language [VSL (ASL)]	9	0.01
Color Overlay	7	0.01
Amplification Device	18	0.03
Small Group Translation	35	0.06
Translator/Interpreter	34	0.05

Table 4-8 Summary Table of Manual Materials

Material	Configuration
<p>DAC/SAC Guide (District Assessment Coordinator/School Assessment Coordinator Guide)</p>	<p>The DAC/SAC Guide is a 39-page handbook that includes the following information:</p> <ul style="list-style-type: none"> • Key dates • Roles and responsibilities • Test security • Accessibility information • Procedures before testing begins • Technology resources • Testing times and schedules • Braille ordering • Overview of testing and test management software • Procedures for once testing is finished • Transferring students • Coordinator checklists • Guidelines and procedures for documenting a test security incident • Multiplication chart (for use with some tests) • Sample test schedules
<p>eDIRECT User Guide: User Management</p>	<p>The Manage Users Guide is a 29-page guide that includes the following information:</p> <ul style="list-style-type: none"> • Managing user’s own eDIRECT account • Adding and editing other eDIRECT users • Adding and removing eDIRECT user permissions
<p>eDIRECT User Guide: Students and Testing</p>	<p>The Students and Testing Guide is a 72-page guide that includes the following information:</p> <ul style="list-style-type: none"> • Adding and editing students and student demographics, accommodations, and testing codes • Viewing, adding, and editing student test session information • Printing and managing student test tickets • Transferring students between schools and districts
<p>Accessibility Guide</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 are also included.</p>
<p>Student/Administrator Tutorials</p>	<p>The Student Tutorial includes 12 videos intended for students in grades 4-10 and 7 videos for students in grade 3. 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 2018 tutorial also includes ten videos for test administrators to familiarize them with the administrative features and functionality of eDIRECT as well as the accessibility features of the Wisconsin Forward Exam.</p>
<p>Item Samplers</p>	<p>The item samplers can be used by both educators and students to gain familiarity with the types of items and functionality. The format appears as a ‘guided practice test’ in the online, PDF and braille versions. Accommodations, universal tools and supports are available in the test engine for the item samplers.</p> <p>Item samplers are accessible through the OTT menu page. They include the answer key and scoring guide for each item.</p>

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

Material	Configuration
<p>TAM (Test Administration Manual) and Test Directions</p>	<p>The TAM is a 54-page document intended for test proctors. It includes the following information:</p> <ul style="list-style-type: none"> • Key dates • Test times and schedules • Test security • Accessibility information • Procedures for before testing • Test ticket management • Test material management • Setting up the testing environment • Procedures for during testing • Procedures for after testing • Proctor checklist and guidelines • Read-aloud protocol • Scribe guidelines <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>Technology User Guide (TUG)</p>	<p>The TUG is an approximately 280-page document, split into 5 volumes, intended for Technology Coordinators. It includes detailed instructions on the installation and configuration of INSIGHT and the Testing Site Manager for all supported platforms. DRC also produces a 336-page version of the TUG for those districts previewing Central Office Services (COS) in 2018.</p>
<p>Interpretive Guide</p>	<p>The Interpretive Guide is a 27-page document that includes the following information:</p> <ul style="list-style-type: none"> • Interpreting Wisconsin Forward Exam scores • Accessing Individual Student Reports (ISRs) and summary reports via the eDIRECT Portal
<p>Technology Readiness Package</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"> • Capacity Estimator • System requirements • Technology overview presentation • Technology Coordinator Checklist • Tech FAQ
<p>Online Tools Training (OTT)</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>Technical Report</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"> • Description of the item pool used in the Wisconsin form-development process • Description of the test administration process and test security • Scoring of various types of items • Summary information of student performance (including means and standard deviations of scale scores, percentage of examinees within each performance level for each content area and grade level, and scale score distribution tables) • Item- and test-level analysis information for each content area and grade level, test scaling procedure, and student scoring process • Measures of scoring reliability for text-dependent analysis items • Evidence of test validity
<p>Data Forensic Report</p>	<p>A separate Data Forensic Report will include analyses of the following:</p> <ul style="list-style-type: none"> • Evaluation of response changes • Evaluation of student response time to items

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. (p. 91)

Part 5 describes

- the scoring process of multiple-choice (MC) and multi-select (MS) items;
- the autoscoring process of technology-enhanced (TE), short-answer (SA), and evidence-based selected response (EBSR) 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 MS 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 MS 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 EBSR 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, and EBSR item scoring included the following:

- A scoring rubric was created for each TE, SA, and EBSR item. It was similar to describing the one-and-only-one correct answer for dichotomously scored items (scored as either right or wrong). For ELA EBSR 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 EBSR 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 2018 administration, the ELA forms in grades 3–8 contained one TDA item at each grade level. As stated in Part 2, Table 2-2, of this report, the TDA prompts are scored using a holistic scoring guideline on a 1–4-point scale. A weight of 2 is later applied to the item scores in computation of the student total test raw scores and scale scores. That is, the TDA prompts will contribute up to 8 raw score points towards the student total test raw score.

The TDA 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 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, and resolving any scores that did not agree. While the engine only requires one score per response to build a model, the second score provides necessary information about how well two humans are able to agree on a score, which is then used as a benchmark for how well the engine’s predictions should agree with the human scores. Once the sample was scored, responses and corresponding scores were delivered to the AI team at MI for model development. MI’s linguistics experts, software developers, psychometricians, and human-computer interaction specialists created task-specific algorithms that were then used to predict how humans would score these responses.

To build a scoring model, the engine analyzes the training set and calculates features that pertain to the content in question. The engine then sends the features to dozens of different algorithms that compete to see which ones can best associate the features with the human-assigned scores. These algorithms draw on many of the latest advances in the field of machine learning to generate both linear and nonlinear models. The strongest models are then automatically blended together to create a final model that retains the best elements from the various algorithms.

When the engine builds a model, it selects the model elements that maximize scoring accuracy for the data in question. Therefore, it is important to choose an agreement statistic on which the engine can optimize its models in such a way that the final model will exhibit reliable, accurate scoring. The inter-rater reliability of two human scorers is often measured via exact and adjacent agreement or the Pearson product-moment correlation coefficient (Pearson’s r). It has also been found that using quadratic weighted kappa, which has become the industry standard for AI scoring as the optimization and evaluation metric, leads to the most reliable and accurate scoring. Quadratic weighted kappa as a metric can detect changes in mean difference and variance between scorers and is, therefore, well suited for comparing the accuracy of AI scoring with that of human scoring as well as measuring the agreement of two independent human scorers.

MI’s AI scoring software flagged 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, contain inappropriate language, or represent a bad-faith effort to complete the test (e.g., repeated text, off-topic text). These 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-scorable 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 who were trained and qualified to score these items.

5.3.5 Anchor Papers and Training Papers

DRC’s project managers and scoring directors began preparations for rangefinding by using the scoring guidelines (rubric) to select a representative sampling of student responses for each score point. The sample reflects the various, common response types produced for the specific item. The responses were then assembled into sample sets and duplicated for all rangefinding participants (project managers and scoring directors). This rangefinding committee read the passage(s) for the first grade/item, read and analyzed the writing prompt, and discussed the holistic scoring guideline. When an understanding of the scoring guideline had been established, participants read, scored, and discussed each response until consensus was reached. The scoring director for the specific grade took detailed notes, capturing scores and specific rationales for each score. Each grade and TDA item progressed in the same manner, using the same process. Once all sets were reviewed and scored, each grade-level scoring director selected responses to create a set of anchor papers, training papers, and qualifying papers. These anchor, training, and qualifying 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. The 2,000 responses were then delivered to the AI vendor to build the AI engine model. Once the model was built, 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. (p. 92)

DRC recruited, trained, and managed personnel to complete all 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, 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. (p. 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. (p. 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. For aberrant or inconsistent scoring, DRC has the capability to wholesale drop scores and have them rescored if deemed necessary.

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 based on the census data. The presented scores, on a 1-4 scale, are from the AI engine supplemented by non-scorable responses resolved by human readers.

Table 5-4 shows the score and non-scorable code distributions for TDA items for responses selected for the second read (handscoring). Table 5-5 shows the associated percentages of scores and non-scorable codes 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 numbers and percentages of the non-scorable responses from the 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 acceptable degree of agreement between the AI engine and the human scorers, with the differences being approximately 2% or less for most score points. Greater differences between the AI engine and the human scorers were found in grade 3 at score points 1 and 2, in grade 6 at score point 3, in grade 7 at score points 1 and 3, and in grade 8 at score points 1, 2, and 3. These discrepancies ranged from approximately 3% to 6%. It was observed that the human scorers had higher percentages of score 1 compared to the AI engine, while the AI engine tended to have higher percentages of scores 3 and 4.

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

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

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"> • are completely erased (so that words are unreadable). • are completely crossed out (so that words are unreadable). • are online and consist solely of “white space” (e.g., spaces, tabs, returns).
R – Refusal	<p>A response indicates a refusal to attempt the task. This includes the following examples:</p> <ul style="list-style-type: none"> • <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> • <i>“I don’t know”; “IDK”; “we never learned this”; “X”; “NA”</i> • <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> • <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> <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"> • responses written entirely in a language other than English. • responses that are completely illegible due to poor handwriting.* • 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>) • responses too insufficient to be assessed by the criteria on the rubric. • (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). • (for TDAs only) responses that consist solely, or almost solely, of text copied directly from the passage(s). <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	63194	32483	19954	3818	247	213	91	5938	351	99
4	6	64354	34677	19104	4111	943	154	54	4978	244	89
5	4	64903	30977	18568	4293	656	105	12	10097	176	19
6	4	63600	27239	22346	7697	504	134	55	5407	204	14
7	4	63140	22886	23837	9018	1786	119	4	5278	194	18
8	5	63248	15384	22925	14916	3263	174	6	6266	293	21

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)	11600	5206	2662	480	22			3230		
	Scorer 2 (Human)	11600	5910	2145	299	16		10	3122	59	39
4	Scorer 1 (AI Engine)	10647	4970	2304	449	100			2824		
	Scorer 2 (Human)	10647	5224	2260	295	44		5	2764	19	36
5	Scorer 1 (AI Engine)	14524	4549	2260	501	77			7137		
	Scorer 2 (Human)	14524	4856	2211	296	24		5	7097	23	12
6	Scorer 1 (AI Engine)	10113	3518	2496	859	63			3177		
	Scorer 2 (Human)	10113	3764	2533	601	38		4	3141	27	5
7	Scorer 1 (AI Engine)	8938	2451	2635	953	165			2734		
	Scorer 2 (Human)	8938	2689	2840	612	63			2705	19	10
8	Scorer 1 (AI Engine)	11746	1806	2636	1697	361	1		5245		
	Scorer 2 (Human)	11746	2314	2987	1051	148	1	2	5100	132	11

Note: TDA items are weighted x 2 in computation of student scores.

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)	11600	44.88	22.95	4.14	0.19			27.84		
	Scorer 2 (Human)	11600	50.95	18.49	2.58	0.14		0.09	26.91	0.51	0.34
4	Scorer 1 (AI Engine)	10647	46.68	21.64	4.22	0.94			26.52		
	Scorer 2 (Human)	10647	49.07	21.23	2.77	0.41		0.05	25.96	0.18	0.34
5	Scorer 1 (AI Engine)	14524	31.32	15.56	3.45	0.53			49.14		
	Scorer 2 (Human)	14524	33.43	15.22	2.04	0.17		0.03	48.86	0.16	0.08
6	Scorer 1 (AI Engine)	10113	34.79	24.68	8.49	0.62			31.42		
	Scorer 2 (Human)	10113	37.22	25.05	5.94	0.38		0.04	31.06	0.27	0.05
7	Scorer 1 (AI Engine)	8938	27.42	29.48	10.66	1.85			30.59		
	Scorer 2 (Human)	8938	30.09	31.77	6.85	0.70			30.26	0.21	0.11
8	Scorer 1 (AI Engine)	11746	15.38	22.44	14.45	3.07	0.01		44.65		
	Scorer 2 (Human)	11746	19.70	25.43	8.95	1.26	0.01	0.02	43.42	1.12	0.09

Note: TDA items are weighted x 2 in computation of student scores.

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 2018 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. (p. 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 2018 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 deriving 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 2018 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

multiple-choice (MC) items and constructed-response (CR) items, respectively. All non-MC items, including technology-enhanced (TE) items, evidence-based selected response (EBSR) items, short-answer (SA) items, and text-dependent analysis (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 θ 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 θ will respond in category k to partial-credit item j is

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

where $z_{jk} = (k - 1)f_j - \sum_{i=0}^{k-1} g_{ji}$ and $g_{j0} = \mathbf{0}$ 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 the 2014–15 administration for ELA and Mathematics) implemented for these content areas in the Wisconsin testing program. Item parameters estimated after the 2017–18 test administration were used to score the responses of Wisconsin students who took these tests.

6.1.2 Calibration Sample

The calibration of the Wisconsin Forward Exam occurred after the Spring 2018 test administration and was based on the student data acquired during the entire testing window. 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. (p. 25)

The calibration sample consisted of the student data acquired during the entire testing window 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 2018 calibration samples consisted of over 99% of the student population and, as such, were 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 2017 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, which is 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

Calibration 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

During calibration, items may not converge, meaning the characteristics of the item will not be able to be determined. When this occurs, items may be suppressed from student scoring and future assessments. In Spring 2018, 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, low-ability students are expected to be 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. Eighteen items were flagged for poor fit for ELA, three items were flagged for Mathematics, and two items were flagged for Social Studies. Most of the flagged items were CR items (TE and EBSR). For example, ELA grade 3 item #24 in calibration was flagged because the observed Z of 195.34 is larger than the critical Z value of 167.74 based on a sample size of 62,902. For many of the flagged items, the observed Z and the critical Z are not very far apart, indicating small misfit; however, it was observed that for some items the misfit was moderate (e.g., item #7 for ELA grade 5 or item #12 for ELA grade 6). 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 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 end of the distribution. Similarly, for CR items, there were, in general, fewer students at the lower score levels, which provides less information at the tail ends of the student distribution. Items that only show misfit at the tail ends 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 assessments from 2018 to the established Wisconsin Forward Exam ELA, Mathematics, Science, and Social Studies scales. Sets of items that were administered to Wisconsin students in previous operational test administrations and that were included in the Spring 2018 assessments served as the anchor sets in each ELA, Mathematics, and Social Studies grade. ELA anchor items were selected from the Spring 2016 and 2017 operational assessments. Mathematics and Social Studies anchor sets were selected from the Spring 2017 operational assessments. The anchor sets constituted at least 25 percent of the Spring 2018 assessments and were representative of the Spring 2018 test content. Since the Spring 2018 Science operational test forms were reused from the Spring 2017 test administration, all operational Science items were used as anchors in Spring 2018. 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 characteristic 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. (p. 105)

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

Test Characteristic Curve Method

The Stocking & Lord (1983) procedure, also called the TCC 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 2018 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 +.10 are considered large and differences between +.07 and +.10 are considered moderate.

Additionally, the maximum absolute difference (Max Abs) was identified. For Max Abs, large differences are those greater than +.15 and moderate differences are all differences between +.125 and +.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 a flagged anchor item 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. Therefore, 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 if 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 percent 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 percent 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 by the dashed red line and the new anchor estimate TCC (after equating) is indicated by 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 3 through 5, 7, and 8; Mathematics grades 5 through 7; and Social Studies grades 4 and 10. Two anchor items were flagged as a -parameter outliers in Mathematics grades 3 and 4 and Science grades 4 and 8. One anchor item was flagged as a b -parameter outlier in each of the following: ELA grades 3, 4, and 8; Mathematics grades 5 through 8; Science grade 4; and Social Studies grades 4 and 10. Two anchor items were flagged as b -parameter outliers in Mathematics grade 4, Science grade 8, and Social Studies grade 8. Overall the number of anchor items flagged using the TCC method was small. No anchor items were flagged using the item-ability regression statistics. No anchor items were excluded from test equating.

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. (p. 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 the scaling constants used in the Spring 2016 scale development. They are presented in Table 6-7. 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 the scale score metric for MC items,
- B_{ss} is a difficulty parameter in the scale score metric for MC items,
- F_{ss} is a discrimination parameter in the scale score metric for CR items,
- G_{ss} is a difficulty level (gamma) for category m_j in the scale score metric for CR items,
- a_{θ} is a discrimination parameter in the original theta metric for MC items,
- b_{θ} is a difficulty parameter in the original theta metric for MC items,
- f_{θ} is a discrimination parameter in the original theta metric for CR items,
- g_{θ} is a difficulty level (gamma) for category m_j in the original theta metric for CR items, and
- C_{ss} and c_{θ} is a guessing parameter in the original theta metric.

ELA Scale

Test Characteristic Curves—Figure 6-18 shows the TCCs for ELA tests. As shown in Figure 6-18, 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 the lower-to-middle part of the ability scale, indicating comparable difficulty of ELA grade 5

and grade 6 assessments for students of low-to-medium ability. The grade 6 and 7 TCCs are ordinal, indicating that the grade 7 ELA assessment is more difficult than the grade 6 ELA assessment. The grade 7 and grade 8 TCCs overlap at all ability levels, indicating that the grade 7 and 8 ELA assessments are of comparable difficulty for students at all ability levels.

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 pattern of scale scores at quartiles (see Growth at Quartiles paragraph below) for grades 3–8, 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 6-19 are U-shaped, indicating smaller errors around ability estimates that are 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 6-20. 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 6-21 shows the TCCs for Mathematics assessments, which are on a vertical scale. As observed in Figure 6-21, the TCCs for Mathematics are ordinal, indicating increasing difficulty of the assessment as the grade level increases.

Standard Error Curves—The SE curves for Mathematics presented in Figure 6-22 are U-shaped (as expected), indicating smaller errors around ability estimates that are 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 6-23. 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 6-24 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 6-25 shows the SE curves for Science grades 4 and 8. The SE curves are U-shaped, indicating smaller errors around ability estimates that are 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 6-26. 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—Although the Social Studies assessments are not vertically scaled, the TCCs for grades 4, 8, and 10 are presented together in Figure 6-27 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 6-28 shows Social Studies SE curves for grades 4, 8, and 10. The SE curves are U-shaped, indicating smaller errors around ability estimates that are 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 6-29. 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 on 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. It is expected 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 they follow a 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, θ 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 θ 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 ($u_1 u_2 \cdots u_n$) on a test with some number of items (n items), the “likelihood function,” or the probability that a student with a given ability value (θ) would produce this particular response pattern, is given by

$$P(u_1 u_2 \cdots u_n | \theta) = \prod_{i=1}^n P(u_i | \theta). \quad (2)$$

The formula indicates that the “estimated maximum likelihood” IRT item-pattern scoring method searches for the ability estimate (θ_0) that maximizes the probability function in (2) and assigns an ability estimate (θ_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 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 2 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

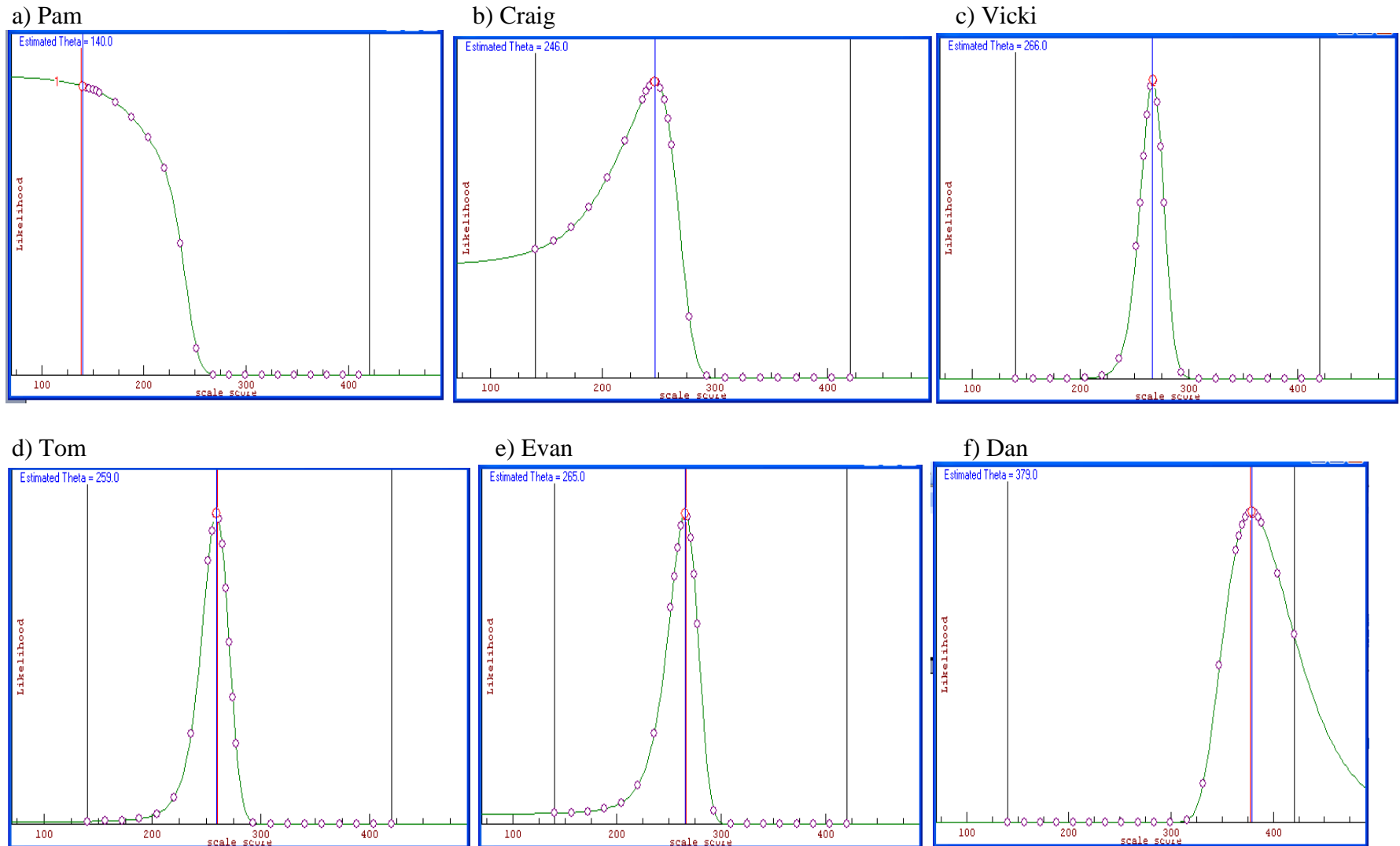
Student	Response Pattern ($u_1u_2 \cdots u_n$)	Raw Score	Item-Pattern Score
Pam	1000011001010000000000000101	7	140
Craig	101010101010101010101010101010	15	246
Vicki	010101010101010101010101010101	15	266
Tom	001100110011001100110011001101	15	259
Evan	110011001100110011001100110010	15	265
Dan	1111111111111111111111111011111	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 in conjunction 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 accurately 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 the students’ abilities. 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 the present purposes.

First, item-pattern scoring generally produces more accurate scores for individual students. Specifically, it produces a smaller 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 it would 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 & 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 a student's score obtained from a single testing will fall within one CSEM of that student's true score 68% of the time and that the obtained score will fall within two CSEMs of the true score 95% of the time.

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. (p. 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 6-19 (for ELA), 6-22 (for Mathematics), 6-25 (for Science), and 6-28 (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 higher 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 score LOSS and HOSS, are the first 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 2018 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 6-8 through 6-24. These tables are provided to illustrate the approximate raw score-to-scale score relationship for each unique raw score and do not include all combinations 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. A high proportion of students at 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 values 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 percentages of students at the LOSS and HOSS were small: less than 1%. However, in some grades and content areas, the LOSS percentages were larger. In Mathematics, all grades had more than 1% of students at the LOSS, ranging from 1.18% in Grade 3 to 5.66% in Grade 7. 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 response patterns of students at the LOSS in Mathematics were investigated after the Spring 2018 test administration. It was found that these students typically answered fewer than ten MC items and none of the non-MC items, which resulted in their LOSS values. For these students to receive a scale score above the LOSS, they would need to correctly answer more items, including some non-MC items. Non-MC items do not assume guessing, so the correct responses tend to represent student ability more accurately. The percentage of students at the LOSS in Mathematics may be reduced in future years by including some additional items, particularly non-MC items, that are less difficult. Also, the percentage of students at the LOSS in Social Studies grade 10 was found to be higher than 1%. The percentages of students scoring at the HOSS were less than 1% in all grades and content areas.

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 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	62904		63194		
Gender					
Male	32196	51.18	32397	51.27	0.08
Female	30708	48.82	30797	48.73	-0.08
Race/Ethnicity					
White	41278	65.62	41366	65.46	-0.16
Black	6938	11.03	7054	11.16	0.13
Hispanic	8540	13.58	8597	13.60	0.03
Asian/Pacific Islander	2729	4.34	2737	4.33	-0.01
American Indian	741	1.18	750	1.19	0.01
Other	2678	4.26	2690	4.26	0.00
LEP					
No	57080	90.74	57340	90.74	0.00
Yes	5824	9.26	5854	9.26	0.00
Disability					
No	55244	87.82	55465	87.77	-0.05
Yes	7660	12.18	7729	12.23	0.05
SES Disadvantaged					
No	34783	55.30	34850	55.15	-0.15
Yes	28121	44.70	28344	44.85	0.15
Grade 4	N	%	N	%	
All Students	64111		64354		
Gender					
Male	32601	50.85	32746	50.88	0.03
Female	31510	49.15	31608	49.12	-0.03
Race/Ethnicity					
White	42420	66.17	42490	66.03	-0.14
Black	7001	10.92	7102	11.04	0.12
Hispanic	8695	13.56	8736	13.57	0.01
Asian/Pacific Islander	2593	4.04	2597	4.04	-0.01
American Indian	759	1.18	762	1.18	0.00
Other	2643	4.12	2667	4.14	0.02
LEP					
No	58358	91.03	58568	91.01	-0.02
Yes	5753	8.97	5786	8.99	0.02
Disability					
No	56398	87.97	56578	87.92	-0.05
Yes	7713	12.03	7776	12.08	0.05
SES Disadvantaged					
No	35524	55.41	35587	55.30	-0.11
Yes	28587	44.59	28767	44.70	0.11

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

Grade 5	Calibration Sample		Population		Difference
	N	%	N	%	%
All Students	64713		64903		
Gender					
Male	33140	51.21	33254	51.24	0.03
Female	31573	48.79	31649	48.76	-0.03
Race/Ethnicity					
White	43196	66.75	43253	66.64	-0.11
Black	6927	10.70	7007	10.80	0.09
Hispanic	8671	13.40	8697	13.40	0.00
Asian/Pacific Islander	2551	3.94	2556	3.94	0.00
American Indian	798	1.23	802	1.24	0.00
Other	2570	3.97	2588	3.99	0.02
LEP					
No	59889	92.55	60063	92.54	0.00
Yes	4824	7.45	4840	7.46	0.00
Disability					
No	56868	87.88	56999	87.82	-0.06
Yes	7845	12.12	7904	12.18	0.06
SES Disadvantaged					
No	36698	56.71	36740	56.61	-0.10
Yes	28015	43.29	28163	43.39	0.10
Grade 6	N	%	N	%	
All Students	63331		63600		
Gender					
Male	32427	51.20	32607	51.27	0.07
Female	30904	48.80	30993	48.73	-0.07
Race/Ethnicity					
White	42551	67.19	42634	67.03	-0.15
Black	6619	10.45	6726	10.58	0.12
Hispanic	8470	13.37	8529	13.41	0.04
Asian/Pacific Islander	2553	4.03	2559	4.02	-0.01
American Indian	796	1.26	797	1.25	0.00
Other	2342	3.70	2355	3.70	0.00
LEP					
No	59761	94.36	60004	94.35	-0.02
Yes	3570	5.64	3596	5.65	0.02
Disability					
No	55777	88.07	55971	88.00	-0.07
Yes	7554	11.93	7629	12.00	0.07
SES Disadvantaged					
No	36521	57.67	36579	57.51	-0.15
Yes	26810	42.33	27021	42.49	0.15

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

Grade 7	Calibration Sample		Population		Difference
	N	%	N	%	%
All Students	62903		63140		
Gender					
Male	32101	51.03	32237	51.06	0.02
Female	30802	48.97	30903	48.94	-0.02
Race/Ethnicity					
White	43249	68.76	43341	68.64	-0.11
Black	6353	10.10	6432	10.19	0.09
Hispanic	7910	12.57	7951	12.59	0.02
Asian/Pacific Islander	2435	3.87	2444	3.87	0.00
American Indian	782	1.24	786	1.24	0.00
Other	2174	3.46	2186	3.46	0.01
LEP					
No	59913	95.25	60127	95.23	-0.02
Yes	2990	4.75	3013	4.77	0.02
Disability					
No	55590	88.37	55757	88.31	-0.07
Yes	7313	11.63	7383	11.69	0.07
SES Disadvantaged					
No	37853	60.18	37925	60.06	-0.11
Yes	25050	39.82	25215	39.94	0.11
Grade 8	N	%	N	%	
All Students	62901		63248		
Gender					
Male	32333	51.40	32528	51.43	0.03
Female	30568	48.60	30720	48.57	-0.03
Race/Ethnicity					
White	43763	69.57	43892	69.40	-0.18
Black	6135	9.75	6259	9.90	0.14
Hispanic	7759	12.34	7811	12.35	0.01
Asian/Pacific Islander	2448	3.89	2454	3.88	-0.01
American Indian	780	1.24	787	1.24	0.00
Other	2016	3.21	2045	3.23	0.03
LEP					
No	60137	95.61	60465	95.60	-0.01
Yes	2764	4.39	2783	4.40	0.01
Disability					
No	55545	88.31	55790	88.21	-0.10
Yes	7356	11.69	7458	11.79	0.10
SES Disadvantaged					
No	38560	61.30	38649	61.11	-0.20
Yes	24341	38.70	24599	38.89	0.20

Table 6-2 Mathematics Calibration Sample Demographics Compared to Population

Grade 3	Calibration Sample		Population		Difference
	N	%	N	%	%
All Students	63220		63314		
Gender					
Male	32397	51.24	32454	51.26	0.01
Female	30823	48.76	30860	48.74	-0.01
Race/Ethnicity					
White	41354	65.41	41384	65.36	-0.05
Black	7006	11.08	7045	11.13	0.05
Hispanic	8659	13.70	8671	13.70	0.00
Asian/Pacific Islander	2767	4.38	2770	4.38	0.00
American Indian	747	1.18	751	1.19	0.00
Other	2687	4.25	2693	4.25	0.00
LEP					
No	57252	90.56	57340	90.56	0.00
Yes	5968	9.44	5974	9.44	0.00
Disability					
No	55505	87.80	55575	87.78	-0.02
Yes	7715	12.20	7739	12.22	0.02
SES Disadvantaged					
No	34879	55.17	34897	55.12	-0.05
Yes	28341	44.83	28417	44.88	0.05
Grade 4	N	%	N	%	
All Students	64354		64462		
Gender					
Male	32748	50.89	32802	50.89	0.00
Female	31606	49.11	31660	49.11	0.00
Race/Ethnicity					
White	42471	66.00	42504	65.94	-0.06
Black	7055	10.96	7102	11.02	0.05
Hispanic	8806	13.68	8818	13.68	0.00
Asian/Pacific Islander	2611	4.06	2613	4.05	0.00
American Indian	758	1.18	761	1.18	0.00
Other	2653	4.12	2664	4.13	0.01
LEP					
No	58472	90.86	58571	90.86	0.00
Yes	5882	9.14	5891	9.14	0.00
Disability					
No	56589	87.93	56679	87.93	-0.01
Yes	7765	12.07	7783	12.07	0.01
SES Disadvantaged					
No	35592	55.31	35619	55.26	-0.05
Yes	28762	44.69	28843	44.74	0.05

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

Grade 5	Calibration Sample		Population		Difference
	N	%	N	%	%
All Students	64915		65021		
Gender					
Male	33262	51.24	33318	51.24	0.00
Female	31653	48.76	31703	48.76	0.00
Race/Ethnicity					
White	43231	66.60	43266	66.54	-0.05
Black	6956	10.72	7000	10.77	0.05
Hispanic	8775	13.52	8790	13.52	0.00
Asian/Pacific Islander	2575	3.97	2577	3.96	0.00
American Indian	799	1.23	803	1.23	0.00
Other	2579	3.97	2585	3.98	0.00
LEP					
No	59971	92.38	60067	92.38	0.00
Yes	4944	7.62	4954	7.62	0.00
Disability					
No	57024	87.84	57103	87.82	-0.02
Yes	7891	12.16	7918	12.18	0.02
SES Disadvantaged					
No	36760	56.63	36789	56.58	-0.05
Yes	28155	43.37	28232	43.42	0.05
Grade 6	N	%	N	%	
All Students	63519		63669		
Gender					
Male	32557	51.26	32645	51.27	0.02
Female	30962	48.74	31024	48.73	-0.02
Race/Ethnicity					
White	42592	67.05	42638	66.97	-0.09
Black	6671	10.50	6730	10.57	0.07
Hispanic	8551	13.46	8584	13.48	0.02
Asian/Pacific Islander	2564	4.04	2568	4.03	0.00
American Indian	794	1.25	796	1.25	0.00
Other	2347	3.69	2353	3.70	0.00
LEP					
No	59857	94.23	59992	94.22	-0.01
Yes	3662	5.77	3677	5.78	0.01
Disability					
No	55916	88.03	56034	88.01	-0.02
Yes	7603	11.97	7635	11.99	0.02
SES Disadvantaged					
No	36562	57.56	36591	57.47	-0.09
Yes	26957	42.44	27078	42.53	0.09

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

Grade 7	Calibration Sample		Population		Difference
	N	%	N	%	%
All Students	63104		63218		
Gender					
Male	32225	51.07	32279	51.06	-0.01
Female	30879	48.93	30939	48.94	0.01
Race/Ethnicity					
White	43283	68.59	43329	68.54	-0.05
Black	6396	10.14	6436	10.18	0.04
Hispanic	8015	12.70	8033	12.71	0.01
Asian/Pacific Islander	2448	3.88	2452	3.88	0.00
American Indian	781	1.24	782	1.24	0.00
Other	2181	3.46	2186	3.46	0.00
LEP					
No	60009	95.10	60112	95.09	-0.01
Yes	3095	4.90	3106	4.91	0.01
Disability					
No	55760	88.36	55838	88.33	-0.04
Yes	7344	11.64	7380	11.67	0.04
SES Disadvantaged					
No	37904	60.07	37946	60.02	-0.04
Yes	25200	39.93	25272	39.98	0.04
Grade 8	N	%	N	%	
All Students	63191		63318		
Gender					
Male	32504	51.44	32566	51.43	-0.01
Female	30687	48.56	30752	48.57	0.01
Race/Ethnicity					
White	43847	69.39	43895	69.32	-0.06
Black	6204	9.82	6252	9.87	0.06
Hispanic	7863	12.44	7878	12.44	0.00
Asian/Pacific Islander	2461	3.89	2465	3.89	0.00
American Indian	787	1.25	790	1.25	0.00
Other	2029	3.21	2038	3.22	0.01
LEP					
No	60326	95.47	60446	95.46	0.00
Yes	2865	4.53	2872	4.54	0.00
Disability					
No	55779	88.27	55858	88.22	-0.05
Yes	7412	11.73	7460	11.78	0.05
SES Disadvantaged					
No	38647	61.16	38679	61.09	-0.07
Yes	24544	38.84	24639	38.91	0.07

Table 6-3 Science Calibration Sample Demographics Compared to Population

Grade 4	Calibration Sample		Population		Difference
	N	%	N	%	%
All Students	64352		64448		
Gender					
Male	32754	50.90	32802	50.90	0.00
Female	31598	49.10	31646	49.10	0.00
Race/Ethnicity					
White	42469	65.99	42500	65.94	-0.05
Black	7057	10.97	7093	11.01	0.04
Hispanic	8807	13.69	8819	13.68	0.00
Asian/Pacific Islander	2607	4.05	2611	4.05	0.00
American Indian	759	1.18	761	1.18	0.00
Other	2653	4.12	2664	4.13	0.01
LEP					
No	58473	90.86	58562	90.87	0.00
Yes	5879	9.14	5886	9.13	0.00
Disability					
No	56593	87.94	56670	87.93	-0.01
Yes	7759	12.06	7778	12.07	0.01
SES Disadvantaged					
No	35584	55.30	35610	55.25	-0.04
Yes	28768	44.70	28838	44.75	0.04
Grade 8	N	%	N	%	
All Students	63099		63272		
Gender					
Male	32451	51.43	32542	51.43	0.00
Female	30648	48.57	30730	48.57	0.00
Race/Ethnicity					
White	43814	69.44	43877	69.35	-0.09
Black	6161	9.76	6230	9.85	0.08
Hispanic	7850	12.44	7875	12.45	0.01
Asian/Pacific Islander	2459	3.90	2464	3.89	0.00
American Indian	788	1.25	790	1.25	0.00
Other	2027	3.21	2036	3.22	0.01
LEP					
No	60239	95.47	60403	95.47	0.00
Yes	2860	4.53	2869	4.53	0.00
Disability					
No	55706	88.28	55826	88.23	-0.05
Yes	7393	11.72	7446	11.77	0.05
SES Disadvantaged					
No	38617	61.20	38669	61.12	-0.09
Yes	24482	38.80	24603	38.88	0.09

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

Grade 4	Calibration Sample		Population		Difference
	N	%	N	%	%
All Students	64359		64456		
Gender					
Male	32757	50.90	32806	50.90	0.00
Female	31602	49.10	31650	49.10	0.00
Race/Ethnicity					
White	42465	65.98	42500	65.94	-0.04
Black	7068	10.98	7097	11.01	0.03
Hispanic	8805	13.68	8818	13.68	0.00
Asian/Pacific Islander	2607	4.05	2611	4.05	0.00
American Indian	760	1.18	762	1.18	0.00
Other	2654	4.12	2668	4.14	0.02
LEP					
No	58479	90.86	58570	90.87	0.00
Yes	5880	9.14	5886	9.13	0.00
Disability					
No	56596	87.94	56671	87.92	-0.02
Yes	7763	12.06	7785	12.08	0.02
SES Disadvantaged					
No	35588	55.30	35613	55.25	-0.04
Yes	28771	44.70	28843	44.75	0.04
Grade 8	N	%	N	%	
All Students	63078		63230		
Gender					
Male	32431	51.41	32525	51.44	0.03
Female	30647	48.59	30705	48.56	-0.03
Race/Ethnicity					
White	43808	69.45	43863	69.37	-0.08
Black	6155	9.76	6209	9.82	0.06
Hispanic	7845	12.44	7871	12.45	0.01
Asian/Pacific Islander	2460	3.90	2463	3.90	0.00
American Indian	786	1.25	789	1.25	0.00
Other	2024	3.21	2035	3.22	0.01
LEP					
No	60225	95.48	60364	95.47	-0.01
Yes	2853	4.52	2866	4.53	0.01
Disability					
No	55693	88.29	55792	88.24	-0.06
Yes	7385	11.71	7438	11.76	0.06
SES Disadvantaged					
No	38608	61.21	38648	61.12	-0.08
Yes	24470	38.79	24582	38.88	0.08

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

Grade 10	Calibration Sample		Population		Difference
	N	%	N	%	%
All Students	61884		62630		
Gender					
Male	31478	50.87	31881	50.90	0.04
Female	30406	49.13	30749	49.10	-0.04
Race/Ethnicity					
White	45146	72.95	45373	72.45	-0.51
Black	4935	7.97	5261	8.40	0.43
Hispanic	6969	11.26	7091	11.32	0.06
Asian/Pacific Islander	2351	3.80	2378	3.80	0.00
American Indian	708	1.14	728	1.16	0.02
Other	1775	2.87	1799	2.87	0.00
LEP					
No	59415	96.01	60102	95.96	-0.05
Yes	2469	3.99	2528	4.04	0.05
Disability					
No	55514	89.71	56079	89.54	-0.17
Yes	6370	10.29	6551	10.46	0.17
SES Disadvantaged					
No	41321	66.77	41559	66.36	-0.42
Yes	20563	33.23	21071	33.64	0.42

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

Content	Grade	Item Number in Calibration	Type	N	Z	Critical Z
ELA	3	24	CR	62902	195.34	167.74
	3	33	CR	62902	192.93	167.74
	3	34*	CR	62902	262.16	167.74
	4	1	CR	64108	181.87	170.95
	4	37	CR	64108	179.72	170.95
	5	7	CR	64705	617.16	172.55
	5	34*	CR	64705	358.68	172.55
	6	6*	CR	63329	246.19	168.88
	6	12*	CR	63329	577.81	168.88
	7	4	MC	62894	204.27	167.72
	7	20	CR	62894	209.79	167.72
	7	24	CR	62894	286.64	167.72
	7	31	CR	62894	255.43	167.72
	7	32*	CR	62894	343.21	167.72
	8	5*	MC	62878	220.33	167.67
	8	24*	CR	62878	993.78	167.67
	8	32	CR	62878	187.82	167.67
8	34	CR	62878	195.31	167.67	
Math	5	10	CR	64873	202.59	172.99
	5	25	CR	64873	205.04	172.99
	5	35*	CR	64873	197.44	172.99
Social Studies	10	11	MC	61668	172.21	164.45
	10	50	MC	61668	312.68	164.45

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-Parameter		b-Parameter		A	B
			# Iterations	F Value	Corr	# RMSD Outliers	Corr	# RMSD Outliers		
ELA	3	16	6	0.0775	0.99	1	0.99	1	0.9432	-1.1417
	4	16	5	0.1082	0.98	1	0.96	1	1.0401	-0.5735
	5	13	7	0.1096	0.98	1	0.97	0	0.9676	-0.1201
	6	15	7	0.0777	0.98	0	0.99	0	1.0238	0.0155
	7	14	8	0.2383	0.99	1	0.98	0	1.1326	0.4440
	8	17	5	0.1259	0.98	1	0.97	1	1.1997	0.6014
Math	3	30	2	0.0340	0.97	2	0.99	0	0.9172	-1.1782
	4	31	10	0.0264	0.98	2	0.99	2	0.9385	-0.7134
	5	20	25	0.0502	0.98	1	0.98	1	0.9086	-0.1875
	6	22	9	0.0648	0.98	1	1.00	1	1.0020	0.0861
	7	24	32	0.0975	0.98	1	0.99	1	1.0329	0.3963
	8	20	30	0.0648	0.99	0	0.96	1	1.0036	0.8061
Science	4	40	14	0.0514	1.00	2	1.00	1	1.0644	-0.1050
	8	40	15	0.0508	0.98	2	1.00	2	1.1132	-0.2170
Social Studies	4	14	31	0.1666	0.99	1	0.96	1	1.0979	-0.1283
	8	14	22	0.0396	0.99	0	0.99	2	1.0507	0.0626
	10	18	5	0.0643	0.99	1	0.99	1	1.0978	-0.0650

Table 6-7 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 6-8 Scoring Table for English Language Arts Grade 3

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	330	92	31	563	12
1	330	92	32	567	12
2	330	92	33	571	12
3	330	92	34	575	12
4	330	92	35	579	12
5	339	83	36	583	13
6	396	42	37	588	13
7	421	32	38	592	13
8	437	27	39	597	13
9	450	24	40	602	14
10	461	21	41	607	14
11	470	20	42	613	15
12	477	18	43	619	16
13	484	17	44	626	17
14	491	16	45	634	18
15	497	16	46	643	19
16	502	15	47	654	21
17	507	14	48	666	24
18	512	14	49	682	26
19	516	14	50	700	30
20	521	13	51	725	35
21	525	13	52	763	47
22	529	13	53	900	161
23	533	12			
24	537	12			
25	541	12			
26	544	12			
27	548	12			
28	552	12			
29	556	12			
30	559	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 6-9 Scoring Table for English Language Arts Grade 4

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	340	63	31	583	13
1	340	63	32	587	13
2	340	63	33	591	12
3	340	63	34	595	12
4	340	63	35	599	13
5	340	63	36	603	13
6	341	63	37	608	13
7	387	44	38	612	13
8	414	36	39	616	13
9	433	32	40	621	13
10	449	29	41	626	14
11	462	26	42	631	14
12	474	25	43	637	15
13	484	23	44	643	15
14	493	22	45	649	16
15	501	21	46	656	17
16	509	20	47	663	18
17	516	19	48	671	19
18	522	18	49	680	20
19	528	17	50	690	21
20	534	17	51	701	23
21	539	16	52	714	26
22	544	15	53	731	30
23	549	15	54	753	37
24	554	14	55	792	54
25	558	14	56	930	176
26	562	14			
27	567	13			
28	571	13			
29	575	13			
30	579	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 6-10 Scoring Table for English Language Arts Grade 5

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	350	99	31	598	12
1	350	99	32	602	12
2	350	99	33	606	12
3	350	99	34	610	12
4	350	99	35	614	12
5	350	99	36	617	12
6	350	99	37	621	12
7	359	92	38	625	12
8	424	51	39	630	12
9	454	38	40	634	12
10	473	32	41	639	13
11	488	28	42	643	13
12	499	25	43	649	14
13	509	22	44	654	14
14	517	21	45	660	15
15	525	19	46	666	16
16	531	18	47	674	17
17	537	17	48	681	18
18	543	16	49	690	19
19	548	15	50	700	21
20	553	15	51	712	23
21	558	14	52	725	25
22	562	14	53	742	28
23	567	14	54	764	35
24	571	13	55	802	50
25	575	13	56	940	173
26	579	13			
27	583	13			
28	587	12			
29	591	12			
30	595	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 6-11 Scoring Table for English Language Arts Grade 6

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	360	64	31	603	14
1	360	64	32	607	14
2	360	64	33	612	14
3	360	64	34	616	14
4	360	64	35	621	15
5	373	58	36	625	15
6	413	42	37	630	15
7	438	35	38	635	15
8	456	31	39	640	15
9	471	28	40	645	16
10	483	26	41	651	16
11	493	24	42	656	16
12	503	22	43	662	17
13	511	21	44	668	17
14	519	20	45	675	18
15	525	19	46	682	18
16	532	18	47	689	19
17	538	18	48	698	20
18	544	17	49	707	21
19	549	16	50	717	23
20	554	16	51	728	24
21	559	15	52	742	27
22	564	15	53	758	30
23	568	15	54	778	34
24	573	15	55	805	41
25	577	14	56	848	56
26	582	14	57	950	133
27	586	14			
28	590	14			
29	594	14			
30	599	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 7

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	370	65	31	624	14
1	370	65	32	629	14
2	370	65	33	633	14
3	370	65	34	637	14
4	370	65	35	641	14
5	370	65	36	645	14
6	393	55	37	650	14
7	428	44	38	654	14
8	453	38	39	659	14
9	473	34	40	664	14
10	488	31	41	669	15
11	502	29	42	674	15
12	513	27	43	679	16
13	524	25	44	685	16
14	533	23	45	691	17
15	541	22	46	698	18
16	549	21	47	706	19
17	556	20	48	714	20
18	562	19	49	724	22
19	568	18	50	735	24
20	574	18	51	748	27
21	579	17	52	764	31
22	584	17	53	784	36
23	589	16	54	812	45
24	594	16	55	859	65
25	599	15	56	960	143
26	603	15			
27	608	15			
28	612	14			
29	616	14			
30	620	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 8

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	380	75	31	623	14
1	380	75	32	627	14
2	380	75	33	631	14
3	380	75	34	636	14
4	380	75	35	640	14
5	380	75	36	645	15
6	380	75	37	649	15
7	430	52	38	654	15
8	461	41	39	659	15
9	482	35	40	664	16
10	498	30	41	670	16
11	510	28	42	675	17
12	521	25	43	681	17
13	530	23	44	688	18
14	539	22	45	694	18
15	546	20	46	702	19
16	553	19	47	709	19
17	559	18	48	717	19
18	565	17	49	726	20
19	570	17	50	737	22
20	575	16	51	748	24
21	580	16	52	763	28
22	585	15	53	782	34
23	589	15	54	808	42
24	594	14	55	851	59
25	598	14	56	970	155
26	602	14			
27	606	14			
28	611	14			
29	615	14			
30	619	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 6-14 Scoring Table for Mathematics Grade 3

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	360	110	26	568	10
1	360	110	27	572	10
2	360	110	28	575	10
3	360	110	29	579	10
4	360	110	30	584	11
5	360	110	31	588	11
6	417	54	32	592	11
7	452	33	33	597	11
8	470	26	34	602	12
9	483	21	35	607	12
10	492	19	36	613	13
11	500	17	37	620	14
12	507	15	38	628	15
13	513	14	39	638	17
14	519	13	40	652	21
15	524	13	41	677	33
16	529	12	42	760	108
17	533	12			
18	537	11			
19	541	11			
20	545	11			
21	549	11			
22	553	10			
23	557	10			
24	560	10			
25	564	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 6-15 Scoring Table for Mathematics Grade 4

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	405	113	26	596	10
1	405	113	27	600	10
2	405	113	28	603	9
3	405	113	29	606	9
4	405	113	30	610	9
5	405	113	31	613	9
6	405	113	32	616	9
7	405	113	33	620	9
8	457	61	34	623	10
9	488	37	35	627	10
10	505	29	36	631	10
11	518	24	37	635	10
12	528	21	38	639	11
13	536	19	39	644	11
14	543	17	40	649	12
15	550	16	41	656	13
16	556	14	42	663	15
17	561	14	43	672	17
18	566	13	44	685	21
19	570	12	45	707	30
20	574	12	46	800	112
21	578	11			
22	582	11			
23	586	10			
24	590	10			
25	593	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 5

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	430	120	26	628	9
1	430	120	27	631	9
2	430	120	28	634	9
3	430	120	29	637	9
4	430	120	30	640	9
5	430	120	31	643	9
6	430	120	32	646	9
7	496	55	33	650	9
8	530	32	34	653	10
9	546	24	35	657	10
10	557	20	36	661	10
11	566	17	37	665	10
12	573	15	38	669	11
13	579	14	39	674	11
14	584	13	40	680	12
15	589	12	41	686	13
16	593	12	42	693	14
17	597	11	43	702	16
18	601	11	44	715	20
19	605	10	45	736	29
20	608	10	46	830	114
21	612	10			
22	615	10			
23	618	9			
24	621	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 6-17 Scoring Table for Mathematics Grade 6

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	440	106	26	642	10
1	440	106	27	646	10
2	440	106	28	649	10
3	440	106	29	653	10
4	440	106	30	656	10
5	440	106	31	660	10
6	440	106	32	664	10
7	498	52	33	667	10
8	528	34	34	671	10
9	545	26	35	675	10
10	558	22	36	679	11
11	568	19	37	684	11
12	576	18	38	688	11
13	584	16	39	694	12
14	590	15	40	699	12
15	596	14	41	706	13
16	601	14	42	714	15
17	606	13	43	723	17
18	611	12	44	737	21
19	615	12	45	761	32
20	620	11	46	870	132
21	624	11			
22	628	11			
23	631	11			
24	635	10			
25	639	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 6-18 Scoring Table for Mathematics Grade 7

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	450	150	26	667	10
1	450	150	27	670	10
2	450	150	28	673	10
3	450	150	29	677	10
4	450	150	30	680	10
5	450	150	31	684	10
6	450	150	32	687	10
7	450	150	33	691	10
8	545	55	34	694	10
9	571	33	35	698	10
10	586	25	36	702	11
11	597	20	37	707	11
12	605	18	38	712	12
13	613	16	39	717	12
14	619	15	40	722	13
15	624	14	41	729	14
16	629	13	42	737	16
17	634	12	43	747	18
18	638	12	44	761	22
19	642	11	45	785	33
20	646	11	46	880	116
21	649	11			
22	653	10			
23	657	10			
24	660	10			
25	663	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 6-19 Scoring Table for Mathematics Grade 8

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	470	136	26	678	11
1	470	136	27	682	11
2	470	136	28	685	11
3	470	136	29	689	11
4	470	136	30	693	11
5	470	136	31	697	11
6	470	136	32	701	11
7	470	136	33	705	11
8	554	52	34	709	11
9	579	33	35	713	12
10	594	25	36	718	12
11	605	21	37	723	12
12	614	19	38	728	13
13	621	17	39	734	13
14	627	16	40	741	14
15	633	14	41	748	15
16	638	14	42	757	17
17	643	13	43	768	20
18	647	12	44	784	25
19	652	12	45	811	37
20	656	11	46	890	103
21	660	11			
22	663	11			
23	667	11			
24	671	11			
25	674	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 6-20 Scoring Table for Science Grade 4

Raw Score	Scale Score	SEM
0	190	106
1	190	106
2	190	106
3	190	106
4	190	106
5	190	106
6	190	106
7	190	106
8	212	84
9	249	49
10	270	35
11	285	28
12	296	23
13	306	21
14	314	19
15	321	17
16	328	16
17	334	16
18	340	15
19	346	15
20	351	15
21	357	14
22	362	14
23	367	14
24	372	14
25	377	14
26	383	14
27	388	14
28	393	14
29	399	14
30	405	14
31	411	15
32	418	15
33	426	17
34	435	18
35	445	20
36	458	23
37	474	27
38	498	34
39	540	54
40	600	103

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 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	393	114
9	449	58
10	472	38
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	545	14
20	549	13
21	554	13
22	559	13
23	564	13
24	569	13
25	573	13
26	578	13
27	583	13
28	589	13
29	594	13
30	600	14
31	606	14
32	613	15
33	620	16
34	628	17
35	638	18
36	650	21
37	664	24
38	685	30
39	720	45
40	770	79

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 Social Studies Grade 4

Raw Score	Scale Score	SEM
0	200	112
1	200	112
2	200	112
3	200	112
4	200	112
5	200	112
6	200	112
7	216	96
8	262	51
9	284	35
10	298	27
11	309	23
12	317	20
13	325	18
14	331	16
15	337	15
16	343	14
17	348	14
18	353	13
19	358	13
20	363	13
21	367	13
22	372	13
23	377	13
24	381	13
25	386	13
26	392	13
27	397	14
28	403	14
29	409	15
30	415	15
31	423	16
32	431	17
33	440	18
34	450	20
35	463	22
36	481	27
37	513	42
38	570	89

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 8

Raw Score	Scale Score	SEM
0	420	85
1	420	85
2	420	85
3	420	85
4	420	85
5	420	85
6	420	85
7	444	61
8	472	37
9	488	28
10	499	23
11	509	20
12	516	18
13	523	17
14	529	16
15	535	15
16	540	14
17	545	14
18	550	13
19	555	13
20	559	13
21	564	12
22	568	12
23	573	12
24	577	12
25	582	12
26	586	12
27	591	13
28	596	13
29	601	13
30	606	14
31	612	14
32	618	15
33	625	15
34	632	16
35	641	18
36	651	20
37	663	22
38	680	27
39	709	39
40	780	95

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 10

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	490	118	26	677	12
1	490	118	27	680	12
2	490	118	28	684	11
3	490	118	29	688	11
4	490	118	30	691	11
5	490	118	31	695	11
6	490	118	32	698	11
7	490	118	33	702	11
8	490	118	34	705	11
9	490	118	35	709	11
10	490	118	36	713	11
11	550	60	37	717	12
12	578	42	38	721	12
13	596	33	39	726	12
14	609	28	40	731	13
15	619	24	41	736	13
16	627	22	42	741	14
17	634	20	43	747	15
18	641	18	44	754	16
19	646	17	45	762	17
20	651	15	46	772	19
21	656	14	47	784	22
22	661	14	48	801	27
23	665	13	49	830	40
24	669	13	50	890	86
25	673	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 6-25 The Number and Percentage of Students at LOSS and HOSS

Content	Grade	LOSS	N	Percent	HOSS	N	Percent
ELA	3	330	23	.04	900	2	.00
	4	340	11	.02	930	3	.00
	5	350	27	.04	940	8	.01
	6	360	11	.02	950	2	.00
	7	370	17	.03	960	9	.01
	8	380	36	.06	970	23	.04
Math	3	360	750	1.18	760	176	.28
	4	405	1398	2.17	800	135	.21
	5	430	3037	4.67	830	40	.06
	6	440	1996	3.13	870	85	.13
	7	450	3578	5.66	880	52	.08
	8	470	2825	4.46	890	88	.14
Science	4	190	135	.21	600	326	.51
	8	390	396	.63	770	306	.48
Social Studies	4	200	531	.82	570	552	.86
	8	420	317	.50	780	622	.98
	10	490	919	1.47	890	217	.35

Figure 6-1 Anchor Set TCCs: ELA Grade 3

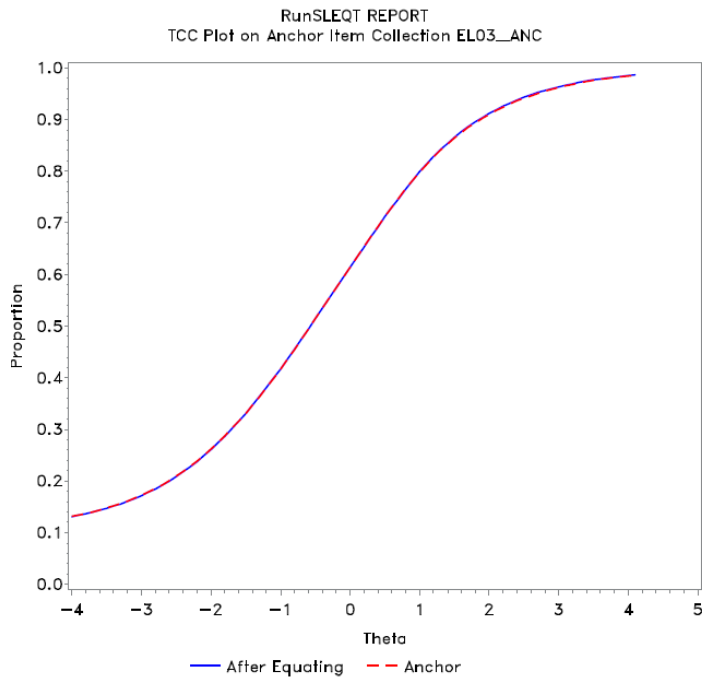


Figure 6-2 Anchor Set TCCs: ELA Grade 4

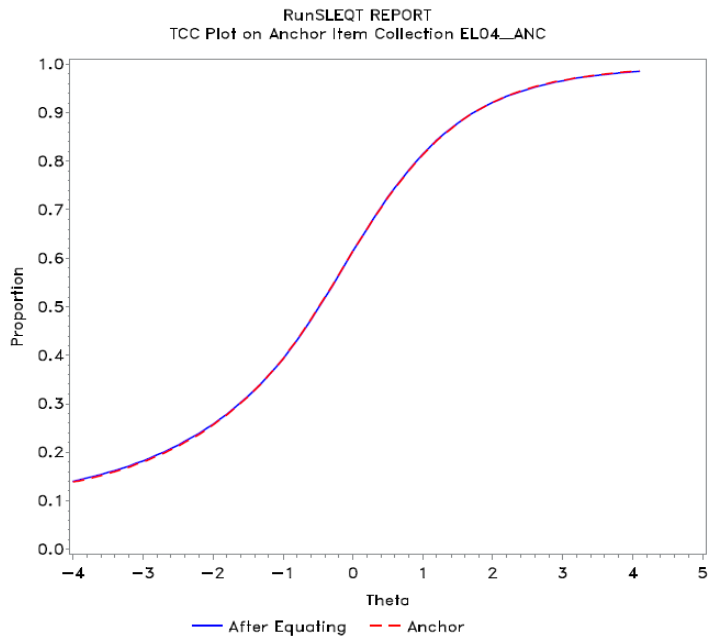


Figure 6-3 Anchor Set TCCs: ELA Grade 5

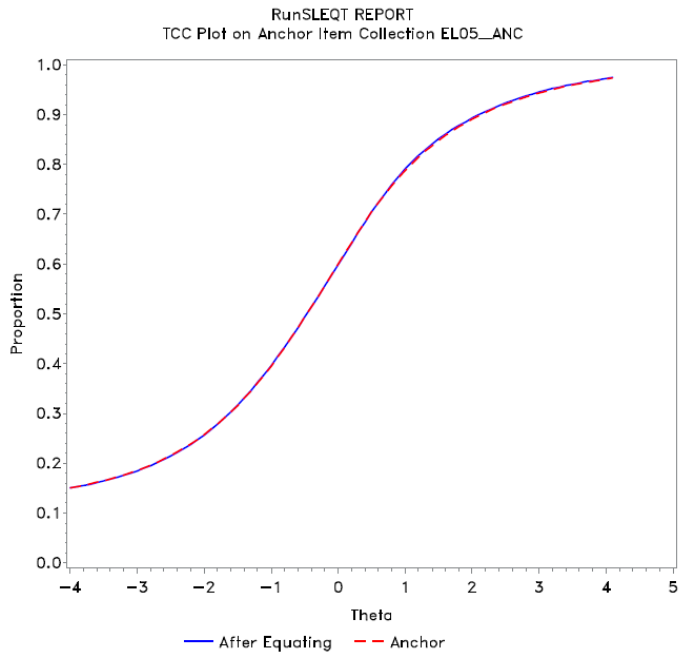


Figure 6-4 Anchor Set TCCs: ELA Grade 6

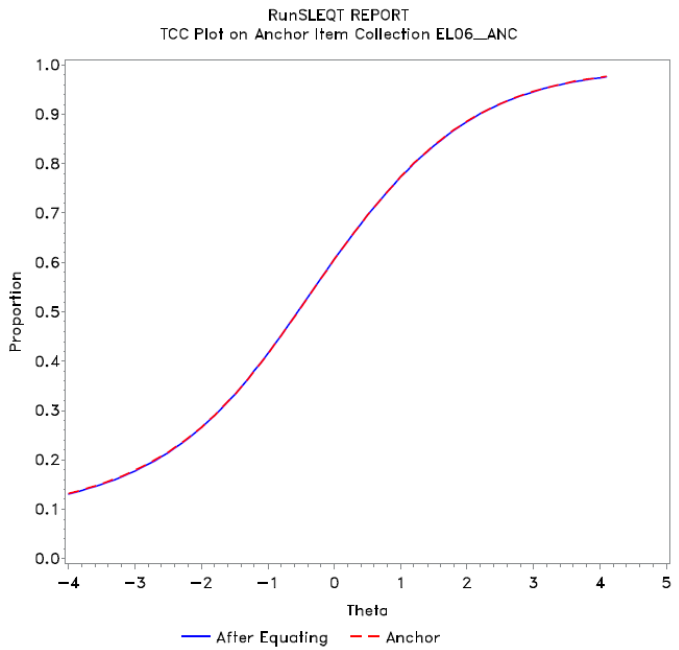


Figure 6-5 Anchor Set TCCs: ELA Grade 7

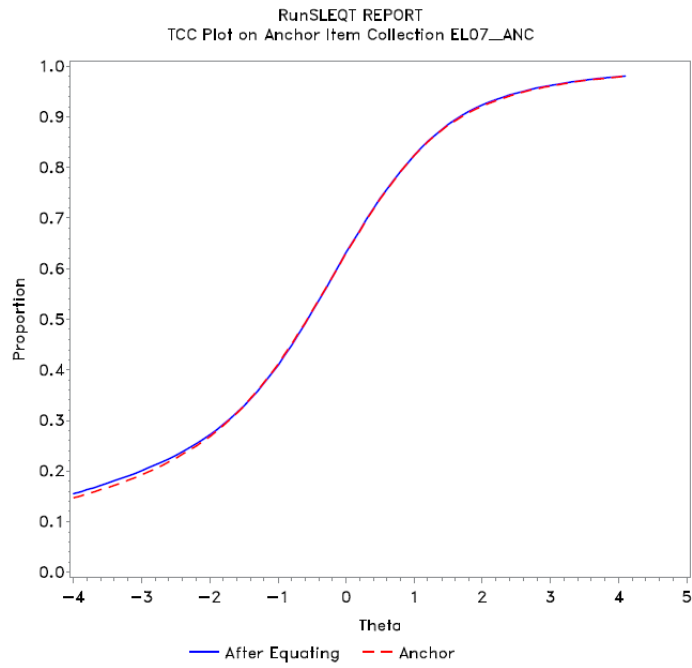


Figure 6-6 Anchor Set TCCs: ELA Grade 8

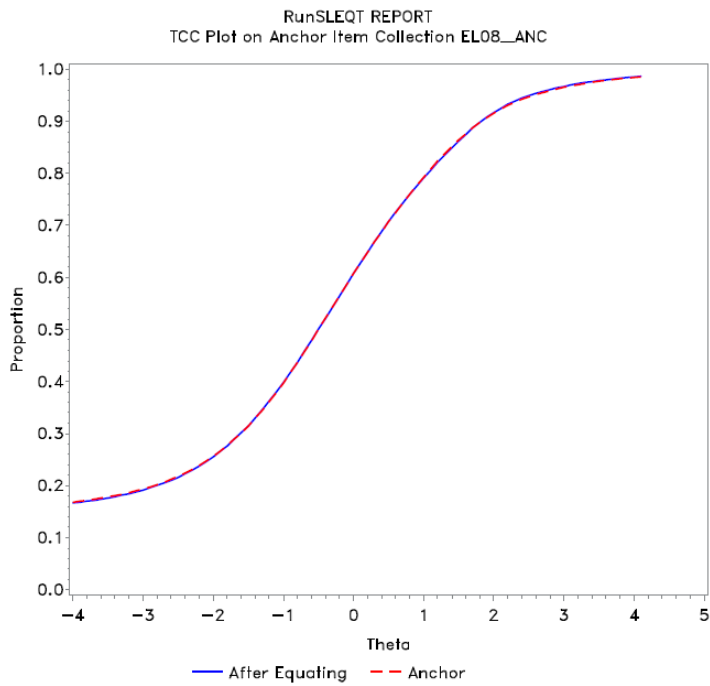


Figure 6-7 Anchor Set TCCs: Mathematics Grade 3

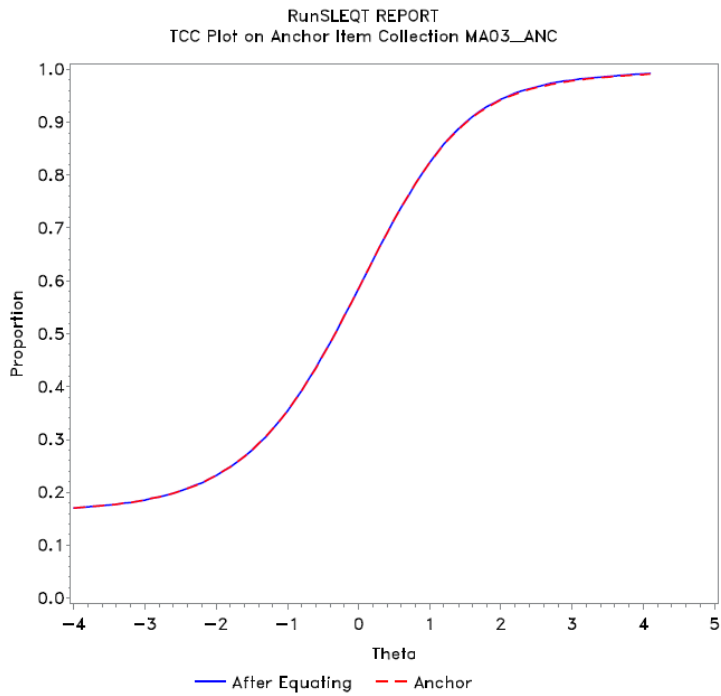


Figure 6-8 Anchor Set TCCs: Mathematics Grade 4

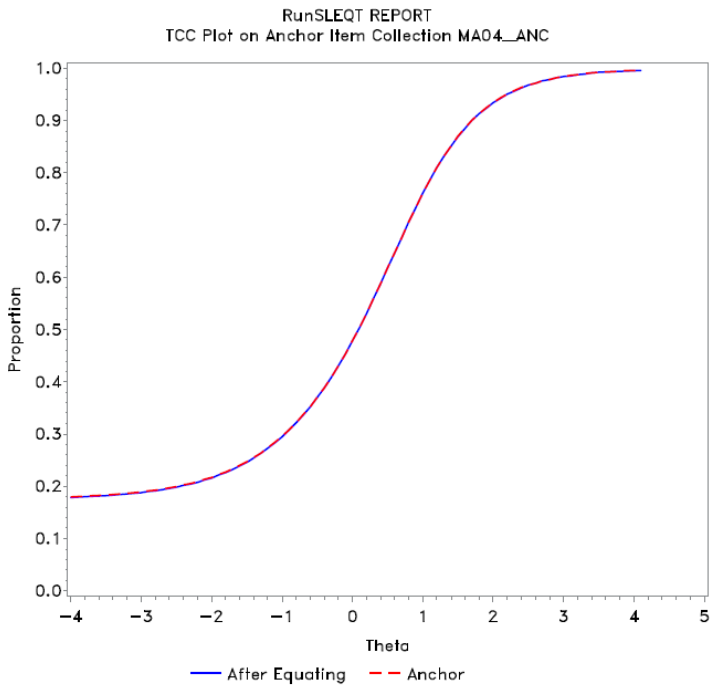


Figure 6-9 Anchor Set TCCs: Mathematics Grade 5

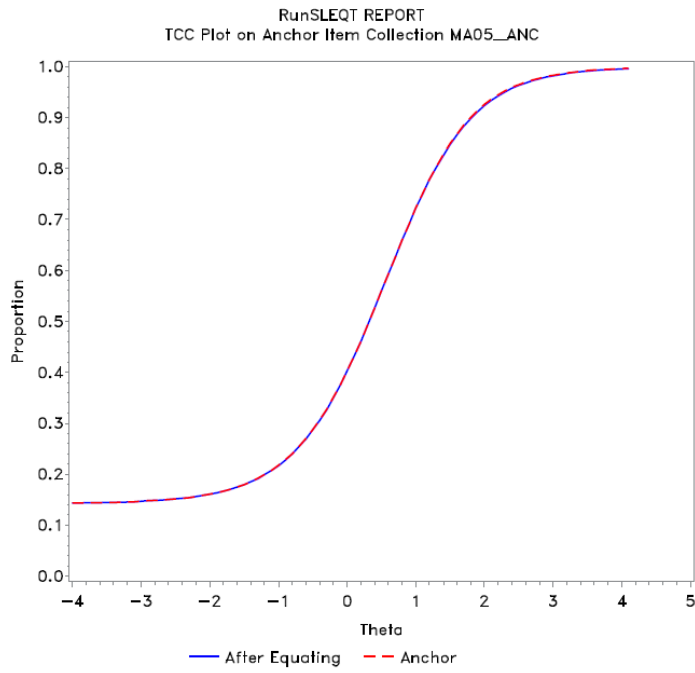


Figure 6-10 Anchor Set TCCs: Mathematics Grade 6

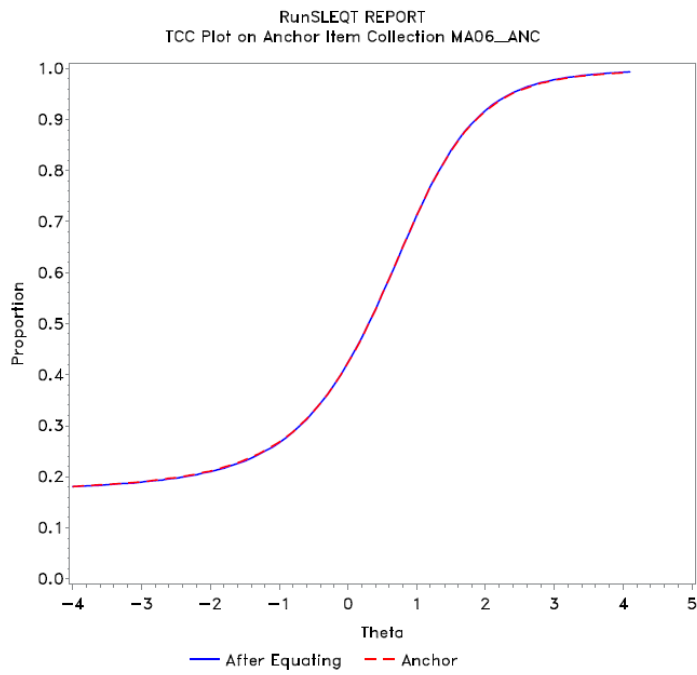


Figure 6-11 Anchor Set TCCs: Mathematics Grade 7

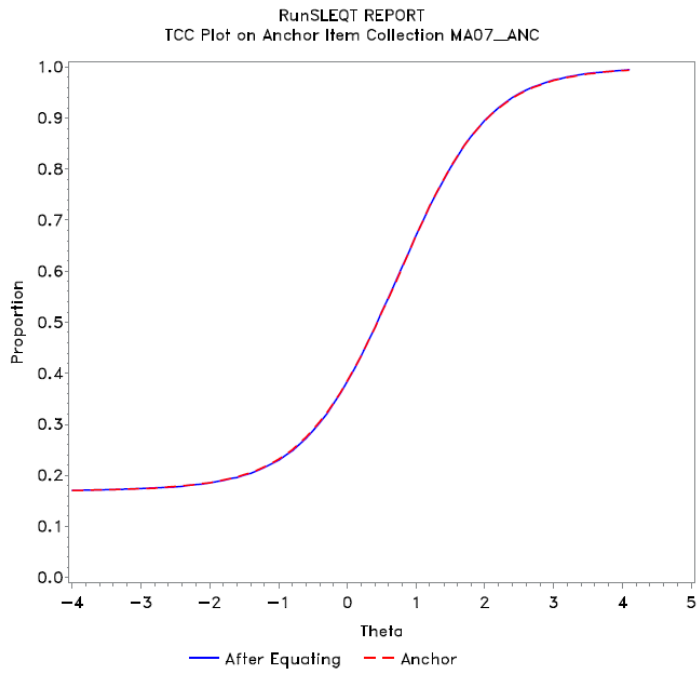


Figure 6-12 Anchor Set TCCs: Mathematics Grade 8

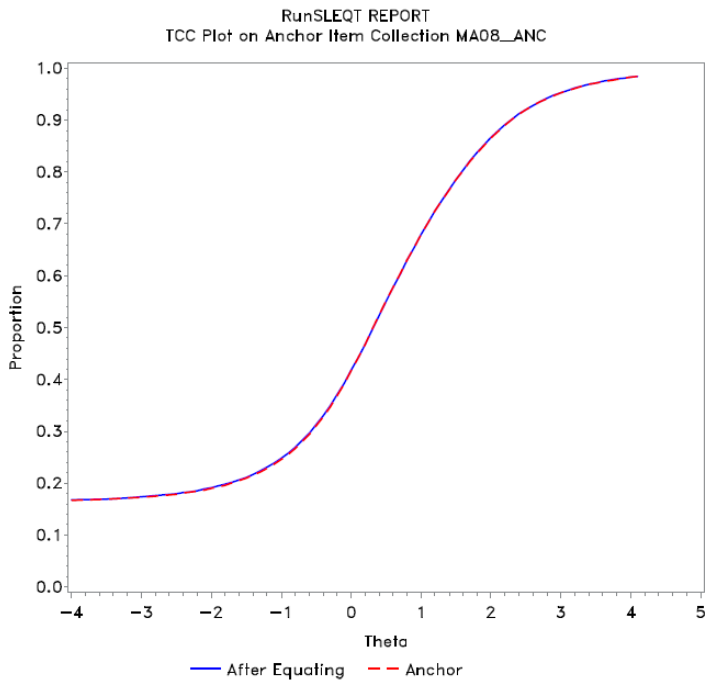


Figure 6-13 Anchor Set TCCs: Science Grade 4

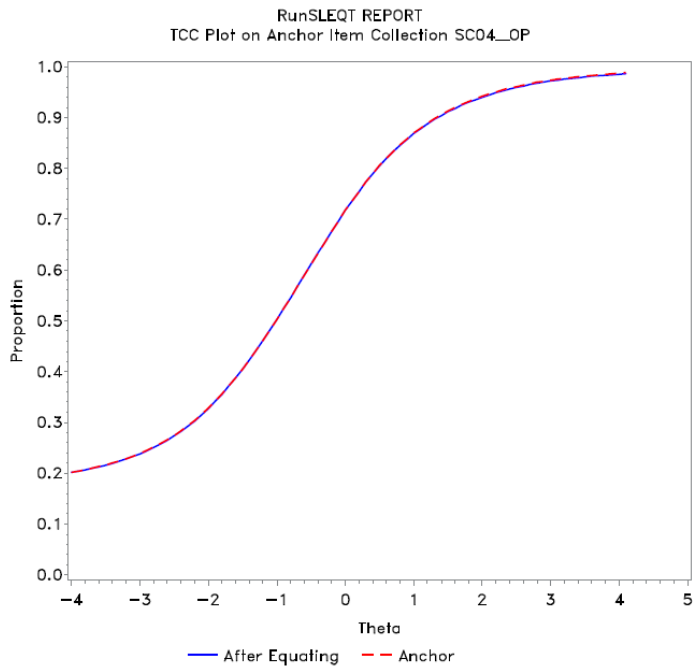


Figure 6-14 Anchor Set TCCs: Science Grade 8

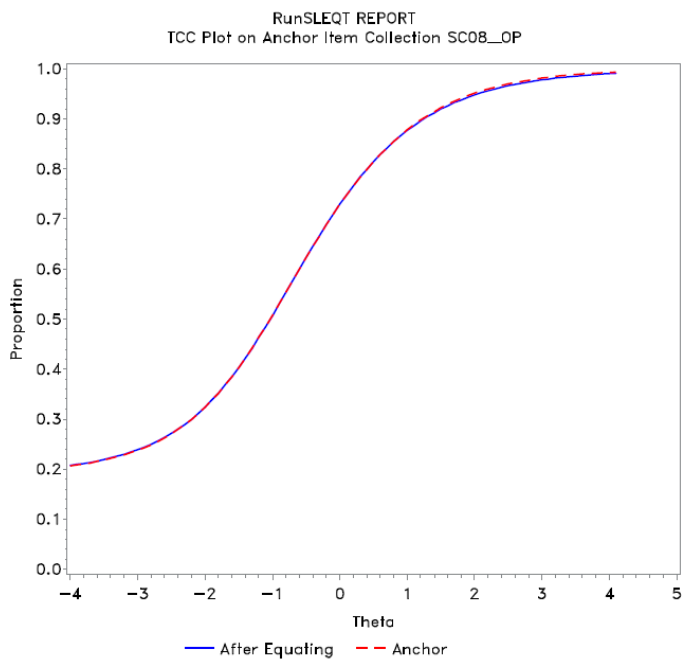


Figure 6-15 Anchor Set TCCs: Social Studies Grade 4

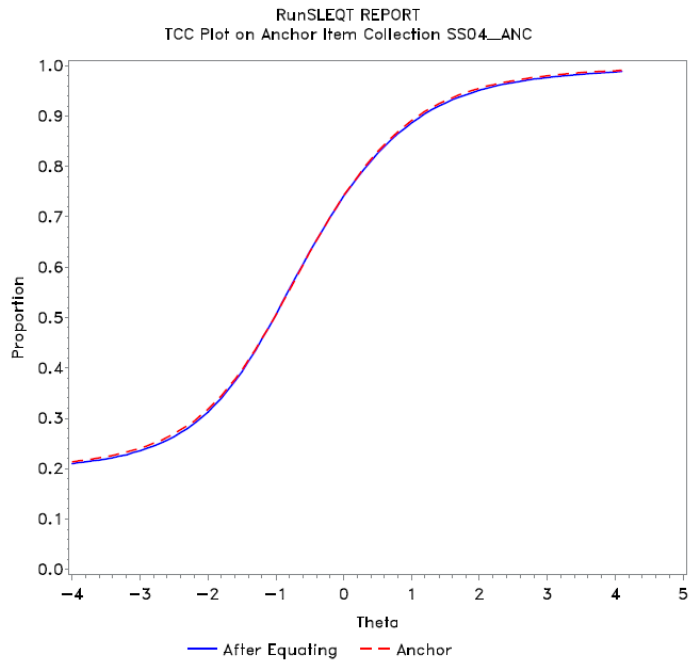


Figure 6-16 Anchor Set TCCs: Social Studies Grade 8

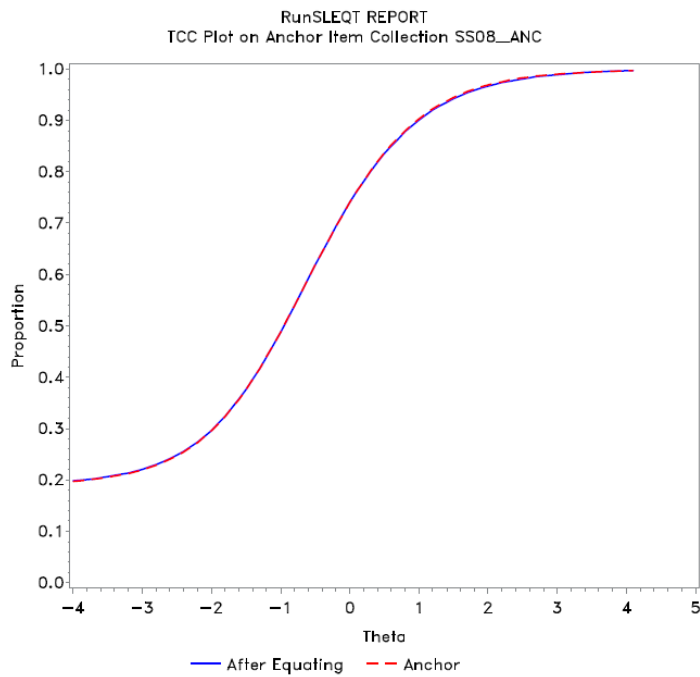


Figure 6-17 Anchor Set TCCs: Social Studies Grade 10

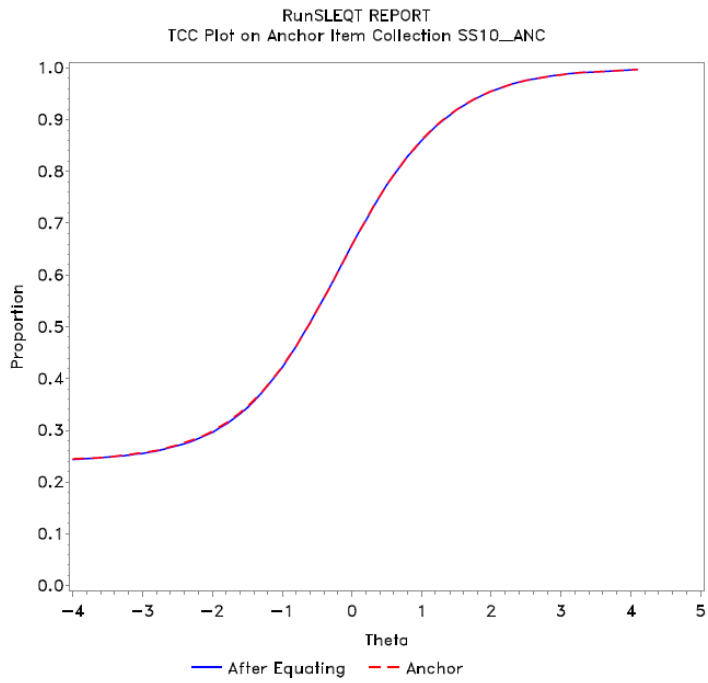


Figure 6-18 English Language Arts Test Characteristic Curves

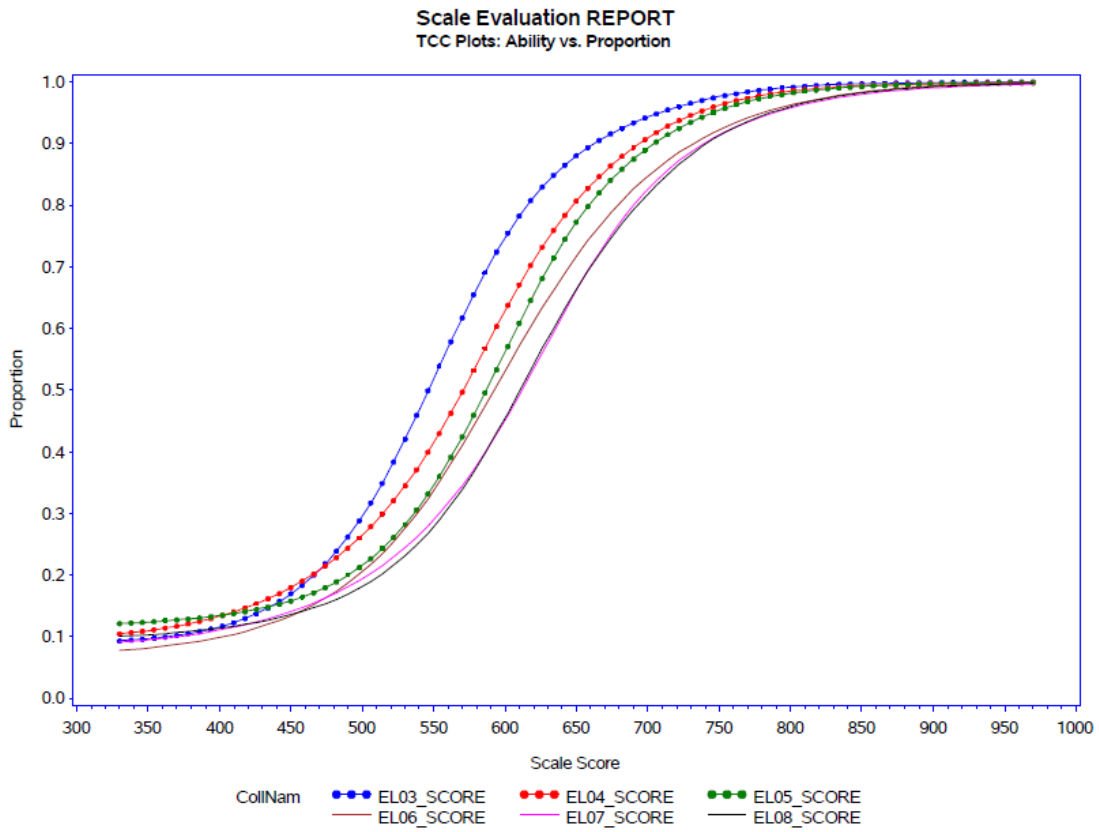


Figure 6-19 English Language Arts Standard Error Curves

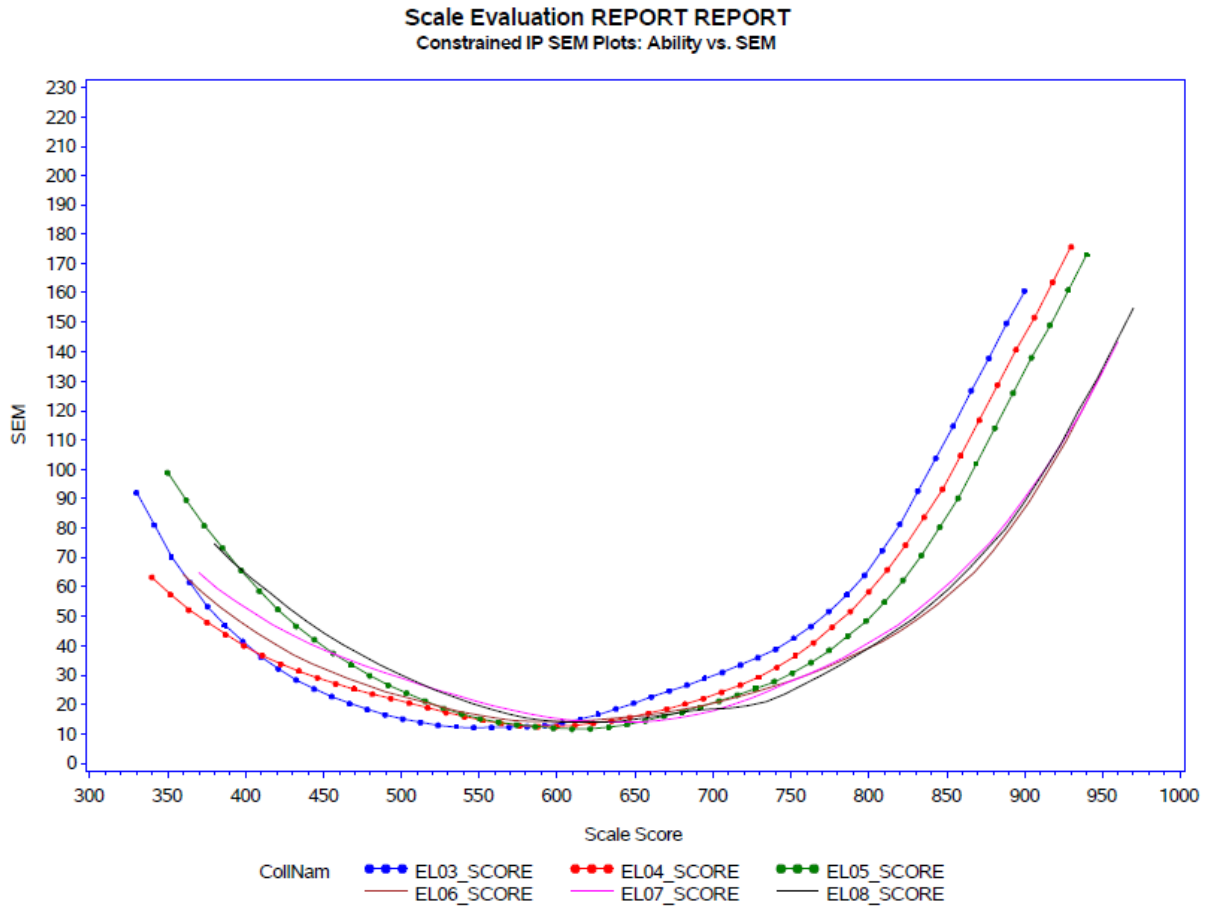


Figure 6-20 English Language Arts Growth at Quartiles

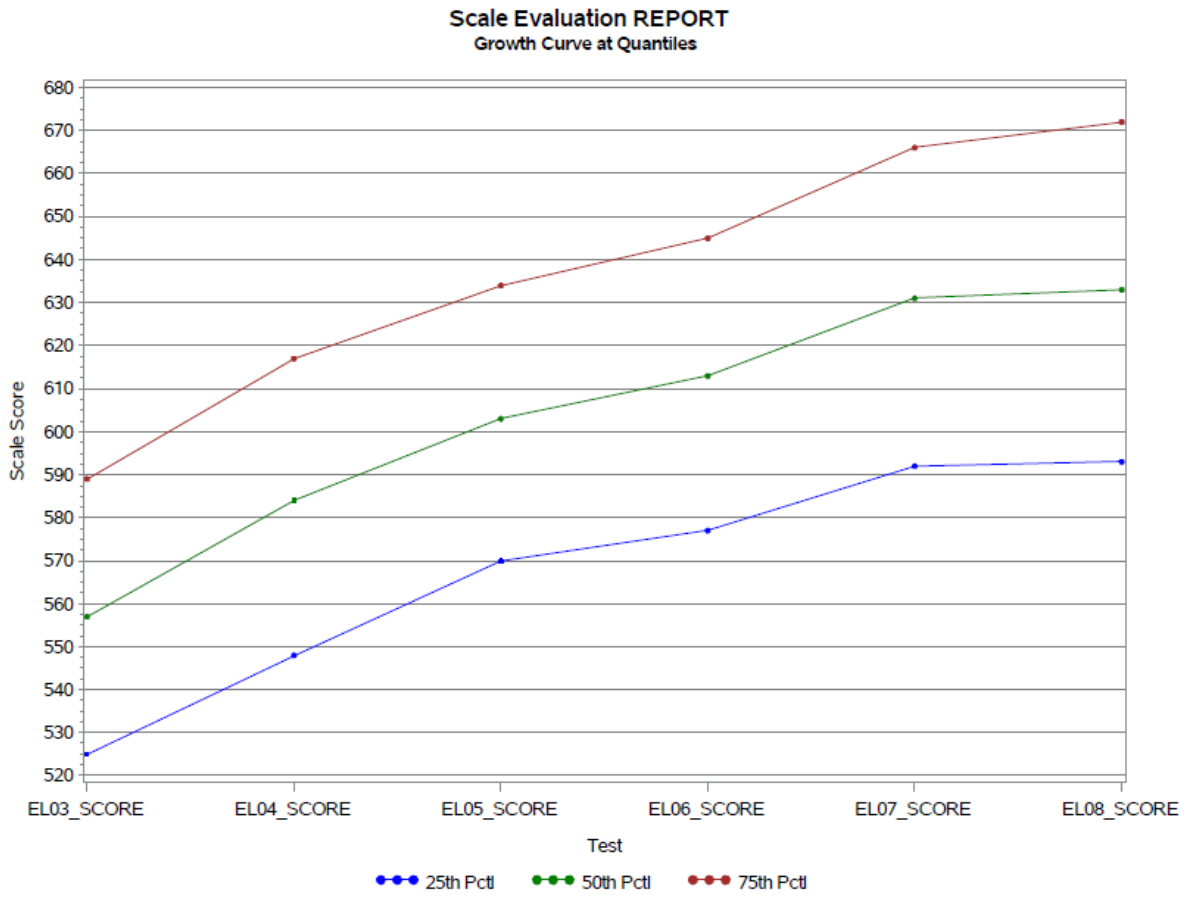


Figure 6-21 Mathematics Test Characteristic Curves

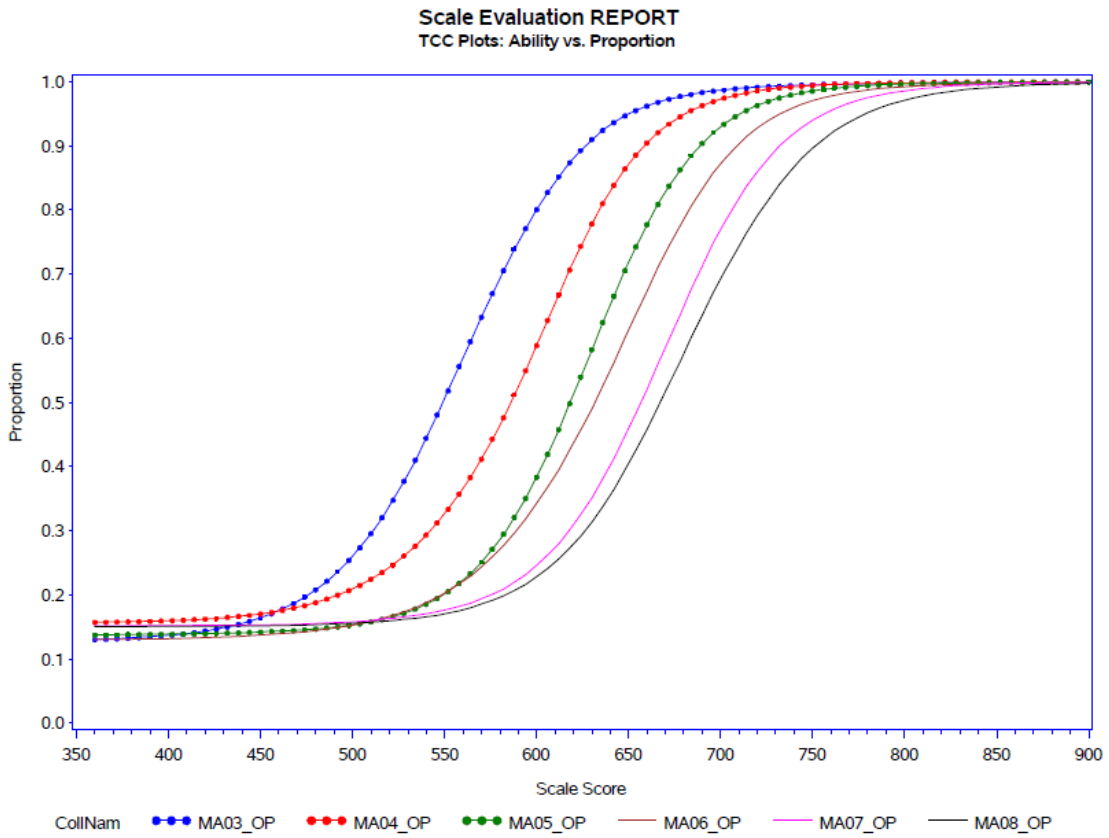


Figure 6-22 Mathematics Standard Error Curves

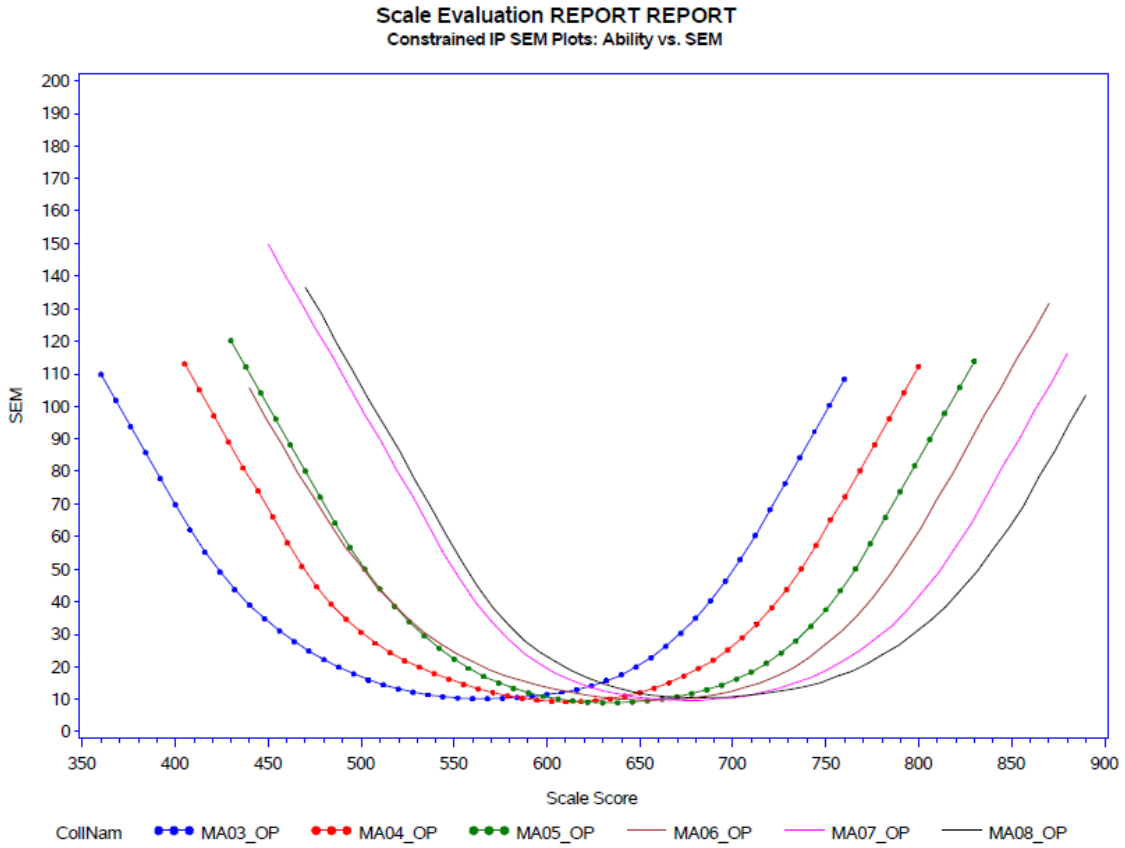


Figure 6-23 Mathematics Growth at Quartiles

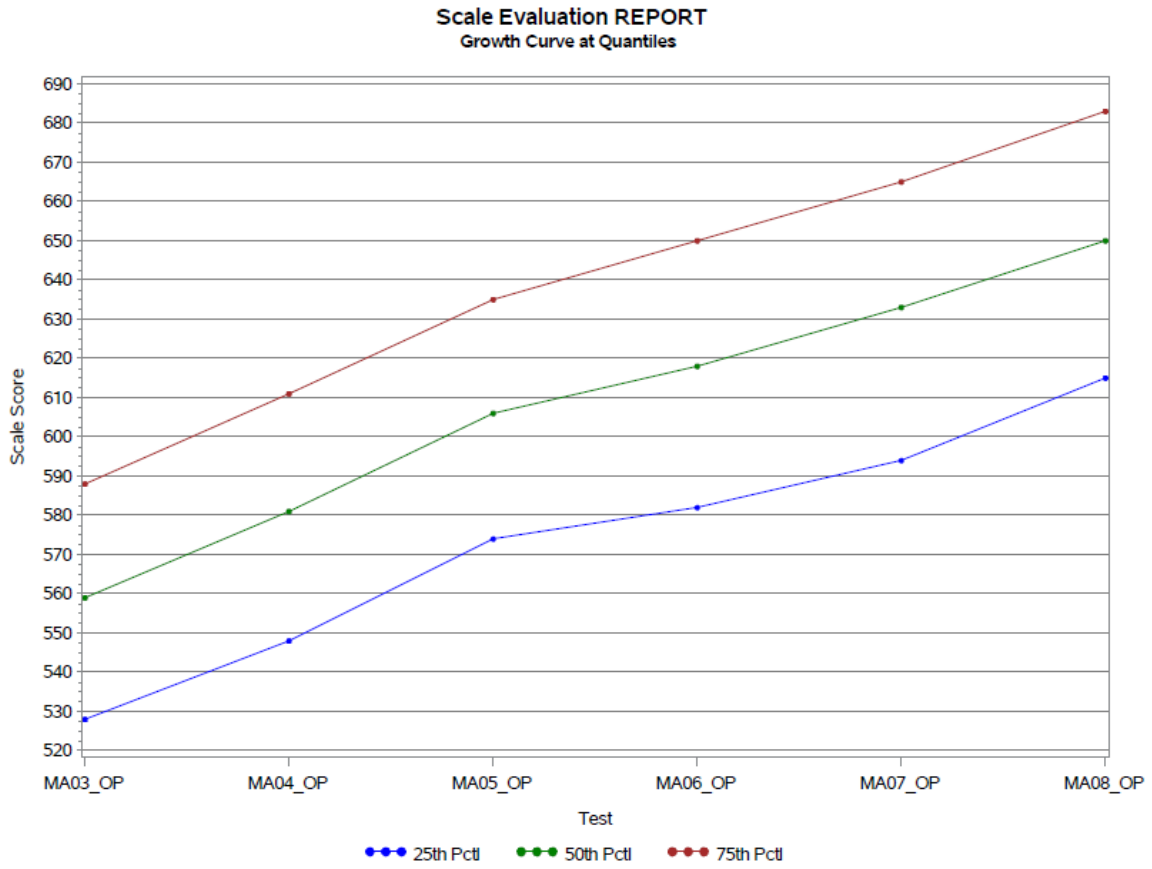


Figure 6-24 Science Test Characteristic Curves

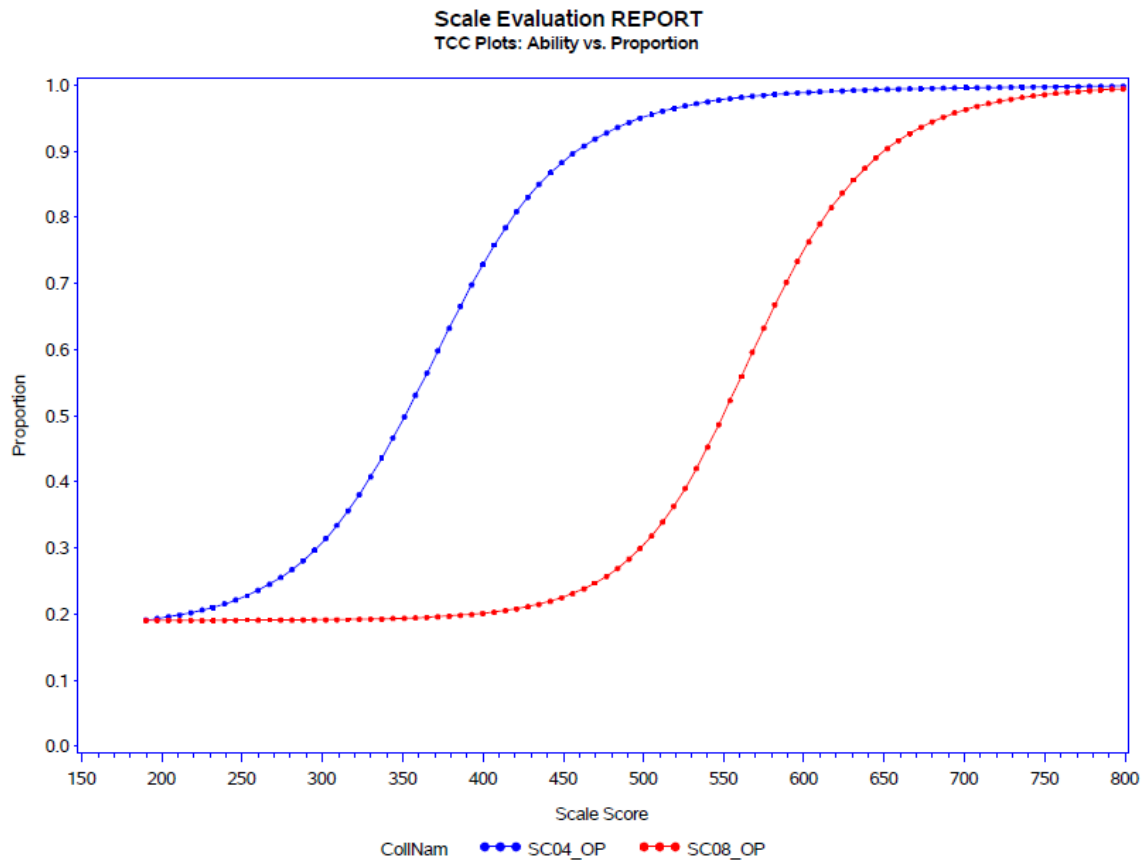


Figure 6-25 Science Standard Error Curves

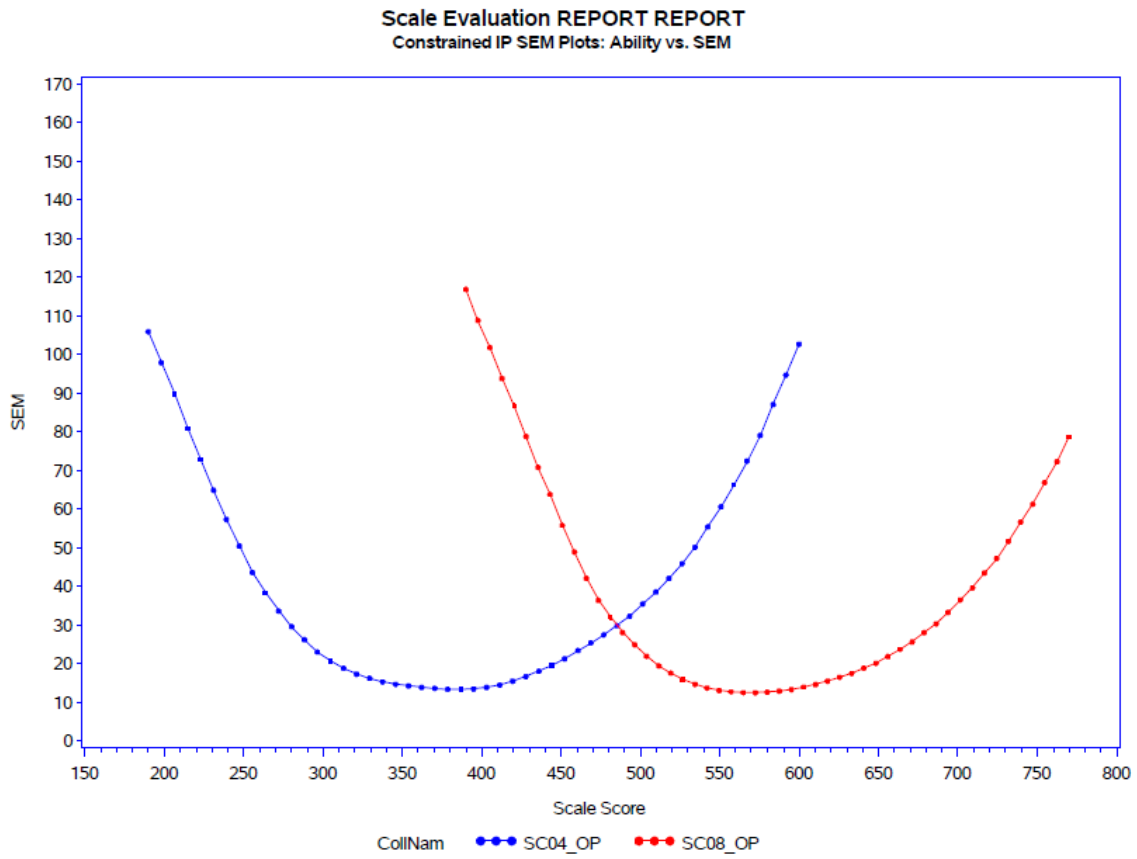


Figure 6-26 Science Growth at Quartiles

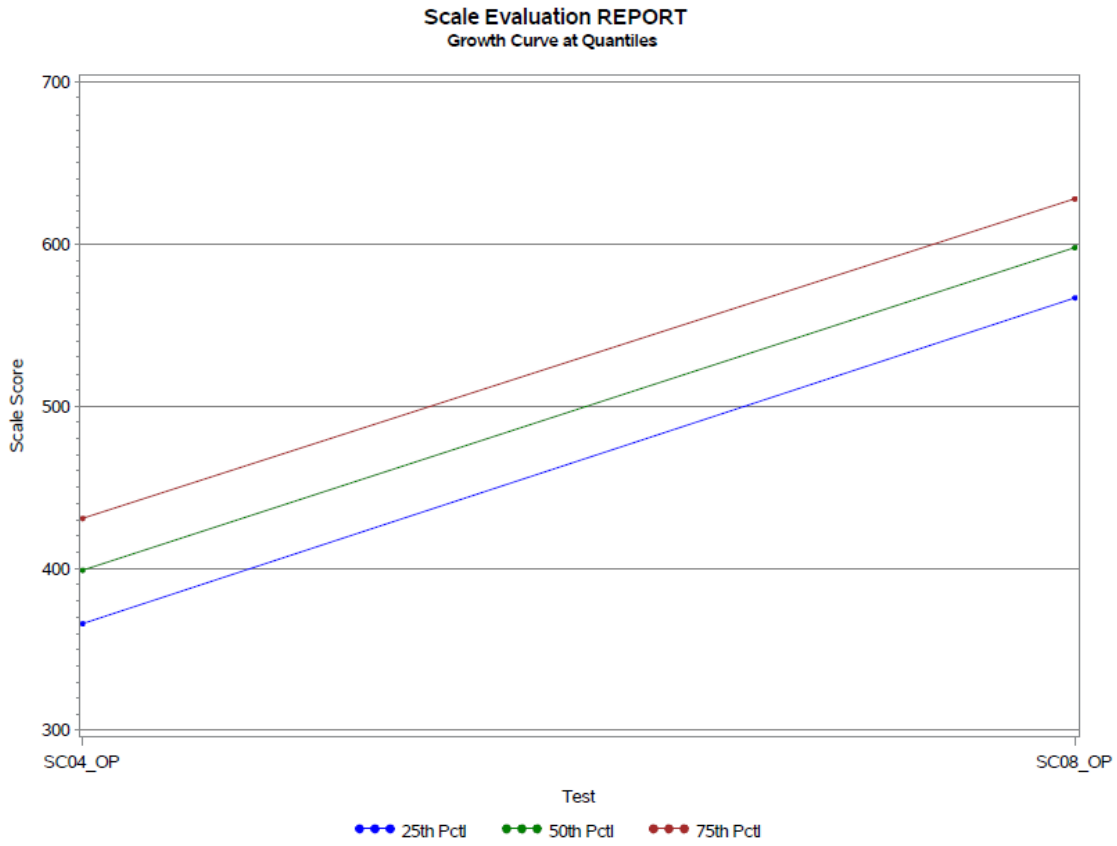


Figure 6-27 Social Studies Test Characteristic Curves

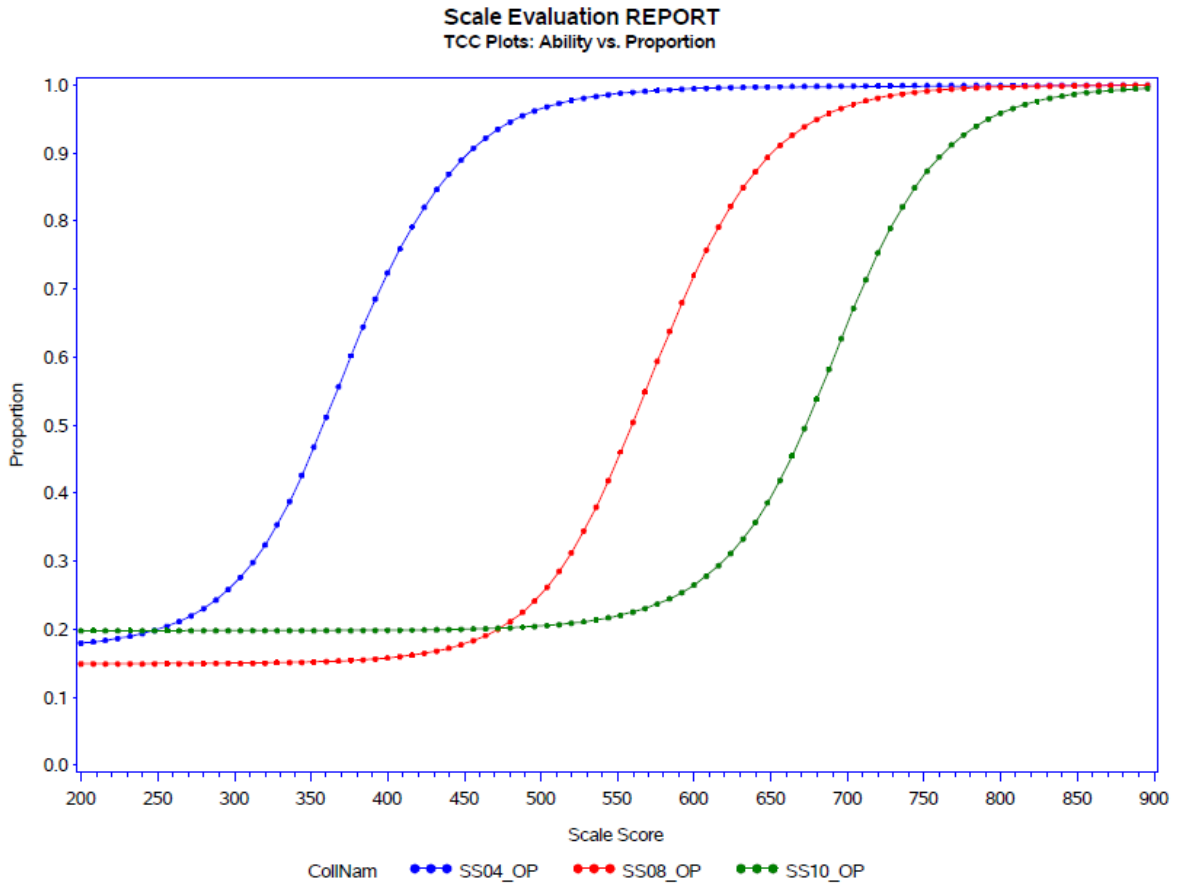


Figure 6-28 Social Studies Standard Error Curves

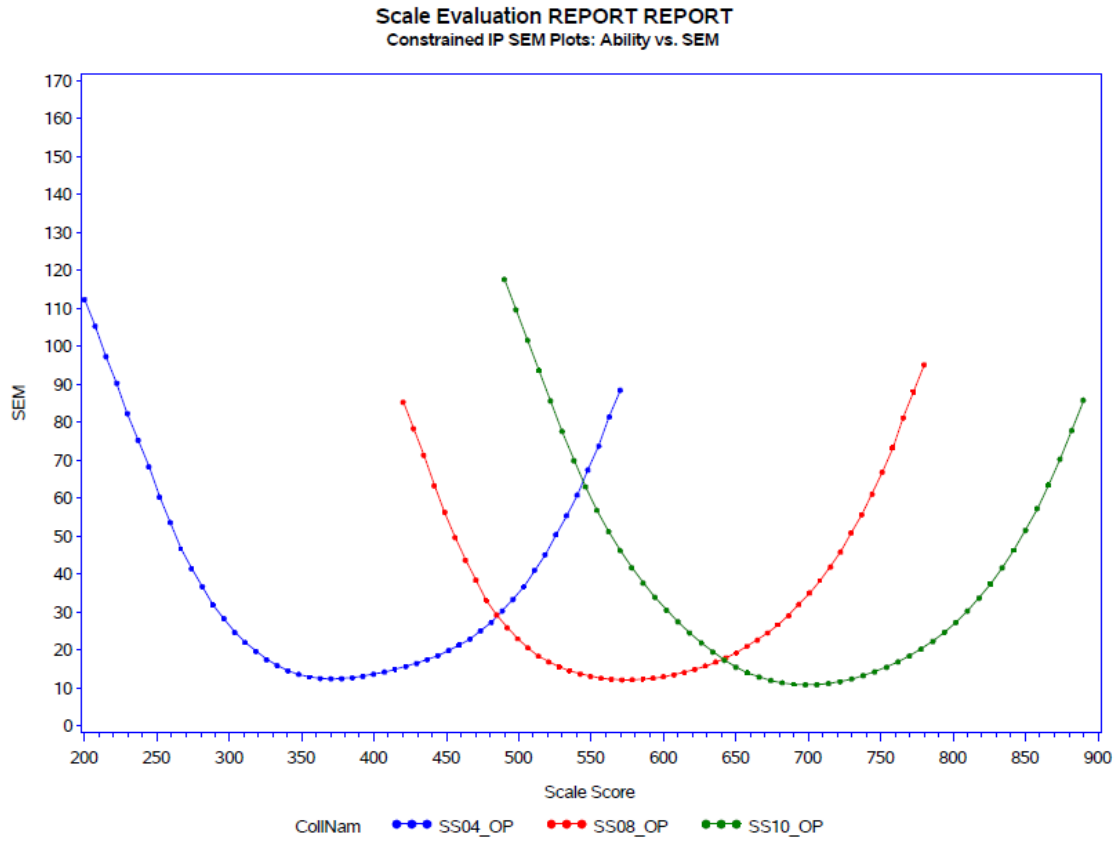
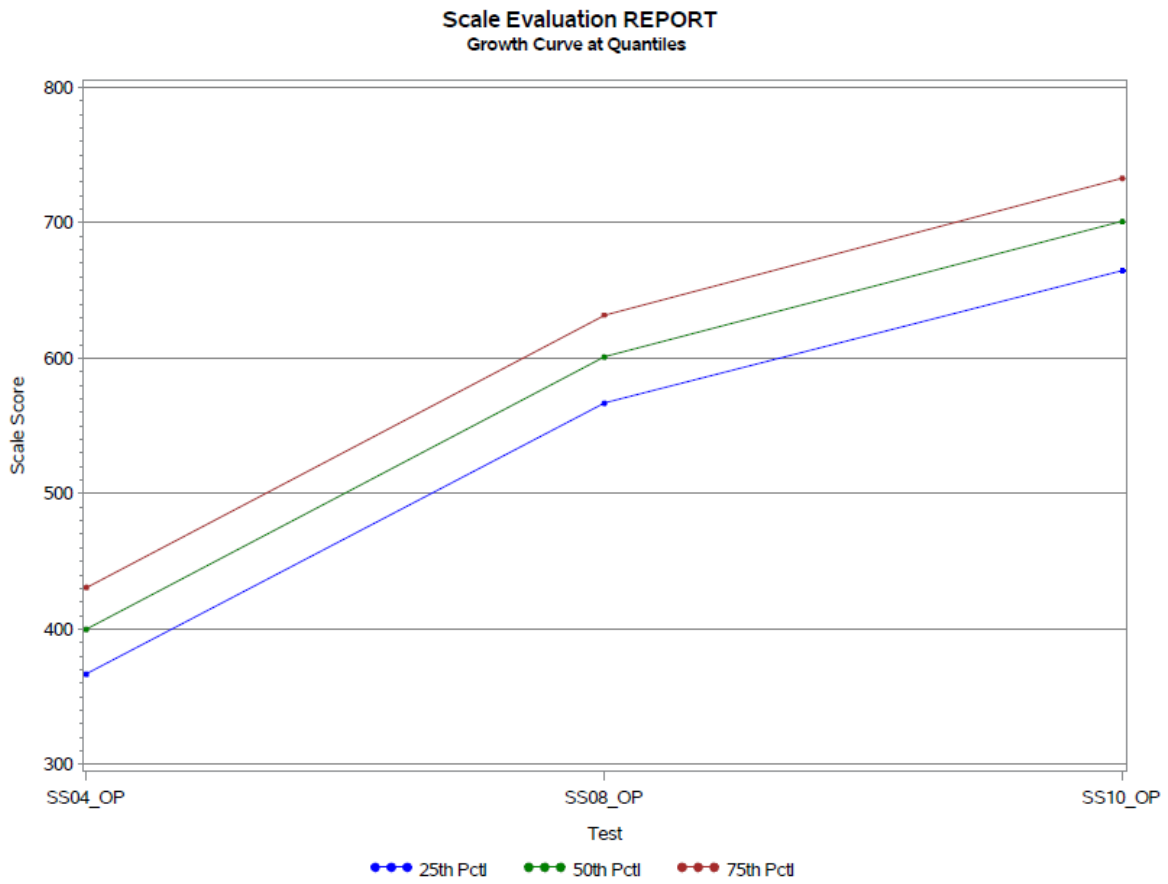


Figure 6-29 Social Studies Growth at Quartiles



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. (p. 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. (p. 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

the Spring 2016 test administration remained unchanged for the Spring 2017 and Spring 2018 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 minimal understanding 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 partial understanding 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 adequate understanding 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 thorough understanding 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 2018 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 American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (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. (p. 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. (p. 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. (p. 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. (p. 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. (p. 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. (p. 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. Therefore, 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.¹ 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 2018 Wisconsin Forward Exam, items were flagged for further investigation according to the following rules:

¹ 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 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.
- 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 (easy items) or very low p -values (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, 38 operational items were flagged on the Spring 2018 Wisconsin Forward Exam operational tests as meeting the investigational criteria bulleted above.

Table 8-A shows the number of scored items in the Spring 2018 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

One item was flagged for item-total test correlation <0.15 for ELA and for Mathematics. Two items were flagged for item-total correlations <0.15 for Science. All of the flagged items had item-total test correlations of at least 0.12.

8.1.3 Flagging for p -Value

Fourteen items were flagged for p -values <0.20 in Mathematics assessments, and all flagged items had p -values between 0.11 and 0.19. 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 of higher than 5%. Most of the items had an omit rate of 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 unspeeed 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 approximately 99% or more 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.

Tables 8-1 through 8-17 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, EBSR).

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 and test reliability statistics were computed on the Spring 2018 Wisconsin Forward Exam data for students with available item responses. These statistics 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 score points corresponding to each test.

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 53, 56, or 57. In Mathematics, the maximum possible raw score is 42 or 46.

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 23.46 and the maximum possible score is 42, resulting in the test mean p -value of approximately 0.56. 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. (The p -value reflects the overall test difficulty to which each test item contributes only once, while the mean raw score reflects the student performance on the test and was computed based on the student test scores that included weighted TDA item scores.)

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 eleven out of seventeen content areas/grades are zero, meaning that at least one student failed all items for each of those tests. The table also shows that the maximum obtained scores are equal to the maximum number of points possible on the test in all grades, 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 0.60.
- Standard deviations ranged from 9.86 to 11.20 raw score points.
- Alpha was relatively high in every grade (0.88 to 0.90).
- SEM ranged from 3.28 to 3.50.

Mathematics

- Test difficulty ranged from 0.41 to 0.56, with generally lower difficulty in lower grades and higher difficulty in higher grades.
- Standard deviations ranged from 9.39 to 10.10 raw score points.
- Alpha was relatively high in every grade (0.91 to 0.92).
- SEM ranged from 2.69 to 2.92.

Science

- Test difficulty was 0.69 in grade 4 and 0.70 in grade 8.
- Standard deviations were 7.28 and 7.44 raw score points for grades 4 and 8, respectively.
- Alpha was 0.88 for both grades.
- SEM was 2.54 and 2.52 for grades 4 and 8, respectively.

Social Studies

- Test difficulty was 0.68 for grades 4 and 8, and 0.63 for grade 10.
- Standard deviations ranged from 7.59 to 10.72 raw score points.
- Alpha ranged from 0.89 to 0.92.
- SEM ranged from 2.50 to 3.03.

8.2.1 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, use of testing accommodations, and English language proficiency. These subgroup comparisons draw from Tables 8-19 through 8-27 and show some consistent performance patterns by subgroups.

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 0.02 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 5 through 8 (differences of 0.1 or 0.0 in test p -value). At grades 3 and 4, the test was slightly easier for male students than for female students, with a difference of 0.02. In Science, the test difficulties were very similar (at 0.01) between male and female students in grades 4 and 8. In Social Studies, the differences in test difficulty between genders were very small (at 0.01 and 0.02) for all grades, with female students performing slightly better than male students in all grades.

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, Hispanic students, American Indian 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 between 0.00 and 0.02 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 0.13 to 0.17 across all grades and content areas, with the largest differences observed in Mathematics.

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 Science grade 4 to 0.24 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 grades 3 and 4, and Mathematics grade 3 to 0.22 in Social Studies grades 8 and 10. Larger differences in student performance were observed for higher grade levels compared to lower grade levels.

When looking at the test difficulty for students using testing accommodations, it should be noted that only 100 or fewer students per grade used ELA testing accommodations and fewer than 40 students per grade used Science or Social Studies testing accommodations. While it was observed that the tests were more difficult for students using testing accommodations, given small numbers of students using testing accommodations in ELA, Science, and Social Studies, the comparisons of the test difficulty for these students with the test difficulty for students not using testing accommodations for the corresponding grades and content areas should be made with caution. The number of students using testing accommodations in Mathematics ranged from 908 in Grade 3 to 2,739 in Grade 6. In all grades, the test was much easier for students not using testing accommodations, with differences in test difficulty ranging from 0.21 to 0.24.

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 8-28 is the summary scale score table based on 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 556.70 for grade 3 to 630.98 for grade 8. This mean scale score pattern supports the ELA vertical scale properties.
- Standard deviations ranged from 46.66 to 59.94 scale score points.
- In each grade level, student scores spanned the full-scale score range from the LOSS to the HOSS.

Mathematics

- Mean scale score increased by grade level, ranging from 555.94 for grade 3 to 644.24 for grade 8. This mean scale score pattern supports the Mathematics vertical scale properties.
- Standard deviations ranged from 50.87 to 65.55 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 399.03 and 595.66 for grades 4 and 8, respectively.
- Standard deviations were 53.13 and 52.26 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 398.23, 599.17, and 695.70 for grades 4, 8, and 10, respectively.
- Standard deviations ranged from 53.25 to 58.24 scale score points.
- In each grade level, student scores spanned the full scale score range from the LOSS to the HOSS.

8.3.1 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, English language proficiency, and accommodation use are presented in Tables 8-29 through 8-37. The scale score statistics were computed based on the census data.

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 8.00 scale score points in grade 3 to 18.24 scale score points in grade 8.
- In Mathematics, male students as a group showed slightly higher mean scale scores in grades 3 and 4 (difference of 2.7 and 3.5 points, respectively) and lower mean scale scores in grades 6 and 8 (difference of 3.9 and 5 scale score points) than female students. The differences between genders were less than half a scale score point in grades 5 and 7.
- In Science, the mean scale scores between genders were very similar for grade 4, with a difference of less than half a scale score point. A difference of 3.23 scale score points was observed in grade 8, with female students performing better than male students.
- There were small differences between mean scale scores by gender in Social Studies, from 2.23 scale score points in grade 4 to 4.36 scale score points in grade 8. Female students performed better than male students in all grades.

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, American Indian students, and African-American students.
- The mean scale scores of African-American students were typically more than one standard deviation lower than the mean scale scores of White students. The mean scale scores of Hispanic and American Indian students were approximately two-thirds of a standard deviation lower than the mean scale scores of White students for most grades and content areas. The mean scale scores of Asian students were typically less than a quarter of a standard deviation lower than the mean scale scores of White 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 33.80 scale score points in ELA grade 3 to 47.49 scale score points in Mathematics grade 7.
- For every grade and content area, the mean scale scores of students who were economically disadvantaged were typically more than two-thirds of a 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 36.19 scale score points in ELA grade 3 to 67.57 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 of a standard deviation to just 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 26.03 scale score points in Mathematics grade 3 to 64.11 scale score points in Mathematics grade 7.
- 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 of a standard deviation to just over one standard deviation.

Accommodation Use

- Students using testing accommodation (listed in section 4.1.3 of this report) performed less well on the tests compared to their peers not using testing accommodations. The differences ranged from 19 scale score points for Social Studies grade 10 to 78.47 for Mathematics grade 7.
- For every grade and content area, the mean scale scores of students using testing accommodations were lower than the mean scale scores of fully English proficient students by about half of a standard deviation to just over one standard deviation.
- The differences in mean scale scores between these two groups should be interpreted with caution for ELA, Science, and Social Studies due to a very low number of students using testing accommodations in these content areas.

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.” (p. 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 8-38 shows the cut scores for each content and grade level. For ease of reference, Tables 8-39 through 8-42 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 37% (grade 8) and 45% (grade 7) of students were either *Proficient* or *Advanced* in ELA.
- Between 7% and 10% 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 55% below *Proficient* in grade 7 to 63% below *Proficient* in grade 8.

Mathematics

- Between approximately 37% (grade 8) and 50% (grade 3) of students were either *Proficient* or *Advanced* in Mathematics.
- The proportion of students who were *Advanced* was between approximately 5% and 12%, depending on the grade level.
- Across all grade levels, 50% or more students were below *Proficient*. These percentages ranged from approximately 50% below *Proficient* in grade 3 to 63% below *Proficient* in grade 8.

Science

- Approximately 51% of students were either *Proficient* or *Advanced* in grade 4 and about 50% 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 15% in grade 8.
- The proportion of students classified as below *Proficient* was approximately 49% in grade 4 and 50% in grade 8.

Social Studies

- More than half of students in grades 4 and 8 were either *Proficient* or *Advanced* in Social Studies. The percentage of *Proficient* or *Advanced* students was approximately 54% in grade 4 and 52% in grade 8. Approximately 48% students in grade 10 were classified as either *Proficient* or *Advanced*.
- Between 20% and 22% of students were *Advanced* in all three grades.
- The percentage of students classified as below *Proficient* was approximately 46% in grade 4, 48% 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 8-43 through 8-46.

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 ranging from approximately 7% to 12%) 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, approximately 30% or more 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 compared to students with disabilities,

with the differences ranging from approximately 25% to 40%, depending on grade level and content area. 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 higher percentages of students who were fully English proficient classified as *Proficient* or *Advanced* compared to students with limited English proficiency, with the differences ranging from approximately 25% to 40%, depending on grade level and content area. There were much lower percentages of fully English proficient students who were classified in the lowest performance level in all grades and content areas.

Performance level results showed that there were higher percentages of students not using testing accommodations who were classified as *Proficient* or above compared to students using testing accommodations. The differences ranged from approximately 10% to 28% for ELA, Science, and Social Studies, depending on the grade level. For Mathematics, these differences were close to just above 40% in all grade levels. The differences in the percentages of students in different performance levels between groups of students using and not using testing accommodations should be interpreted with caution for ELA, Science, and Social Studies due to a very low number of students using testing accommodations in these 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, 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 because 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 the percentage 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, or 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 8-47 through 8-50 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 8-47 through 8-50 are summarized below. Tables 8-51 through 8-54 identify the SPI cut scores for each content area reporting category and grade level.

English Language Arts

Tables 8-47a and 8-47b 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 43.88 to 70.29 for content standards and from 51.78 to 65.81 for domains, indicating that the items were moderately difficult to easy for examinees. In general, content standard D (Writing/Language—Text Types and Purposes) was the most difficult in all grades. These content standards contained the TDA item, which was generally difficult for students. Content standard C (Reading—Vocabulary Use) was one of the easiest in grades 5 through 8. The Listening domain was easier than other domains for students in grades 3, 5, 7, and 8. The Reading domain was the easiest in grades 4 and 6. The Writing/Language domain was the most difficult domain for students in all grades.

Mathematics

Table 8-48 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 36.54 (Ratios and Proportional Relationships in grade 6) to 62.07 (Geometry in grade 3). 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 grades 4 and 5. Content standard F (Ratios and Proportional Relationships) was the most difficult in grade 6. Content standard E (Geometry) was the most difficult in grade 7, and content standard G (The Number System) was the most difficult in grade 8.

Science

Table 8-49 presents Science *p*-values and SPI scores across grades and content standards. The mean Science SPI scores across both grades and all content standards ranged from 57.76 to 76.78, 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 8-50 presents Social Studies *p*-values and SPI scores across grades and content standards. The mean Social Studies SPI scores across all grades and content standards ranged from 57.30 to 79.31, 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, the most difficult standard was content standard D (Economics); in grade 8, the most difficult standard was content standard C (Political Science and Citizenship); and in grade 10, the most difficult standard was content standard E (Behavioral Sciences).

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 of >75 were the following:

- Grade 4 Science content standards GH (Science Application & Personal and Social Perspective)
- Grade 8 Science content standards AB (Science Connections & Nature of Science) and D (Physical Science)
- Grade 4 Social Studies content standard E (Behavioral Sciences)

There were no SPI mean scores of <25 in the Wisconsin Spring 2018 test administration.

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, three years of the test scores on the same reporting scales are available, and the state-level mean scale scores and standard deviations for the 2016, 2017, and 2018 administrations are presented for ELA, Mathematics, Science, and Social Studies in Tables 8-55 through 8-58. The statistics presented in these tables are based on the total population of Wisconsin students, including students attending public, choice, and private schools.

It was observed that the mean scale score for ELA decreased for all grade levels except for grade 7 between the last two test administrations. The score decrease for grades ranged from approximately 2 scale score points for grades 3 and 5 to over 6 scale score points for grade 8. The scale score increase for grade 7 was small and less than 1 scale score point.

The mean scale score for Mathematics increased between 1 and 3 scale score points for grades 3, 4 and 8 between the last two test administrations. The mean scale score decreased between approximately 1 point for grades 5 and 6, and by approximately 5 points for grade 7 between Spring 2017 and 2018 administrations.

For Science, the mean scale score remained practically unchanged for grade 4 and increased by approximately 1 and a half scale score points for grade 8 between the 2017 and 2018 test administrations.

For Social Studies, the mean scale score increased just over 1 scale score point between the 2017 and 2018 test administrations for grades 4 and 8. A decrease of approximately 1 point in the scale score mean was observed for grade 10.

Tables 8-59 through 8-62 show the percentages of students in each achievement level in the Spring 2016, 2017, and 2018 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, a decrease in the percentage of students at or above *Proficient* was observed for grades 3 through 6 and grade 8, ranging from approximately 2% for grade 3 to close to 4% for grade 8. A small increase in the percentage of students at or above *Proficient* was observed for grade 7 (approximately 1.5%).

For Mathematics, an increase in the percentage of students at or above *Proficient* was observed for grades 3, 4, 5, and 8. This increase ranged between 1% and 2%. There was no practical change in the percentage of students at or above *Proficient* for grades 6 and 7.

For Science, there was no practical change in the percentage of students at or above *Proficient* for grade 4. A small increase in the percentage of students at or above *Proficient* was observed for grade 8 (approximately 1%).

For Social Studies, approximately a 2% increase in the percentage of students at or above *Proficient* was observed for grades 4 and 8. There was no practical change in the percentage of students at or above *Proficient* for grade 10.

Overall, the percentages of students classified in any of the four performance level categories were found to be comparable between the Spring 2017 and 2018 test administrations across all grade levels and content areas. With the exception of ELA grade 4 (*Below Basic*), grade 6 (*Below Basic*), grade 7 (*Proficient*), and grade 8 (*Below Basic* and *Advanced*), the changes between the percentage of students in Spring 2017 and Spring 2018 in any performance level category, grade, or content area was 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. Therefore, 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 Wisconsin Forward Exam

Content	Grade	# of Items Flagged	Number of Flags			
			Correlation <0.15	Distractor Correlation >0	Omit >5%	<i>p</i> -Value <0.20
ELA	3	2	0	2	0	0
	4	2	0	2	0	0
	5	0	0	0	0	0
	6	3	1	3	0	0
	7	1	0	1	0	0
	8	0	0	0	0	0
MA	3	2	0	1	0	1
	4	3	0	1	0	2
	5	3	0	1	0	2
	6	7	1	2	0	4
	7	7	0	4	0	3
	8	3	0	1	0	2
SC	4	2	1	2	0	0
	8	2	1	2	0	0
SS	4	0	0	0	0	0
	8	0	0	0	0	0
	10	1	0	1	0	0
Total		38	4	23	0	14

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	EL	27	MC	0.55	0.22	0.71		+	0.05		
3	EL	37	MC	0.35	0.23	0.36		+	0.05		
4	EL	2	MC	0.39	0.22	0.08		+	0.01		
4	EL	10	MC	0.57	0.18	0.16		+	0.03		
6	EL	10	MC	0.37	0.12	0.19	+	+	0.16		
6	EL	14	MC	0.56	0.34	0.08		+	0.02		
6	EL	18	MC	0.25	0.18	0.21		+	0.05		
7	EL	26	MC	0.48	0.21	0.27		+	0.07		

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	8	MC	0.19	0.30	0.14					+
3	MA	30	MC	0.39	0.20	0.20		+	0.00		
4	MA	4	SA	0.19	0.48	0.18					+
4	MA	36	TE	0.19	0.49	0.45					+
4	MA	37	MC	0.52	0.31	0.18		+	0.02		
5	MA	3	MC	0.59	0.19	0.14		+	0.02		
5	MA	31	SA	0.11	0.42	0.43					+
5	MA	44	SA	0.12	0.44	0.26					+
6	MA	11	SA	0.19	0.51	0.47					+
6	MA	13	SA	0.14	0.46	0.23					+
6	MA	28	MC	0.33	0.34	0.44		+	0.03		
6	MA	30	TE	0.15	0.36	0.23					+
6	MA	31	SA	0.15	0.49	0.56					+
6	MA	33	MC	0.31	0.13	0.34	+				
6	MA	43	MC	0.33	0.23	0.31		+	0.01		
7	MA	1	MC	0.23	0.40	0.10		+	0.01		
7	MA	7	MC	0.35	0.27	0.08		+	0.04		
7	MA	24	TE	0.15	0.40	0.79					+
7	MA	27	MC	0.29	0.30	0.55		+	0.04		
7	MA	39	MC	0.31	0.24	0.65		+	0.02		
7	MA	41	SA	0.19	0.47	1.06					+
7	MA	44	TE	0.19	0.57	1.18					+
8	MA	29	TE	0.11	0.42	0.66					+
8	MA	34	MC	0.36	0.18	0.64		+	0.04		
8	MA	43	SA	0.15	0.37	1.17					+

Table 8-D Science and Social Studies 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	
4	SC	9	MC	0.35	0.22	0.53		+	0.03		
	SC	25	MC	0.46	0.13	0.09	+	+	0.09		
8	SC	32	MC	0.32	0.14	0.24	+	+	0.02		
8	SC	33	MC	0.36	0.22	0.23		+	0.03		
10	SS	46	MC	0.43	0.23	0.48		+	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	99.64	99.74	99.72	99.83	99.66	99.67	
Mathematics	98.48	99.84	99.85	99.70	99.46	99.48	
Science		99.81				99.73	
Social Studies		99.89				99.80	99.53

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	TDA	0.34	0.53	0.00				
2	MC	0.67	0.31	0.10				
3	MC	0.59	0.25	0.17				
4	TE	0.70	0.50	0.14				
5	MC	0.65	0.32	0.16				
6	MC	0.83	0.45	0.18				
7	MC	0.66	0.37	0.19				
8	TE	0.57	0.45	0.26				
9	MC	0.60	0.43	0.19				
10	MC	0.54	0.32	0.20				
11	MC	0.49	0.31	0.28				
12	MC	0.39	0.39	0.22				
13	MC	0.72	0.54	0.22				
14	TE	0.67	0.53	0.19				
15	MC	0.71	0.36	0.10				
16	TE	0.65	0.53	0.16				
17	MC	0.73	0.46	0.19				
18	EBSR	0.58	0.52	0.12				
19	MC	0.56	0.31	0.18				
20	MC	0.62	0.39	0.28				
21	TE	0.41	0.54	2.16				
22	MC	0.81	0.31	0.22				
23	MC	0.57	0.31	0.26				
24	EBSR	0.43	0.40	0.14				
25	MC	0.54	0.24	0.28				
26	MC	0.66	0.41	0.55				
27	MC	0.55	0.22	0.71		+		
28	TE	0.67	0.65	0.31				
29	MC	0.47	0.37	0.31				
30	MC	0.45	0.40	0.59				
31	MC	0.59	0.52	0.77				
32	MC	0.48	0.34	0.32				
33	EBSR	0.47	0.54	0.23				
34	TE	0.68	0.53	0.37				
35	MC	0.41	0.33	0.37				
36	MC	0.49	0.35	0.34				
37	MC	0.35	0.23	0.36		+		

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	TDA	0.35	0.52	0.00				
2	MC	0.39	0.22	0.08		+		
3	TE	0.64	0.43	0.14				
4	MC	0.75	0.36	0.12				
5	MC	0.68	0.35	0.16				
6	TE	0.47	0.45	0.23				
7	TE	0.63	0.46	0.21				
8	MC	0.58	0.46	0.18				
9	TE	0.53	0.48	0.16				
10	MC	0.57	0.18	0.16		+		
11	TE	0.70	0.41	0.47				
12	MC	0.70	0.46	0.16				
13	TE	0.80	0.35	0.13				
14	MC	0.58	0.28	0.06				
15	MC	0.76	0.34	0.14				
16	EBSR	0.33	0.34	0.05				
17	EBSR	0.62	0.44	0.06				
18	MC	0.74	0.39	0.13				
19	MC	0.38	0.29	0.15				
20	MC	0.47	0.43	0.16				
21	MC	0.78	0.49	0.18				
22	TE	0.58	0.58	0.24				
23	MC	0.60	0.43	0.21				
24	TE	0.54	0.49	0.29				
25	MC	0.64	0.51	0.26				
26	MC	0.59	0.41	0.30				
27	MC	0.77	0.42	0.36				
28	MC	0.71	0.51	0.26				
29	TE	0.26	0.49	0.60				
30	TE	0.68	0.50	0.21				
31	MC	0.53	0.40	0.43				
32	MC	0.50	0.37	0.87				
33	MC	0.54	0.39	0.33				
34	MC	0.42	0.27	0.28				
35	MC	0.59	0.49	0.31				
36	MC	0.57	0.49	0.31				
37	TE	0.46	0.41	0.28				
38	MC	0.55	0.42	0.28				
39	MC	0.60	0.43	0.26				

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	TDA	0.32	0.57	0.00				
2	MC	0.66	0.23	0.08				
3	MC	0.70	0.28	0.10				
4	TE	0.50	0.30	0.19				
5	MC	0.79	0.45	0.12				
6	TE	0.58	0.44	0.15				
7	TE	0.52	0.51	0.30				
8	MC	0.63	0.43	0.13				
9	MC	0.66	0.40	0.12				
10	MC	0.51	0.31	0.17				
11	MC	0.76	0.45	0.18				
12	MC	0.58	0.31	0.15				
13	TE	0.71	0.40	0.12				
14	MC	0.85	0.41	0.12				
15	MC	0.74	0.44	0.06				
16	MC	0.71	0.50	0.11				
17	TE	0.62	0.54	0.11				
18	TE	0.61	0.45	0.10				
19	MC	0.63	0.39	0.12				
20	MC	0.77	0.46	0.13				
21	MC	0.50	0.28	0.15				
22	TE	0.66	0.54	0.13				
23	MC	0.69	0.40	0.21				
24	MC	0.61	0.42	0.25				
25	TE	0.72	0.17	0.44				
26	MC	0.58	0.39	0.26				
27	MC	0.64	0.48	0.27				
28	MC	0.50	0.48	0.48				
29	MC	0.67	0.42	0.27				
30	MC	0.58	0.46	0.26				
31	MC	0.58	0.40	0.26				
32	MC	0.51	0.31	0.23				
33	MC	0.73	0.49	0.47				
34	TE	0.42	0.39	0.81				
35	MC	0.50	0.33	0.36				
36	MC	0.40	0.34	0.35				
37	MC	0.50	0.50	0.31				
38	MC	0.52	0.31	0.34				
39	MC	0.50	0.45	0.32				
40	MC	0.54	0.33	0.32				
41	MC	0.39	0.22	0.28				

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	TDA	0.38	0.53	0.00				
2	MC	0.70	0.38	0.06				
3	MC	0.68	0.24	0.12				
4	MC	0.73	0.46	0.14				
5	MC	0.57	0.30	0.15				
6	TE	0.62	0.27	0.14				
7	TE	0.75	0.48	0.16				
8	MC	0.46	0.32	0.13				
9	TE	0.51	0.28	0.17				
10	MC	0.37	0.12	0.19	+	+		
11	TE	0.61	0.42	0.19				
12	EBSR	0.43	0.36	0.15				
13	MC	0.52	0.35	0.16				
14	MC	0.56	0.34	0.08		+		
15	MC	0.57	0.33	0.17				
16	TE	0.61	0.39	0.17				
17	EBSR	0.71	0.53	0.10				
18	MC	0.25	0.18	0.21		+		
19	MC	0.77	0.48	0.20				
20	MC	0.45	0.21	0.19				
21	TE	0.66	0.46	0.19				
22	MC	0.51	0.34	0.25				
23	TE	0.59	0.44	0.30				
24	MC	0.55	0.33	0.29				
25	MC	0.81	0.51	0.30				
26	TE	0.57	0.43	0.25				
27	MC	0.62	0.45	0.36				
28	TE	0.63	0.39	0.42				
29	MC	0.37	0.31	0.23				
30	MC	0.74	0.46	0.24				
31	MC	0.74	0.43	0.43				
32	TE	0.54	0.46	0.39				
33	MC	0.68	0.43	0.31				
34	TE	0.66	0.51	0.31				
35	TE	0.58	0.39	0.33				
36	MC	0.51	0.34	0.29				
37	EBSR	0.58	0.50	0.17				

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	TDA	0.42	0.56	0.00				
2	TE	0.64	0.27	0.15				
3	TE	0.52	0.38	0.11				
4	MC	0.72	0.25	0.11				
5	MC	0.75	0.39	0.12				
6	TE	0.66	0.32	0.23				
7	MC	0.79	0.46	0.22				
8	MC	0.78	0.37	0.14				
9	MC	0.51	0.39	0.14				
10	MC	0.64	0.38	0.13				
11	MC	0.61	0.29	0.16				
12	EBSR	0.51	0.48	0.13				
13	MC	0.47	0.30	0.18				
14	MC	0.61	0.44	0.18				
15	TE	0.77	0.45	0.04				
16	MC	0.71	0.47	0.17				
17	MC	0.59	0.40	0.19				
18	MC	0.50	0.43	0.13				
19	MC	0.57	0.33	0.16				
20	EBSR	0.38	0.43	0.09				
21	TE	0.74	0.35	0.39				
22	MC	0.60	0.52	0.25				
23	MC	0.76	0.41	0.22				
24	EBSR	0.50	0.38	0.11				
25	MC	0.50	0.35	0.28				
26	MC	0.48	0.21	0.27		+		
27	MC	0.58	0.48	0.49				
28	TE	0.64	0.44	3.21				
29	TE	0.66	0.56	0.42				
30	MC	0.42	0.28	0.33				
31	TE	0.52	0.33	0.33				
32	EBSR	0.55	0.55	0.33				
33	MC	0.57	0.51	0.39				
34	MC	0.66	0.46	0.37				
35	MC	0.51	0.48	0.37				
36	MC	0.59	0.49	0.37				
37	TE	0.71	0.42	0.50				
38	MC	0.48	0.42	0.34				
39	MC	0.60	0.43	0.34				

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	TDA	0.47	0.61	0.00				
2	MC	0.64	0.34	0.04				
3	TE	0.61	0.32	0.27				
4	MC	0.68	0.47	0.11				
5	MC	0.22	0.18	0.18				
6	TE	0.52	0.48	0.63				
7	MC	0.73	0.47	0.20				
8	MC	0.68	0.38	0.16				
9	MC	0.81	0.35	0.16				
10	MC	0.55	0.40	0.17				
11	TE	0.49	0.35	0.53				
12	MC	0.54	0.29	0.25				
13	MC	0.69	0.44	0.23				
14	MC	0.57	0.43	0.20				
15	MC	0.53	0.39	0.20				
16	MC	0.60	0.25	0.05				
17	MC	0.87	0.41	0.13				
18	EBSR	0.55	0.52	0.04				
19	EBSR	0.55	0.47	0.05				
20	MC	0.70	0.49	0.14				
21	MC	0.47	0.36	0.12				
22	TE	0.47	0.38	0.66				
23	MC	0.56	0.54	0.33				
24	EBSR	0.46	0.41	0.11				
25	MC	0.63	0.45	0.31				
26	MC	0.62	0.25	0.26				
27	MC	0.60	0.49	0.24				
28	MC	0.50	0.27	0.28				
29	TE	0.69	0.57	0.43				
30	MC	0.76	0.54	0.45				
31	MC	0.47	0.24	0.41				
32	EBSR	0.61	0.53	0.16				
33	MC	0.54	0.39	0.52				
34	EBSR	0.58	0.63	0.22				
35	MC	0.78	0.53	0.36				
36	MC	0.67	0.44	0.33				
37	MC	0.61	0.35	0.35				
38	TE	0.57	0.42	0.34				
39	MC	0.75	0.49	0.36				
40	MC	0.59	0.43	0.33				

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	0.64	0.46	0.11				
2	MC	0.74	0.44	0.08				
3	MC	0.75	0.48	0.12				
4	MC	0.76	0.45	0.12				
5	MC	0.53	0.46	0.12				
6	TE	0.43	0.61	0.35				
7	SA	0.84	0.39	0.58				
8	MC	0.19	0.30	0.14				+
9	MC	0.67	0.40	0.18				
10	SA	0.61	0.56	0.12				
11	MC	0.44	0.36	0.18				
12	SA	0.21	0.48	0.15				
13	MC	0.62	0.42	0.51				
14	MC	0.47	0.32	0.60				
15	MC	0.43	0.32	0.25				
16	SA	0.33	0.51	0.24				
17	MC	0.40	0.44	0.16				
18	MC	0.59	0.42	0.18				
19	MC	0.72	0.49	0.17				
20	TE	0.87	0.36	1.09				
21	MC	0.82	0.45	0.14				
22	MC	0.47	0.39	0.10				
23	MC	0.44	0.39	0.13				
24	MC	0.62	0.45	0.21				
25	MC	0.61	0.39	0.14				
26	SA	0.64	0.57	0.10				
27	TE	0.63	0.33	0.34				
28	MC	0.68	0.45	0.64				
29	MC	0.42	0.38	0.25				
30	MC	0.39	0.20	0.20		+		
31	MC	0.61	0.49	0.18				
32	MC	0.41	0.39	0.23				
33	SA	0.52	0.56	0.22				
34	MC	0.49	0.48	0.59				
35	MC	0.66	0.51	0.59				

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	TE	0.24	0.47	0.67				
37	MC	0.65	0.49	0.20				
38	SA	0.59	0.60	0.22				
39	MC	0.80	0.43	0.19				
40	MC	0.75	0.39	0.19				
41	MC	0.52	0.57	0.24				
42	TE	0.35	0.47	1.52				

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	0.68	0.45	0.32				
2	MC	0.46	0.30	0.08				
3	MC	0.57	0.49	0.07				
4	SA	0.19	0.48	0.18				+
5	MC	0.72	0.45	0.07				
6	TE	0.38	0.49	0.25				
7	MC	0.64	0.41	0.13				
8	MC	0.64	0.37	0.34				
9	MC	0.73	0.39	0.09				
10	MC	0.37	0.32	0.14				
11	MC	0.59	0.37	0.14				
12	TE	0.69	0.36	0.10				
13	MC	0.47	0.58	0.15				
14	MC	0.41	0.28	0.17				
15	MC	0.40	0.34	0.48				
16	SA	0.36	0.52	0.46				
17	MC	0.88	0.32	0.14				
18	MC	0.42	0.46	0.19				
19	MC	0.58	0.33	0.17				
20	TE	0.26	0.57	0.50				
21	MC	0.59	0.42	0.14				
22	MC	0.83	0.31	0.17				
23	MC	0.43	0.46	0.17				
24	MC	0.82	0.40	0.33				
25	MC	0.86	0.33	0.12				
26	MC	0.51	0.41	0.10				
27	MC	0.49	0.49	0.15				
28	MC	0.33	0.57	0.14				
29	TE	0.25	0.44	0.15				
30	SA	0.23	0.43	0.15				
31	MC	0.32	0.41	0.33				
32	MC	0.40	0.24	0.16				
33	MC	0.53	0.59	0.14				
34	MC	0.43	0.43	0.14				
35	MC	0.63	0.49	0.16				

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	TE	0.19	0.49	0.45				+
37	MC	0.52	0.31	0.18		+		
38	MC	0.58	0.41	0.36				
39	MC	0.38	0.51	0.37				
40	MC	0.49	0.49	0.16				
41	MC	0.32	0.53	0.20				
42	MC	0.49	0.38	0.19				
43	SA	0.50	0.44	0.22				
44	MC	0.28	0.40	0.17				
45	MC	0.56	0.40	0.16				
46	MC	0.55	0.39	0.16				

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	SA	0.71	0.46	0.30				
2	MC	0.47	0.52	0.06				
3	MC	0.59	0.19	0.14		+		
4	MC	0.35	0.40	0.09				
5	MC	0.42	0.35	0.14				
6	TE	0.34	0.61	0.25				
7	TE	0.33	0.55	0.62				
8	MC	0.43	0.35	0.35				
9	MC	0.67	0.43	0.10				
10	SA	0.61	0.23	0.17				
11	MC	0.76	0.47	0.15				
12	SA	0.34	0.51	0.27				
13	MC	0.55	0.56	0.14				
14	MC	0.56	0.45	0.15				
15	TE	0.23	0.56	0.94				
16	SA	0.32	0.35	0.35				
17	MC	0.47	0.34	0.16				
18	MC	0.46	0.50	0.24				
19	MC	0.44	0.46	0.20				
20	TE	0.59	0.46	0.22				
21	MC	0.39	0.26	0.22				
22	SA	0.32	0.60	0.28				
23	MC	0.69	0.44	0.24				
24	MC	0.56	0.33	0.28				
25	SA	0.48	0.47	0.14				
26	MC	0.40	0.37	0.10				
27	MC	0.57	0.44	0.13				
28	MC	0.37	0.17	0.20				
29	TE	0.44	0.50	0.32				
30	SA	0.56	0.53	0.25				
31	SA	0.11	0.42	0.43				+
32	MC	0.54	0.49	0.15				
33	SA	0.25	0.45	0.21				
34	SA	0.36	0.52	0.24				
35	TE	0.23	0.49	1.24				

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	0.52	0.49	0.17				
37	TE	0.34	0.55	0.13				
38	MC	0.50	0.43	0.50				
39	MC	0.54	0.29	0.32				
40	MC	0.58	0.45	0.18				
41	MC	0.29	0.41	0.19				
42	MC	0.44	0.42	0.16				
43	MC	0.54	0.47	0.20				
44	SA	0.12	0.44	0.26				+
45	MC	0.28	0.37	0.27				
46	MC	0.79	0.26	0.15				

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	0.08				
2	SA	0.52	0.49	0.29				
3	MC	0.63	0.40	0.15				
4	TE	0.59	0.52	0.41				
5	MC	0.54	0.34	0.11				
6	SA	0.25	0.52	0.45				
7	MC	0.74	0.45	0.15				
8	MC	0.52	0.33	0.21				
9	SA	0.51	0.55	0.24				
10	MC	0.80	0.41	0.14				
11	SA	0.19	0.51	0.47				+
12	MC	0.40	0.29	0.24				
13	SA	0.14	0.46	0.23				+
14	MC	0.50	0.35	0.23				
15	TE	0.38	0.59	0.57				
16	SA	0.20	0.51	0.43				
17	MC	0.77	0.39	0.20				
18	MC	0.68	0.38	0.26				
19	SA	0.24	0.59	0.64				
20	MC	0.36	0.51	0.16				
21	MC	0.71	0.28	0.21				
22	MC	0.60	0.53	0.24				
23	MC	0.51	0.43	0.27				
24	TE	0.73	0.34	1.03				
25	MC	0.42	0.32	0.25				
26	MC	0.52	0.26	0.48				
27	SA	0.31	0.44	0.67				
28	MC	0.33	0.34	0.44		+		
29	MC	0.58	0.32	0.52				
30	TE	0.15	0.36	0.23				+
31	SA	0.15	0.49	0.56				+
32	MC	0.64	0.46	0.31				
33	MC	0.31	0.13	0.34	+			
34	MC	0.44	0.46	0.27				
35	MC	0.61	0.48	0.26				

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	0.46	0.27				
37	TE	0.39	0.51	0.80				
38	MC	0.40	0.25	0.43				
39	MC	0.40	0.26	0.49				
40	MC	0.35	0.55	0.58				
41	SA	0.22	0.50	0.56				
42	MC	0.60	0.46	0.29				
43	MC	0.33	0.23	0.31		+		
44	TE	0.45	0.59	0.53				
45	MC	0.64	0.41	0.39				
46	MC	0.32	0.49	0.30				

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	0.40	0.10		+		
2	MC	0.47	0.54	0.08				
3	MC	0.52	0.33	0.08				
4	SA	0.24	0.54	0.24				
5	MC	0.50	0.43	0.12				
6	MC	0.40	0.44	0.11				
7	MC	0.35	0.27	0.08		+		
8	MC	0.49	0.37	0.08				
9	TE	0.20	0.45	0.13				
10	MC	0.47	0.38	0.11				
11	MC	0.50	0.43	0.09				
12	MC	0.49	0.41	0.15				
13	MC	0.62	0.48	0.26				
14	SA	0.25	0.56	0.87				
15	MC	0.72	0.26	0.29				
16	MC	0.25	0.49	0.39				
17	MC	0.36	0.48	0.26				
18	MC	0.42	0.22	0.25				
19	TE	0.38	0.61	0.76				
20	MC	0.37	0.28	0.27				
21	SA	0.37	0.47	0.52				
22	MC	0.30	0.26	0.34				
23	MC	0.61	0.27	0.45				
24	TE	0.15	0.40	0.79				+
25	MC	0.32	0.27	0.51				
26	SA	0.20	0.54	1.15				
27	MC	0.29	0.30	0.55		+		
28	MC	0.50	0.47	0.41				
29	MC	0.55	0.47	0.41				
30	MC	0.38	0.44	0.48				
31	MC	0.60	0.55	0.42				
32	MC	0.55	0.18	0.40				
33	SA	0.38	0.55	0.47				
34	MC	0.51	0.51	0.58				
35	MC	0.65	0.39	0.55				

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	0.57	0.33	0.54				
37	MC	0.40	0.45	0.68				
38	MC	0.55	0.25	0.62				
39	MC	0.31	0.24	0.65		+		
40	TE	0.30	0.64	0.71				
41	SA	0.19	0.47	1.06				+
42	MC	0.37	0.45	0.59				
43	MC	0.51	0.52	0.57				
44	TE	0.19	0.57	1.18				+
45	MC	0.65	0.36	0.57				
46	MC	0.43	0.31	0.54				

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	0.24	0.20	0.11				
2	SA	0.42	0.62	0.42				
3	MC	0.54	0.39	0.13				
4	MC	0.47	0.23	0.10				
5	SA	0.21	0.53	0.37				
6	MC	0.35	0.44	0.16				
7	MC	0.41	0.31	0.09				
8	MC	0.50	0.49	0.06				
9	MC	0.55	0.34	0.13				
10	MC	0.46	0.33	0.18				
11	TE	0.25	0.49	0.44				
12	MC	0.41	0.36	0.13				
13	SA	0.20	0.51	0.51				
14	MC	0.35	0.41	0.23				
15	MC	0.64	0.27	0.31				
16	TE	0.49	0.51	0.45				
17	MC	0.61	0.45	0.37				
18	MC	0.58	0.43	0.27				
19	MC	0.45	0.40	0.28				
20	MC	0.56	0.50	0.25				
21	MC	0.50	0.43	0.34				
22	MC	0.44	0.44	0.37				
23	MC	0.35	0.27	0.29				
24	TE	0.20	0.46	0.91				
25	MC	0.55	0.39	0.52				
26	MC	0.67	0.41	0.54				
27	SA	0.30	0.54	1.26				
28	MC	0.55	0.57	0.43				
29	TE	0.11	0.42	0.66				+
30	MC	0.30	0.39	0.52				
31	MC	0.49	0.27	0.44				
32	TE	0.34	0.51	0.85				
33	MC	0.57	0.40	0.44				
34	MC	0.36	0.18	0.64		+		
35	MC	0.51	0.31	0.54				

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	0.56	0.47	0.59				
37	TE	0.36	0.37	0.88				
38	MC	0.70	0.48	0.62				
39	MC	0.42	0.39	0.68				
40	MC	0.75	0.43	0.50				
41	MC	0.36	0.30	0.60				
42	MC	0.51	0.54	0.51				
43	SA	0.15	0.37	1.17				+
44	MC	0.56	0.33	0.53				
45	MC	0.49	0.37	0.50				
46	MC	0.56	0.47	0.52				

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	0.38	0.09				
2	MC	0.87	0.44	0.07				
3	MC	0.92	0.18	0.05				
4	MC	0.94	0.31	0.08				
5	MC	0.57	0.42	0.09				
6	MC	0.79	0.43	0.11				
7	MC	0.79	0.46	0.08				
8	MC	0.73	0.46	0.25				
9	MC	0.35	0.22	0.53		+		
10	MC	0.66	0.50	0.23				
11	TE	0.41	0.24	0.61				
12	MC	0.60	0.49	0.12				
13	MC	0.89	0.45	0.09				
14	MC	0.62	0.36	0.12				
15	MC	0.87	0.28	0.06				
16	MC	0.81	0.38	0.06				
17	MC	0.80	0.44	0.24				
18	MC	0.57	0.36	0.29				
19	MC	0.76	0.28	0.09				
20	MC	0.53	0.42	0.10				
21	MC	0.85	0.35	0.10				
22	MC	0.92	0.24	0.06				
23	MC	0.84	0.42	0.11				
24	MC	0.63	0.36	0.09				
25	MC	0.46	0.13	0.09	+	+		
26	MC	0.58	0.34	0.11				
27	MC	0.64	0.39	0.19				
28	MC	0.55	0.29	0.08				
29	MC	0.75	0.32	0.26				
30	MC	0.67	0.41	0.44				
31	MC	0.72	0.37	0.25				
32	MC	0.74	0.48	0.17				
33	MC	0.71	0.49	0.15				
34	MC	0.79	0.44	0.13				
35	MC	0.82	0.44	0.15				

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	0.40	0.20	0.12				
37	MC	0.67	0.34	0.12				
38	MC	0.46	0.33	0.23				
39	MC	0.56	0.45	0.38				
40	MC	0.77	0.47	0.19				

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	0.85	0.41	0.04				
2	MC	0.87	0.40	0.09				
3	MC	0.91	0.38	0.07				
4	MC	0.79	0.38	0.09				
5	MC	0.86	0.29	0.09				
6	MC	0.85	0.38	0.10				
7	MC	0.78	0.44	0.09				
8	MC	0.84	0.41	0.16				
9	MC	0.62	0.46	0.44				
10	MC	0.46	0.20	0.28				
11	MC	0.66	0.45	0.23				
12	MC	0.45	0.26	0.21				
13	MC	0.77	0.29	0.10				
14	MC	0.90	0.32	0.12				
15	MC	0.75	0.35	0.10				
16	MC	0.75	0.37	0.09				
17	MC	0.61	0.29	0.15				
18	MC	0.69	0.45	0.18				
19	MC	0.87	0.50	0.16				
20	MC	0.68	0.44	0.16				
21	MC	0.77	0.34	0.13				
22	MC	0.72	0.36	0.10				
23	MC	0.74	0.49	0.14				
24	MC	0.56	0.26	0.12				
25	MC	0.54	0.27	0.16				
26	MC	0.78	0.50	0.15				
27	MC	0.62	0.33	0.15				
28	MC	0.64	0.48	0.12				
29	MC	0.73	0.45	0.18				
30	MC	0.67	0.52	0.41				
31	MC	0.50	0.33	0.41				
32	MC	0.32	0.14	0.24	+	+		
33	MC	0.36	0.22	0.23		+		
34	MC	0.86	0.46	0.16				
35	MC	0.72	0.42	0.16				

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	0.51	0.16				
37	MC	0.84	0.51	0.15				
38	MC	0.69	0.38	0.18				
39	MC	0.71	0.46	0.31				
40	MC	0.74	0.45	0.27				

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	0.74	0.38	0.05				
2	MC	0.84	0.42	0.12				
3	MC	0.85	0.32	0.11				
4	MC	0.59	0.40	0.14				
5	MC	0.58	0.44	0.12				
6	TE	0.69	0.41	0.13				
7	MC	0.81	0.46	0.11				
8	MC	0.71	0.43	0.14				
9	MC	0.78	0.46	0.11				
10	TE	0.56	0.38	0.11				
11	MC	0.54	0.23	0.12				
12	MC	0.61	0.32	0.27				
13	MC	0.44	0.41	0.14				
14	MC	0.34	0.27	0.15				
15	MC	0.73	0.40	0.10				
16	MC	0.83	0.36	0.12				
17	MC	0.66	0.47	0.13				
18	MC	0.60	0.35	0.13				
19	MC	0.85	0.48	0.10				
20	MC	0.83	0.54	0.06				
21	MC	0.80	0.37	0.09				
22	MC	0.83	0.47	0.10				
23	MC	0.58	0.36	0.12				
24	MC	0.61	0.29	0.10				
25	MC	0.61	0.32	0.10				
26	TE	0.54	0.40	0.12				
27	MC	0.58	0.38	0.47				
28	TE	0.45	0.47	0.19				
29	MC	0.75	0.46	0.16				
30	MC	0.64	0.20	0.14				
31	MC	0.66	0.47	0.15				
32	MC	0.73	0.48	0.13				
33	MC	0.69	0.44	0.31				
34	MC	0.83	0.46	0.43				
35	MC	0.70	0.45	0.20				
36	MC	0.82	0.46	0.14				
37	MC	0.85	0.47	0.14				
38	MC	0.78	0.51	0.11				

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	0.85	0.48	0.09				
2	MC	0.79	0.40	0.13				
3	MC	0.82	0.44	0.10				
4	MC	0.82	0.42	0.09				
5	MC	0.80	0.40	0.09				
6	MC	0.83	0.51	0.13				
7	MC	0.60	0.44	0.17				
8	MC	0.56	0.33	0.29				
9	MC	0.76	0.28	0.18				
10	MC	0.74	0.48	0.14				
11	MC	0.52	0.40	0.13				
12	TE	0.41	0.44	0.18				
13	MC	0.64	0.30	0.14				
14	MC	0.67	0.53	0.18				
15	MC	0.81	0.50	0.29				
16	MC	0.71	0.42	0.20				
17	TE	0.51	0.58	0.18				
18	MC	0.66	0.40	0.17				
19	MC	0.59	0.49	0.15				
20	MC	0.74	0.49	0.13				
21	MC	0.80	0.44	0.20				
22	TE	0.73	0.39	0.29				
23	MC	0.71	0.42	0.15				
24	MC	0.59	0.42	0.27				
25	MC	0.59	0.32	0.23				
26	MC	0.77	0.39	0.20				
27	TE	0.57	0.41	0.42				
28	MC	0.60	0.27	0.23				
29	MC	0.46	0.39	0.20				
30	MC	0.90	0.46	0.19				
31	MC	0.83	0.46	0.18				
32	MC	0.65	0.44	0.23				
33	MC	0.74	0.39	0.28				
34	MC	0.52	0.33	0.23				
35	MC	0.62	0.45	0.21				
36	MC	0.69	0.55	0.23				
37	MC	0.58	0.40	0.20				
38	MC	0.51	0.31	0.23				
39	MC	0.82	0.51	0.21				
40	MC	0.83	0.48	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	0.61	0.25	0.08				
2	MC	0.84	0.38	0.17				
3	MC	0.65	0.40	0.26				
4	MC	0.52	0.37	0.15				
5	MC	0.80	0.45	0.11				
6	MC	0.51	0.30	0.12				
7	TE	0.88	0.33	0.60				
8	MC	0.80	0.37	0.15				
9	TE	0.40	0.35	0.55				
10	MC	0.64	0.40	0.31				
11	MC	0.56	0.32	0.35				
12	MC	0.62	0.40	0.20				
13	MC	0.64	0.37	0.23				
14	MC	0.55	0.29	0.28				
15	MC	0.56	0.41	0.27				
16	MC	0.66	0.41	0.32				
17	MC	0.70	0.46	0.39				
18	MC	0.64	0.45	0.43				
19	MC	0.64	0.50	0.32				
20	MC	0.51	0.45	0.35				
21	MC	0.59	0.53	0.33				
22	MC	0.62	0.46	0.34				
23	MC	0.58	0.39	0.33				
24	MC	0.69	0.33	0.29				
25	MC	0.69	0.41	0.30				
26	MC	0.73	0.43	0.16				
27	MC	0.75	0.48	0.29				
28	MC	0.58	0.55	0.35				
29	MC	0.46	0.47	0.23				
30	TE	0.77	0.47	0.54				
31	MC	0.68	0.48	0.28				
32	MC	0.61	0.50	0.26				
33	MC	0.64	0.46	0.32				
34	MC	0.72	0.50	0.35				
35	MC	0.47	0.40	0.46				

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	0.62	0.48	0.54				
37	TE	0.38	0.43	0.34				
38	MC	0.57	0.50	0.38				
39	MC	0.71	0.34	0.42				
40	MC	0.80	0.55	0.37				
41	MC	0.66	0.43	0.37				
42	MC	0.68	0.51	0.34				
43	MC	0.58	0.36	0.40				
44	MC	0.57	0.41	0.64				
45	MC	0.50	0.25	0.50				
46	MC	0.43	0.23	0.48		+		
47	MC	0.68	0.49	0.49				
48	MC	0.74	0.45	0.46				
49	MC	0.68	0.44	0.46				
50	MC	0.72	0.36	0.47				

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
English Language Arts	3	63121	29.02	0.58	10.14	-0.11	-0.90	0	53	53	0.89	3.34
	4	64284	30.79	0.58	10.32	-0.10	-0.89	0	56	56	0.90	3.28
	5	64827	31.60	0.60	10.32	-0.13	-0.87	1	56	56	0.90	3.31
	6	63484	32.13	0.58	9.86	-0.29	-0.70	1	57	57	0.88	3.43
	7	63045	31.83	0.60	10.67	-0.19	-0.86	2	56	56	0.89	3.48
	8	63127	32.27	0.60	11.20	-0.23	-0.89	0	56	56	0.90	3.50
Mathematics	3	63254	23.46	0.56	9.39	-0.07	-1.00	0	42	42	0.92	2.69
	4	64400	22.89	0.50	9.93	0.32	-0.87	1	46	46	0.92	2.86
	5	64951	20.80	0.45	10.10	0.40	-0.82	0	46	46	0.92	2.88
	6	63562	20.96	0.46	9.79	0.42	-0.71	0	46	46	0.92	2.85
	7	63138	18.95	0.41	9.72	0.64	-0.46	0	46	46	0.91	2.89
	8	63222	20.25	0.44	9.68	0.49	-0.64	0	46	46	0.91	2.92
Science	4	64384	27.71	0.69	7.28	-0.57	-0.45	1	40	40	0.88	2.54
	8	63175	28.07	0.70	7.44	-0.75	-0.15	0	40	40	0.88	2.52
Social Studies	4	64394	25.98	0.68	7.59	-0.64	-0.40	1	38	38	0.89	2.50
	8	63138	27.27	0.68	8.31	-0.63	-0.52	0	40	40	0.91	2.56
	10	62421	31.40	0.63	10.72	-0.31	-0.92	0	50	50	0.92	3.03

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	32354	28.22	.57	10.01	.89	30767	29.86	.59	10.21	.89
	4	32708	29.85	.56	10.25	.90	31576	31.76	.60	10.30	.90
	5	33216	30.46	.58	10.26	.90	31611	32.79	.62	10.25	.90
	6	32538	30.89	.57	10.00	.88	30946	33.44	.60	9.54	.87
	7	32194	30.31	.57	10.70	.89	30851	33.41	.62	10.41	.89
	8	32474	30.61	.57	11.22	.90	30653	34.04	.63	10.91	.90
Mathematics	3	32417	23.75	.57	9.58	.92	30837	23.16	.55	9.17	.91
	4	32773	23.34	.51	10.16	.92	31627	22.41	.49	9.65	.91
	5	33282	21.02	.46	10.41	.92	31669	20.56	.45	9.76	.91
	6	32582	20.88	.46	10.03	.92	30980	21.04	.46	9.52	.91
	7	32242	19.10	.42	9.88	.91	30896	18.79	.41	9.56	.91
	8	32522	20.00	.44	9.93	.91	30700	20.52	.45	9.41	.90
Science	4	32772	27.76	.70	7.43	.88	31612	27.65	.69	7.13	.87
	8	32494	27.85	.70	7.77	.89	30681	28.30	.71	7.07	.87
Social Studies	4	32775	25.80	.68	7.77	.90	31619	26.17	.69	7.40	.89
	8	32475	26.89	.67	8.63	.91	30663	27.67	.69	7.94	.90
	10	31766	31.27	.63	11.12	.93	30655	31.54	.64	10.30	.91

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	41341	31.44	.62	9.41	.88
	4	42466	33.16	.62	9.61	.89
	5	43226	33.88	.64	9.63	.89
	6	42594	34.32	.62	9.10	.86
	7	43291	33.90	.63	9.96	.88
	8	43840	34.48	.64	10.45	.89
African American	3	7030	20.67	.42	9.10	.86
	4	7087	22.01	.42	8.98	.86
	5	6984	23.01	.44	9.15	.86
	6	6689	23.74	.44	9.09	.85
	7	6407	23.11	.44	9.91	.88
	8	6219	22.83	.44	10.09	.87
Hispanic	3	8582	24.78	.49	9.32	.87
	4	8721	26.89	.50	9.52	.88
	5	8682	27.80	.53	9.72	.88
	6	8500	28.46	.51	9.28	.86
	7	7938	27.97	.52	10.28	.88
	8	7797	27.83	.52	10.69	.89
Asian	3	2733	29.13	.58	10.07	.89
	4	2594	31.07	.58	10.33	.90
	5	2554	31.56	.59	10.39	.90
	6	2556	32.55	.58	9.93	.88
	7	2442	33.18	.62	10.61	.89
	8	2450	33.72	.62	11.15	.90
American Indian	3	748	24.93	.50	8.95	.85
	4	760	26.11	.49	9.22	.87
	5	800	26.83	.51	9.01	.86
	6	796	26.88	.49	9.22	.86
	7	784	26.64	.51	9.87	.87
	8	783	26.72	.51	10.79	.89
Two or More	3	2687	28.17	.56	10.02	.89
	4	2656	30.25	.57	10.43	.90
	5	2581	30.86	.59	10.17	.89
	6	2349	31.01	.56	9.81	.88
	7	2183	30.60	.58	10.75	.89
	8	2038	31.14	.58	11.31	.90

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	41364	25.80	.62	8.74	.91
	4	42480	25.25	.55	9.56	.91
	5	43237	23.09	.50	9.85	.91
	6	42600	23.18	.51	9.56	.91
	7	43287	20.88	.46	9.64	.91
	8	43853	22.19	.48	9.50	.90
African American	3	7026	15.61	.38	7.70	.87
	4	7087	14.78	.32	6.91	.83
	5	6979	12.71	.28	6.94	.85
	6	6690	13.10	.29	6.63	.83
	7	6416	11.61	.26	6.19	.81
	8	6219	12.50	.28	6.40	.81
Hispanic	3	8661	19.40	.46	8.48	.89
	4	8809	18.48	.40	8.44	.88
	5	8780	16.54	.36	8.54	.89
	6	8564	16.67	.36	8.01	.87
	7	8021	14.76	.32	7.88	.87
	8	7867	15.91	.35	7.92	.87
Asian	3	2767	23.97	.57	9.66	.92
	4	2611	23.55	.51	10.74	.93
	5	2575	21.64	.47	10.45	.93
	6	2565	21.97	.48	10.45	.93
	7	2450	20.42	.45	10.96	.93
	8	2462	21.93	.48	10.54	.93
American Indian	3	748	18.80	.45	8.55	.90
	4	759	18.14	.40	7.97	.87
	5	800	15.92	.35	8.32	.88
	6	795	15.25	.33	7.61	.86
	7	781	14.26	.31	7.35	.85
	8	788	14.96	.33	7.88	.87
Two or More	3	2688	21.90	.52	9.36	.92
	4	2654	21.92	.48	9.78	.91
	5	2580	19.38	.42	9.96	.92
	6	2348	19.41	.42	9.47	.91
	7	2183	17.61	.39	9.34	.91
	8	2033	18.84	.41	9.40	.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	42474	29.61	.74	6.34	.85
	8	43836	29.85	.75	6.44	.86
African American	4	7080	20.92	.53	7.18	.85
	8	6197	20.60	.52	7.60	.86
Hispanic	4	8809	24.59	.62	7.11	.85
	8	7864	24.56	.62	7.49	.87
Asian	4	2608	27.08	.68	7.32	.88
	8	2461	28.03	.70	7.24	.88
American Indian	4	759	24.29	.61	7.18	.86
	8	788	24.59	.62	7.52	.87
Two or More	4	2654	27.26	.68	7.33	.88
	8	2029	27.39	.69	7.47	.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	42477	27.86	.73	6.60	.87
	8	43824	29.04	.73	7.51	.89
	10	45292	33.31	.67	10.03	.91
African American	4	7087	19.09	.51	7.77	.87
	8	6180	19.49	.49	8.09	.88
	10	5192	21.67	.45	9.96	.89
Hispanic	4	8806	23.02	.61	7.56	.88
	8	7858	23.98	.60	8.15	.89
	10	7054	27.07	.55	10.29	.91
Asian	4	2608	25.36	.67	7.73	.89
	8	2460	27.72	.69	7.87	.90
	10	2371	31.57	.64	10.40	.91
American Indian	4	760	22.58	.60	7.35	.87
	8	788	23.05	.58	8.36	.89
	10	718	26.44	.53	10.22	.90
Two or More	4	2656	25.69	.68	7.66	.89
	8	2028	26.48	.67	8.45	.91
	10	1794	30.21	.61	11.01	.92

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	28285	24.91	.50	9.61	.88	34836	32.35	.64	9.30	.88
	4	28719	26.47	.50	9.68	.88	35565	34.28	.65	9.47	.89
	5	28111	27.16	.52	9.74	.88	36716	34.99	.66	9.43	.88
	6	26931	27.84	.51	9.52	.86	36553	35.30	.64	8.86	.86
	7	25155	27.16	.51	10.24	.88	37890	34.92	.65	9.80	.88
	8	24504	27.34	.52	10.86	.89	38623	35.41	.65	10.25	.89
Mathematics	3	28367	19.51	.47	8.71	.90	34887	26.68	.64	8.66	.91
	4	28798	18.61	.41	8.51	.89	35602	26.35	.57	9.63	.91
	5	28184	16.44	.36	8.63	.89	36767	24.14	.53	9.88	.91
	6	26992	16.57	.36	8.19	.88	36570	24.19	.53	9.60	.91
	7	25225	14.64	.32	7.81	.87	37913	21.81	.48	9.82	.91
	8	24565	15.86	.35	8.03	.87	38657	23.04	.50	9.61	.91
Science	4	28792	24.74	.62	7.39	.87	35592	30.11	.75	6.24	.85
	8	24531	24.66	.62	7.80	.88	38644	30.24	.76	6.31	.86
Social Studies	4	28801	22.86	.60	7.76	.88	35593	28.50	.75	6.42	.86
	8	24509	23.54	.59	8.49	.90	38629	29.64	.74	7.26	.89
	10	20930	26.48	.54	10.55	.91	41491	33.88	.68	9.93	.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	7714	22.11	.45	9.46	.87	55407	29.98	.60	9.86	.89
	4	7767	23.41	.45	9.62	.88	56517	31.80	.60	10.00	.89
	5	7887	22.77	.44	9.24	.87	56940	32.82	.62	9.86	.89
	6	7604	22.41	.42	8.90	.85	55880	33.46	.61	9.22	.86
	7	7362	21.36	.41	8.95	.85	55683	33.21	.62	10.10	.88
	8	7423	20.96	.41	9.26	.86	55704	33.78	.63	10.56	.89
Mathematics	3	7725	17.28	.41	9.04	.91	55529	24.32	.58	9.11	.91
	4	7775	16.76	.37	8.69	.89	56625	23.73	.52	9.79	.91
	5	7902	13.77	.30	8.15	.89	57049	21.77	.47	9.96	.92
	6	7613	13.01	.29	7.20	.86	55949	22.04	.48	9.59	.91
	7	7360	11.56	.25	6.34	.82	55778	19.92	.43	9.67	.91
	8	7430	12.28	.27	6.48	.82	55792	21.31	.47	9.54	.91
Science	4	7767	23.08	.58	7.90	.88	56617	28.34	.71	6.96	.87
	8	7416	20.76	.52	7.83	.87	55759	29.04	.73	6.82	.87
Social Studies	4	7777	20.60	.54	8.36	.90	56617	26.72	.70	7.17	.88
	8	7403	18.87	.47	8.25	.88	55735	28.39	.71	7.66	.89
	10	6498	21.78	.44	9.82	.89	55923	32.52	.65	10.25	.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	5844	23.59	.47	8.66	.84	57277	29.57	.59	10.12	.89
	4	5772	25.11	.47	8.60	.85	58512	31.35	.59	10.31	.90
	5	4830	23.95	.45	8.01	.81	59997	32.21	.61	10.24	.90
	6	3584	23.26	.42	7.47	.78	59900	32.66	.59	9.73	.88
	7	3007	21.54	.40	7.96	.81	60038	32.34	.61	10.53	.89
	8	2776	20.86	.40	7.97	.80	60351	32.80	.61	11.05	.90
Mathematics	3	5968	18.74	.45	8.23	.89	57286	23.95	.57	9.36	.92
	4	5885	17.18	.37	7.81	.87	58515	23.46	.51	9.93	.92
	5	4947	14.15	.31	6.65	.82	60004	21.35	.47	10.14	.92
	6	3671	13.14	.29	5.80	.77	59891	21.44	.47	9.78	.92
	7	3101	10.99	.24	5.19	.73	60037	19.36	.42	9.73	.91
	8	2867	12.13	.27	5.37	.73	60355	20.63	.45	9.67	.91
Science	4	5880	23.21	.58	6.82	.83	58504	28.16	.71	7.17	.88
	8	2865	19.79	.50	6.64	.81	60310	28.46	.71	7.24	.88
Social Studies	4	5881	21.48	.57	7.35	.86	58513	26.43	.70	7.46	.89
	8	2861	18.79	.47	6.87	.83	60277	27.67	.69	8.16	.90
	10	2514	20.45	.42	7.76	.82	59907	31.86	.64	10.59	.92

Table 8-27 Raw Score Descriptive Statistics by Accommodation Use

Content	Grade	Students Using Testing Accommodations					Students Not Using Accommodations				
		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	68	23.31	.47	11.18	.92	63053	29.02	.58	10.14	.89
	4	77	25.13	.48	11.62	.92	64207	30.80	.58	10.32	.90
	5	100	24.88	.47	10.71	.90	64727	31.61	.60	10.32	.90
	6	91	23.46	.44	10.05	.88	63393	32.15	.58	9.86	.88
	7	80	24.09	.46	11.18	.90	62965	31.84	.60	10.67	.89
	8	86	24.74	.47	12.17	.92	63041	32.29	.60	11.20	.90
Mathematics	3	908	13.37	.32	6.47	.82	62346	23.61	.56	9.34	.92
	4	2182	13.09	.29	5.57	.75	62218	23.23	.51	9.87	.92
	5	2627	10.86	.24	5.50	.77	62324	21.22	.46	10.04	.92
	6	2739	10.62	.23	4.76	.69	60823	21.42	.47	9.70	.91
	7	2720	9.76	.21	4.00	.58	60418	19.36	.42	9.70	.91
	8	2512	10.35	.23	4.20	.59	60710	20.66	.45	9.63	.91
Science	4	32	20.94	.52	9.20	.91	64352	27.71	.69	7.28	.88
	8	34	19.59	.49	9.56	.92	63141	28.08	.70	7.44	.88
Social Studies	4	32	19.41	.51	10.03	.93	64362	25.98	.68	7.59	.89
	8	34	20.47	.52	9.76	.92	63104	27.27	.68	8.31	.91
	10	20	27.50	.55	9.66	.89	62401	31.40	.63	10.72	.92

Table 8-28 Scale Score Descriptive Statistics

Content	Grade	N Count	Mean	SD	Skewness	Kurtosis	Min	Max	LOSS	HOSS
English Language Arts	3	63194	556.70	46.66	-0.03	0.31	330	900	330	900
	4	64354	580.90	51.81	-0.22	0.29	340	930	340	930
	5	64903	600.78	48.35	-0.17	0.75	350	940	350	940
	6	63600	609.61	50.18	-0.24	0.24	360	950	360	950
	7	63140	627.43	56.56	-0.25	0.67	370	960	370	960
	8	63248	630.98	59.94	-0.11	0.61	380	970	380	970
Mathematics	3	63314	555.94	50.87	-0.57	2.24	360	760	360	760
	4	64462	576.76	52.99	-0.58	1.72	405	800	405	800
	5	65021	598.82	56.65	-1.04	1.87	430	830	430	830
	6	63669	611.97	57.64	-0.66	1.47	440	870	440	870
	7	63218	622.82	65.55	-0.88	1.11	450	880	450	880
	8	63318	644.24	60.78	-0.76	1.65	470	890	470	890
Science	4	64448	399.03	53.13	0.13	1.36	190	600	190	600
	8	63272	595.66	52.26	-0.38	1.70	390	770	390	770
Social Studies	4	64456	398.23	53.72	-0.26	1.61	200	570	200	570
	8	63230	599.17	53.25	0.00	1.05	420	780	420	780
	10	62630	695.70	58.24	-0.63	1.73	490	890	490	890

Table 8-29 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	32397	552.80	45.84	330	900	30797	560.80	47.15	330	900
	4	32746	575.97	51.62	340	930	31608	586.01	51.50	340	821
	5	33254	595.67	48.01	350	803	31649	606.16	48.13	350	940
	6	32607	603.15	50.81	360	950	30993	616.40	48.60	360	950
	7	32237	619.33	56.95	370	960	30903	635.88	54.89	370	960
	8	32528	622.13	59.43	380	970	30720	640.36	59.05	380	970
Mathematics	3	32454	557.25	52.86	360	760	30860	554.57	48.65	360	760
	4	32802	578.46	54.98	405	800	31660	575.00	50.79	405	800
	5	33318	598.88	59.01	430	830	31703	598.77	54.06	430	830
	6	32645	610.06	60.74	440	870	31024	613.98	54.12	440	870
	7	32279	623.01	66.97	450	880	30939	622.63	64.04	450	880
	8	32566	641.80	63.61	470	890	30752	646.82	57.53	470	890
Science	4	32802	399.14	54.79	190	600	31646	398.91	51.36	190	600
	8	32542	594.09	55.24	390	770	30730	597.32	48.85	390	770
Social Studies	4	32806	397.13	54.85	200	570	31650	399.36	52.50	200	570
	8	32525	597.06	55.54	420	780	30705	601.42	50.62	420	780
	10	31881	694.53	61.76	490	890	30749	696.91	54.33	490	890

Table 8-30 Scale Score Descriptive Statistics for English Language Arts by Race/Ethnicity

Race/Ethnicity	Grade	N Count	Mean	SD	Min	Max
White	3	41366	567.73	43.06	330	900
	4	42490	592.89	47.17	340	930
	5	43253	611.53	44.53	350	940
	6	42634	620.75	46.17	360	950
	7	43341	638.48	51.84	370	960
	8	43892	642.70	55.67	380	970
African-American	3	7054	518.11	44.04	330	705
	4	7102	535.59	49.87	340	783
	5	7007	559.85	46.55	350	739
	6	6726	566.64	47.88	360	730
	7	6432	580.23	57.10	370	796
	8	6259	580.45	56.75	380	873
Hispanic	3	8597	537.53	42.27	330	721
	4	8736	561.88	47.96	349	930
	5	8697	583.30	45.10	350	940
	6	8529	591.01	46.29	360	761
	7	7951	607.65	54.50	370	871
	8	7811	607.76	55.90	380	970
Asian	3	2737	557.84	47.07	388	900
	4	2597	582.31	51.91	393	791
	5	2556	600.48	49.76	350	940
	6	2559	612.03	50.81	370	849
	7	2444	634.45	56.66	404	871
	8	2454	638.85	61.19	410	970
American Indian	3	750	538.59	40.44	330	648
	4	762	557.93	46.37	424	703
	5	802	579.54	41.14	414	720
	6	797	583.62	46.06	414	722
	7	786	600.18	54.83	370	817
	8	787	601.78	57.30	435	841
Two or More	3	2690	553.29	45.56	396	707
	4	2667	578.01	53.08	343	756
	5	2588	597.66	46.63	425	776
	6	2355	604.06	50.38	360	816
	7	2186	621.11	57.41	374	960
	8	2045	624.70	61.77	380	970

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

Race/Ethnicity	Grade	N Count	Mean	SD	Min	Max
White	3	41384	568.22	45.08	360	760
	4	42504	589.67	46.37	405	800
	5	43266	611.86	49.06	430	830
	6	42638	625.71	50.58	440	870
	7	43329	636.89	57.72	450	880
	8	43895	656.85	54.13	470	890
African-American	3	7045	512.55	52.16	360	670
	4	7102	530.23	53.38	405	800
	5	7000	547.83	62.99	430	733
	6	6730	558.99	59.86	440	742
	7	6436	565.66	67.77	450	778
	8	6252	589.94	62.88	470	804
Hispanic	3	8671	535.70	48.02	360	760
	4	8818	554.26	51.95	405	800
	5	8790	577.34	56.74	430	830
	6	8584	588.46	55.16	440	870
	7	8033	595.20	67.17	450	880
	8	7878	618.92	59.34	470	890
Asian	3	2770	560.03	52.33	360	760
	4	2613	581.28	57.01	405	800
	5	2577	605.46	53.25	430	830
	6	2568	619.27	59.10	440	870
	7	2452	629.94	71.54	450	880
	8	2465	655.46	65.62	470	890
American Indian	3	751	531.93	51.36	360	760
	4	761	553.45	48.43	405	705
	5	803	574.58	56.58	430	830
	6	796	579.22	55.95	440	753
	7	782	593.62	63.73	450	717
	8	790	609.48	63.87	470	771
Two or More	3	2693	548.47	52.15	360	760
	4	2664	571.60	54.34	405	800
	5	2585	592.73	56.62	430	830
	6	2353	603.58	58.10	440	870
	7	2186	616.14	64.74	450	880
	8	2038	636.99	59.43	470	890

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

Race/Ethnicity	Grade	N Count	Mean	SD	Min	Max
White	4	42500	412.10	48.82	190	600
	8	43877	607.33	46.84	390	770
African-American	4	7093	353.24	49.13	190	600
	8	6230	547.00	52.42	390	770
Hispanic	4	8819	376.97	47.66	190	600
	8	7875	572.52	50.00	390	770
Asian	4	2611	394.99	53.75	190	600
	8	2464	596.02	52.17	390	770
American Indian	4	761	375.98	47.69	190	581
	8	790	571.82	49.77	390	770
Two or More	4	2664	395.92	52.32	190	600
	8	2036	591.24	52.47	390	770

Table 8-33 Scale Score Descriptive Statistics for Social Studies by Race/Ethnicity

Race/Ethnicity	Grade	N Count	Mean	SD	Min	Max
White	4	42500	410.86	48.23	200	570
	8	43863	610.15	50.02	420	780
	10	45373	705.94	53.17	490	890
African-American	4	7097	351.62	55.43	200	570
	8	6209	551.98	48.79	420	780
	10	5261	642.35	63.24	490	844
Hispanic	4	8818	378.27	49.88	200	570
	8	7871	578.25	47.63	420	780
	10	7091	673.27	56.97	490	890
Asian	4	2611	395.33	55.09	200	570
	8	2463	601.81	50.13	420	780
	10	2378	697.25	56.04	490	890
American Indian	4	762	375.62	47.31	200	533
	8	789	573.67	48.50	420	780
	10	728	670.38	57.19	490	890
Two or More	4	2668	396.18	53.62	200	570
	8	2035	594.25	53.18	420	780
	10	1799	690.12	59.99	490	890

Table 8-34 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	28344	538.06	43.99	330	761	34850	571.85	43.14	330	900
	4	28767	559.40	49.47	340	930	35587	598.28	46.89	340	930
	5	28163	580.15	45.93	350	803	36740	616.60	44.00	350	940
	6	27021	587.87	48.06	360	813	36579	625.67	45.44	360	950
	7	25215	603.01	55.07	370	871	37925	643.67	51.47	370	960
	8	24599	604.77	57.68	380	970	38649	647.67	55.20	380	970
Mathematics	3	28417	535.32	50.14	360	760	34897	572.74	44.93	360	760
	4	28843	554.64	52.09	405	800	35619	594.68	46.55	405	800
	5	28232	575.22	59.12	430	750	36789	616.94	47.27	430	830
	6	27078	586.44	57.72	440	870	36591	630.87	49.76	440	870
	7	25272	594.32	66.95	450	880	37946	641.81	57.20	450	880
	8	24639	617.09	61.35	470	890	38679	661.53	53.68	470	890
Science	4	28838	378.12	50.10	190	600	35610	415.96	49.35	190	600
	8	24603	572.67	52.18	390	770	38669	610.28	46.75	390	770
Social Studies	4	28843	376.90	52.42	200	570	35613	415.50	48.28	200	570
	8	24582	575.86	50.41	420	780	38648	614.00	49.58	420	780
	10	21071	669.75	59.44	490	890	41559	708.85	52.98	490	890

Table 8-35 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	7729	524.93	44.68	330	779	55465	561.12	45.19	330	900
	4	7776	542.63	52.38	340	782	56578	586.16	49.46	340	930
	5	7904	558.61	47.25	350	803	56999	606.63	45.52	350	940
	6	7629	560.10	47.52	360	747	55971	616.35	46.62	360	950
	7	7383	570.93	54.10	370	831	55757	634.91	52.50	370	960
	8	7458	571.38	53.35	380	782	55790	638.95	56.17	380	970
Mathematics	3	7739	519.72	59.27	360	760	55575	560.99	47.44	360	760
	4	7783	539.40	59.57	405	800	56679	581.89	49.88	405	800
	5	7918	552.78	65.91	430	830	57103	605.21	52.13	430	830
	6	7635	557.33	62.08	440	870	56034	619.42	52.80	440	870
	7	7380	564.30	68.61	450	880	55838	630.56	61.07	450	880
	8	7460	588.09	63.17	470	890	55858	651.74	56.37	470	890
Science	4	7778	367.21	55.09	190	600	56670	403.40	51.34	190	600
	8	7446	547.33	55.04	390	770	55826	602.10	48.35	390	770
Social Studies	4	7785	362.15	59.61	200	570	56671	403.18	50.90	200	570
	8	7438	549.58	50.96	420	780	55792	605.79	49.96	420	780
	10	6551	642.54	63.50	490	890	56079	701.91	54.31	490	890

Table 8-36 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	5854	532.15	39.27	330	703	57340	559.20	46.63	330	900
	4	5786	553.00	43.84	340	758	58568	583.66	51.72	340	930
	5	4840	565.64	38.51	350	723	60063	603.62	47.95	350	940
	6	3596	565.77	38.74	360	703	60004	612.23	49.57	360	950
	7	3013	574.12	46.73	370	763	60127	630.10	55.68	370	960
	8	2783	571.61	45.58	380	738	60465	633.72	59.10	380	970
Mathematics	3	5974	532.19	47.26	360	760	57340	558.42	50.59	360	760
	4	5891	547.14	51.35	405	800	58571	579.74	52.23	405	800
	5	4954	564.38	54.77	430	726	60067	601.66	55.86	430	830
	6	3677	565.03	53.69	440	753	59992	614.85	56.62	440	870
	7	3106	561.86	65.68	450	750	60112	625.97	63.99	450	880
	8	2872	591.57	56.81	470	890	60446	646.74	59.82	470	890
Science	4	5886	367.35	44.70	190	600	58562	402.21	52.86	190	600
	8	2869	542.38	46.49	390	770	60403	598.19	51.15	390	770
Social Studies	4	5886	368.17	47.71	200	570	58570	401.25	53.36	200	570
	8	2866	549.58	39.87	420	708	60364	601.53	52.66	420	780
	10	2528	638.13	53.42	490	769	60102	698.12	57.18	490	890

Table 8-37 Scale Score Descriptive Statistics by Accommodation Use

Content	Grade	Students Using Testing Accommodations					Students Not Using Accommodations				
		N Count	Mean	SD	Min	Max	N Count	Mean	SD	Min	Max
English Language Arts	3	68	530.21	50.76	404	636	63126	556.72	46.65	330	900
	4	77	549.13	63.75	340	683	64277	580.94	51.78	340	930
	5	100	568.29	52.53	439	738	64803	600.83	48.33	350	940
	6	91	563.33	53.63	434	686	63509	609.67	50.14	360	950
	7	80	587.41	63.37	466	811	63060	627.48	56.53	370	960
	8	86	587.26	67.99	395	721	63162	631.04	59.91	380	970
Mathematics	3	910	496.66	53.04	360	650	62404	556.81	50.32	360	760
	4	2184	515.99	52.53	405	647	62278	578.89	51.73	405	800
	5	2633	530.80	61.85	430	694	62388	601.70	54.59	430	830
	6	2747	539.03	55.13	440	712	60922	615.26	55.54	440	870
	7	2724	547.73	62.66	450	708	60494	626.20	63.63	450	880
	8	2523	571.77	58.67	470	729	60795	647.25	58.97	470	890
Science	4	32	350.91	64.74	245	470	64416	399.05	53.11	190	600
	8	34	541.68	66.71	390	699	63238	595.69	52.23	390	770
Social Studies	4	32	352.16	73.54	200	497	64424	398.25	53.70	200	570
	8	34	559.59	55.37	450	657	63196	599.20	53.24	420	780
	10	20	676.70	45.14	578	749	62610	695.70	58.25	490	890

Table 8-38 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 8-39 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	22.78	37.47	32.58	7.17	39.75
4	340–545	546–591	592–649	650–930	24.04	32.06	35.72	8.19	43.91
5	350–563	564–609	610–669	670–940	21.53	34.30	37.40	6.77	44.17
6	360–571	572–621	622–670	671–950	22.06	35.08	32.73	10.12	42.86
7	370–584	585–637	638–696	697–960	21.29	33.57	35.72	9.43	45.15
8	380–591	592–651	652–707	708–970	24.66	38.01	27.93	9.40	37.33

Table 8-40 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	18.68	31.48	38.47	11.37	49.83
4	405–535	536–587	588–632	633–800	18.37	37.17	32.71	11.74	44.46
5	430–573	574–610	611–657	658–830	24.73	29.32	35.05	10.90	45.95
6	440–581	582–625	626–687	688–870	24.78	31.27	37.78	6.18	43.96
7	450–605	606–646	647–711	712–880	31.36	29.67	34.33	4.64	38.97
8	470–619	620–666	667–717	718–890	27.95	35.44	28.71	7.90	36.61

Table 8-41 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	15.24	34.07	34.43	16.26	50.69
8	390–551	552–599	600–644	645–770	17.18	33.96	34.16	14.70	48.86

Table 8-42 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	22.14	24.20	31.69	21.97	53.66
8	420–562	563–598	599–639	640–780	22.84	24.95	31.85	20.36	52.21
10	490–669	670–702	703–740	741–890	28.19	23.61	28.01	20.18	48.20

Table 8-43 Percentage of Students in Each Performance Level by Subgroup, English Language Arts

Grade	Performance Level	Examinees		Gender		Race/Ethnicity						ELP		Disability		SES		Accommodations	
		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	Using Accommodations	No Accommodations
3	BB	14397	22.8	20.8	24.7	14.0	55.6	36.5	22.0	33.7	25.5	21.1	39.5	49.5	19.1	36.1	12.0	45.6	22.8
	B	23678	37.5	36.0	38.8	37.3	32.0	41.3	38.9	43.5	39.3	36.8	43.8	34.7	37.9	40.2	35.2	25.0	37.5
	P	20591	32.6	34.4	30.9	39.4	11.3	20.1	30.5	21.2	28.3	34.3	15.6	14.0	35.2	21.1	41.9	27.9	32.6
	A	4528	7.2	8.8	5.6	9.3	1.1	2.2	8.6	1.6	6.9	7.8	1.1	1.9	7.9	2.6	10.9	1.5	7.2
Total		63194	100.0	30797	32397	41366	7054	8597	2737	750	2690	57340	5854	7729	55465	28344	34850	68	63126
4	BB	15468	24.0	21.0	27.0	15.6	58.3	35.8	23.1	38.8	25.4	22.3	41.4	52.5	20.1	37.9	12.8	46.8	24.0
	B	20630	32.1	31.3	32.8	31.2	28.8	37.3	33.7	37.3	33.3	31.3	39.5	29.6	32.4	35.6	29.2	22.1	32.1
	P	22986	35.7	37.9	33.6	42.6	11.9	23.9	34.0	21.9	33.4	37.4	18.5	15.5	38.5	23.8	45.4	28.6	35.7
	A	5270	8.2	9.9	6.6	10.5	1.0	3.0	9.2	2.0	7.8	8.9	.7	2.3	9.0	2.7	12.6	2.6	8.2
Total		64354	100.0	31608	32746	42490	7102	8736	2597	762	2667	58568	5786	7776	56578	28767	35587	77	64277
5	BB	13975	21.5	18.2	24.7	13.8	53.7	32.7	20.7	34.4	23.0	19.6	45.5	56.5	16.7	35.2	11.0	50.0	21.5
	B	22259	34.3	33.6	34.9	33.0	32.7	39.6	37.8	43.6	35.7	33.6	42.9	29.7	34.9	38.5	31.1	31.0	34.3
	P	24274	37.4	39.7	35.2	44.5	12.7	25.1	33.8	20.0	35.0	39.5	11.4	12.6	40.8	24.3	47.5	16.0	37.4
	A	4395	6.8	8.5	5.1	8.6	.9	2.5	7.6	2.0	6.3	7.3	.2	1.2	7.5	2.1	10.4	3.0	6.8
Total		64903	100.0	31649	33254	43253	7007	8697	2556	802	2588	60063	4840	7904	56999	28163	36740	100	64803

Table 8-43 Percentage of Students in Each Performance Level by Subgroup, English Language Arts (cont.)

Grade	Performance Level	Examinees		Gender		Race/Ethnicity						ELP		Disability		SES		Accommodations	
		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	Using Accommodations	No Accommodations
6	BB	14032	22.1	17.6	26.3	14.4	54.1	33.1	21.2	40.2	24.9	20.1	54.4	61.3	16.7	36.0	11.8	62.6	22.0
	B	22312	35.1	34.5	35.6	34.1	32.5	40.7	34.7	39.0	38.6	34.8	39.2	28.6	36.0	38.9	32.3	20.9	35.1
	P	20818	32.7	35.6	30.0	38.6	12.1	22.5	32.4	18.1	27.7	34.3	6.1	8.6	36.0	21.6	40.9	14.3	32.8
	A	6438	10.1	12.2	8.1	12.9	1.2	3.7	11.7	2.8	8.8	10.7	.4	1.6	11.3	3.5	15.0	2.2	10.1
Total		63600	100.0	30993	32607	42634	6726	8529	2559	797	2355	60004	3596	7629	55971	27021	36579	91	63509
7	BB	13440	21.3	16.6	25.8	14.2	53.2	32.9	17.3	36.3	25.5	19.5	57.1	60.2	16.1	35.5	11.8	50.0	21.2
	B	21193	33.6	32.2	34.9	33.3	30.6	37.3	33.3	37.2	33.5	33.4	35.9	29.5	34.1	37.0	31.3	30.0	33.6
	P	22555	35.7	39.5	32.1	41.0	14.4	25.7	36.7	23.8	33.3	37.2	6.7	9.2	39.2	24.2	43.4	17.5	35.7
	A	5952	9.4	11.7	7.3	11.6	1.9	4.1	12.6	2.8	7.7	9.9	.3	1.1	10.5	3.3	13.5	2.5	9.4
Total		63140	100.0	30903	32237	43341	6432	7951	2444	786	2186	60127	3013	7383	55757	25215	37925	80	63060
8	BB	15595	24.7	19.6	29.4	17.2	58.5	38.1	20.5	42.1	27.5	22.8	66.1	66.0	19.1	40.3	14.7	50.0	24.6
	B	24040	38.0	37.3	38.7	38.3	31.3	40.7	38.5	39.6	40.1	38.3	30.7	27.1	39.5	39.1	37.3	33.7	38.0
	P	17665	27.9	31.0	25.0	32.9	8.7	17.5	28.0	15.9	23.9	29.1	3.0	6.1	30.8	17.1	34.8	12.8	28.0
	A	5948	9.4	12.1	6.8	11.5	1.5	3.7	13.0	2.4	8.5	9.8	.2	.8	10.6	3.5	13.2	3.5	9.4
Total		63248	100.0	30720	32528	43892	6259	7811	2454	787	2045	60465	2783	7458	55790	24599	38649	86	63162

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

Grade	Performance Level	Examinees		Gender		Race/Ethnicity						ELP		Disability		SES		Accommodations	
		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	Using Accommodations	No Accommodations
3	BB	11829	18.7	18.8	18.6	10.5	49.4	30.4	17.9	33.2	22.6	17.3	32.3	43.9	15.2	30.6	8.9	61.4	18.1
	B	19933	31.5	32.9	30.1	29.0	34.1	39.2	32.3	38.5	35.6	30.5	40.5	31.6	31.5	37.3	26.7	30.2	31.5
	P	24355	38.5	38.5	38.5	45.8	15.4	26.5	34.7	24.0	32.9	39.9	24.3	20.4	41.0	27.9	47.0	8.0	38.9
	A	7197	11.4	9.8	12.8	14.7	1.1	3.9	15.0	4.4	8.9	12.3	2.9	4.1	12.4	4.1	17.3	0.3	11.5
Total		63314	100.0	30860	32454	41384	7045	8671	2770	751	2693	57340	5974	7739	55575	28417	34897	910	62404
4	BB	11842	18.4	18.4	18.4	10.4	49.7	30.1	17.2	29.6	20.6	16.7	34.5	44.4	14.8	30.3	8.7	61.9	16.8
	B	23962	37.2	39.6	34.9	34.9	38.9	44.9	38.1	48.5	39.7	36.3	46.3	36.1	37.3	43.9	31.7	32.8	37.3
	P	21088	32.7	32.1	33.3	39.7	10.1	20.8	28.7	19.1	29.1	34.3	16.6	15.1	35.1	21.6	41.7	5.1	33.7
	A	7570	11.7	9.9	13.5	15.0	1.3	4.2	16.0	2.9	10.6	12.7	2.6	4.3	12.8	4.2	17.9	0.2	12.1
Total		64462	100.0	31660	32802	42504	7102	8818	2613	761	2664	58571	5891	7783	56679	28843	35619	2184	62278
5	BB	16082	24.7	23.9	25.5	15.7	60.1	39.1	21.5	43.6	28.8	22.9	47.5	57.2	20.2	40.1	13.0	73.1	22.7
	B	19065	29.3	30.8	27.9	28.4	27.4	33.9	31.3	31.3	32.1	28.7	36.9	25.7	29.8	32.6	26.8	20.4	29.7
	P	22787	35.0	35.8	34.3	42.0	11.4	23.0	32.9	22.0	29.5	36.7	14.6	14.0	38.0	23.6	43.8	6.2	36.3
	A	7087	10.9	9.5	12.2	13.9	1.1	3.9	14.2	3.1	9.6	11.7	1.0	3.1	12.0	3.7	16.4	0.4	11.3
Total		65021	100.0	31703	33318	43266	7000	8790	2577	803	2585	60067	4954	7918	57103	28232	36789	2633	62388

Table 8-44 Percentage of Students in Each Performance Level by Subgroup, Mathematics (cont.)

Grade	Performance Level	Examinees		Gender		Race/Ethnicity						ELP		Disability		SES		Accommodations	
		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	Using Accommodations	No Accommodations
6	BB	15774	24.8	22.9	26.6	15.7	61.1	38.8	21.5	47.1	30.3	22.8	56.4	63.7	19.5	40.3	13.3	78.7	22.3
	B	19908	31.3	32.4	30.1	30.5	27.5	36.8	32.6	34.4	33.1	31.0	35.0	24.5	32.2	34.9	28.6	18.1	31.9
	P	24053	37.8	39.1	36.6	45.9	10.8	22.8	35.6	17.3	31.7	39.6	8.2	10.6	41.5	23.0	48.7	3.1	39.3
	A	3934	6.2	5.6	6.7	7.9	0.5	1.7	10.4	1.1	4.8	6.5	0.4	1.2	6.9	1.7	9.5	0.1	6.5
Total		63669	100.0	31024	32645	42638	6730	8584	2568	796	2353	59992	3677	7635	56034	27078	36591	2747	60922
7	BB	19823	31.4	31.0	31.7	21.8	70.1	48.9	30.4	50.3	36.4	29.2	72.9	72.0	26.0	49.4	19.3	84.3	29.0
	B	18758	29.7	30.8	28.6	31.0	20.7	30.4	26.5	31.5	30.9	30.1	21.4	18.7	31.1	30.0	29.4	13.2	30.4
	P	21703	34.3	34.0	34.7	41.5	8.9	19.4	33.3	18.0	28.4	35.8	5.4	8.6	37.7	19.4	44.3	2.5	35.8
	A	2934	4.6	4.3	5.0	5.7	0.3	1.3	9.8	.3	4.3	4.9	0.3	0.7	5.2	1.2	6.9	0.0	4.9
Total		63218	100.0	30939	32279	43329	6436	8033	2452	782	2186	60112	3106	7380	55838	25272	37946	2724	60494
8	BB	17697	27.9	25.3	30.4	19.0	66.8	44.8	23.2	50.5	33.2	26.0	67.9	69.2	22.4	45.7	16.7	82.2	25.7
	B	22439	35.4	37.1	33.9	36.5	25.8	37.4	34.6	33.5	36.2	35.8	27.7	23.8	37.0	36.0	35.1	16.4	36.2
	P	18181	28.7	30.2	27.3	34.7	6.6	15.3	29.4	13.9	24.6	29.9	4.1	5.9	31.8	16.1	36.8	1.3	29.8
	A	5001	7.9	7.4	8.4	9.8	0.8	2.5	12.8	2.0	6.0	8.3	0.2	1.1	8.8	2.3	11.5	0.0	8.2
Total		63318	100.0	30752	32566	43895	6252	7878	2465	790	2038	60446	2872	7460	55858	24639	38679	2523	60795

Table 8-45 Percentage of Students in Each Performance Level by Subgroup, Science

Grade	Performance Level	Examinees		Gender		Race/Ethnicity						ELP		Disability		SES		Accommodations	
		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	Using Accommodations	No Accommodations
4	BB	9820	15.2	14.6	15.9	7.9	44.6	25.3	16.8	26.4	16.8	13.7	30.5	35.6	12.4	25.5	6.9	59.4	15.2
	B	21957	34.1	34.5	33.6	30.8	39.0	43.6	37.5	44.4	35.5	32.8	46.8	37.6	33.6	41.5	28.1	12.5	34.1
	P	22192	34.4	35.3	33.6	40.3	13.9	25.0	31.4	22.5	33.0	35.9	19.9	19.9	36.4	25.5	41.7	18.8	34.4
	A	10479	16.3	15.6	16.9	21.0	2.6	6.1	14.4	6.7	14.8	17.6	2.7	6.9	17.5	7.5	23.4	9.4	16.3
Total		64448	100.0	31646	32802	42500	7093	8819	2611	761	2664	58562	5886	7778	56670	28838	35610	32	64416
8	BB	10871	17.2	15.0	19.3	9.9	50.8	29.4	16.0	30.1	19.7	15.4	53.9	52.2	12.5	30.4	8.7	55.9	17.2
	B	21487	34.0	35.4	32.6	31.7	35.9	42.9	36.9	41.9	36.0	33.7	39.1	32.8	34.1	39.8	30.3	20.6	34.0
	P	21611	34.2	35.6	32.8	40.0	10.9	22.2	32.8	23.3	32.2	35.5	6.7	11.8	37.1	23.6	40.9	20.6	34.2
	A	9303	14.7	14.1	15.3	18.4	2.4	5.5	14.3	4.7	12.1	15.4	0.3	3.2	16.2	6.2	20.1	2.9	14.7
Total		63272	100.0	30730	32542	43877	6230	7875	2464	790	2036	60403	2869	7446	55826	24603	38669	34	63238

Table 8-46 Percentage of Students in Each Performance Level by Subgroup, Social Studies

Grade	Performance Level	Examinees		Gender		Race/Ethnicity						ELP		Disability		SES		Accommodations	
		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	Using Accommodations	No Accommodations
4	BB	14268	22.1	20.4	23.8	13.5	55.9	34.5	24.6	35.7	23.5	20.2	41.3	48.1	18.6	35.4	11.4	53.1	22.1
	B	15599	24.2	24.8	23.6	22.6	25.1	29.4	27.5	31.9	24.3	23.5	31.3	24.2	24.2	28.5	20.7	18.8	24.2
	P	20429	31.7	33.1	30.3	36.2	14.3	25.9	27.1	24.0	32.0	32.7	22.0	18.7	33.5	25.5	36.7	15.6	31.7
	A	14160	22.0	21.6	22.3	27.7	4.7	10.2	20.7	8.4	20.2	23.6	5.5	9.0	23.8	10.7	31.1	12.5	22.0
Total		64456	100.0	31650	32806	42500	7097	8818	2611	762	2668	58570	5886	7785	56671	28843	35613	32	64424
8	BB	14444	22.8	20.5	25.1	15.3	58.2	35.5	19.2	40.9	25.7	21.0	61.7	62.4	17.6	38.0	13.2	47.1	22.8
	B	15774	24.9	26.0	24.0	23.4	26.0	31.2	27.6	27.0	26.2	24.8	29.0	22.1	25.3	28.8	22.5	29.4	24.9
	P	20139	31.9	33.3	30.4	36.0	12.7	24.3	33.2	24.3	31.1	33.0	8.5	11.2	34.6	24.4	36.6	11.8	31.9
	A	12873	20.4	20.2	20.5	25.2	3.1	9.0	20.1	7.7	17.0	21.3	.8	4.2	22.5	8.8	27.7	11.8	20.4
Total		63230	100.0	30705	32525	43863	6209	7871	2463	789	2035	60364	2866	7438	55792	24582	38648	34	63196
10	BB	17655	28.2	26.4	29.9	21.0	65.5	43.3	27.3	45.6	33.6	26.3	72.9	66.8	23.7	45.7	19.3	40.0	28.2
	B	14790	23.6	24.9	22.4	23.6	19.0	26.1	25.8	27.7	22.2	23.7	20.7	18.4	24.2	24.9	23.0	25.0	23.6
	P	17544	28.0	29.9	26.2	31.3	11.4	21.4	26.8	18.4	25.8	29.0	5.5	10.3	30.1	20.4	31.9	30.0	28.0
	A	12641	20.2	18.9	21.5	24.0	4.0	9.2	20.1	8.2	18.3	21.0	.9	4.5	22.0	9.0	25.8	5.0	20.2
Total		62630	100.0	30749	31881	45373	5261	7091	2378	728	1799	60102	2528	6551	56079	21071	41559	20	62610

Table 8-47a 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	63194	A	Reading—Key Ideas and Details	4	3	10	6.07	0.61	2.66	60.57	23.89
	63194	B	Reading—Craft & Structure/Integration of Knowledge & Ideas	6	1	8	3.63	0.46	1.91	46.01	18.68
	63194	C	Reading—Vocabulary Use	3	1	4	2.15	0.54	1.26	54.00	24.44
	63194	D	Writing/Language—Text Types and Purposes	2	2	12	5.34	0.58	2.17	44.81	14.70
	63194	E	Writing/Language—Research	4	1	6	3.88	0.66	1.55	64.29	20.45
	63194	F	Writing/Language—Language Conventions	4	1	6	3.47	0.56	1.64	57.93	22.17
	63194	G	Listening	3	2	7	4.47	0.65	1.95	63.06	22.84
4	64354	A	Reading—Key Ideas and Details	4	4	12	6.86	0.58	2.83	57.07	21.22
	64354	B	Reading—Craft & Structure/Integration of Knowledge & Ideas	7	0	7	4.20	0.60	1.94	60.05	23.58
	64354	C	Reading—Vocabulary Use	4	1	5	2.53	0.51	1.50	50.97	24.69
	64354	D	Writing/Language—Text Types and Purposes	3	2	12	5.22	0.57	2.22	43.88	15.06
	64354	E	Writing/Language—Research	1	3	6	3.37	0.55	1.54	56.15	20.34
	64354	F	Writing/Language—Language Conventions	2	2	6	4.25	0.69	1.28	70.29	15.31
	64354	G	Listening	4	2	8	4.35	0.57	1.89	54.21	18.06

Table 8-47a 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	64903	A	Reading—Key Ideas and Details	8	2	12	6.63	0.56	2.83	55.45	21.27
	64903	B	Reading—Craft & Structure/Integration of Knowledge & Ideas	8	0	8	4.35	0.55	1.96	54.59	19.85
	64903	C	Reading—Vocabulary Use	2	1	4	2.50	0.60	1.02	62.67	16.66
	64903	D	Writing/Language—Text Types and Purposes	3	2	12	5.22	0.61	2.31	43.88	15.65
	64903	E	Writing/Language—Research	2	2	6	3.48	0.60	1.66	57.93	21.81
	64903	F	Writing/Language—Language Conventions	4	1	6	4.10	0.68	1.49	67.68	19.04
	64903	G	Listening	4	2	8	5.30	0.68	2.03	65.81	21.61
6	63600	A	Reading—Key Ideas and Details	0	6	11	6.53	0.60	2.29	59.31	17.92
	63600	B	Reading—Craft & Structure/Integration of Knowledge & Ideas	5	2	9	5.32	0.59	2.24	59.09	21.19
	63600	C	Reading—Vocabulary Use	5	0	5	3.09	0.62	1.35	61.88	21.54
	63600	D	Writing/Language—Text Types and Purposes	4	1	12	5.72	0.62	2.32	47.93	15.61
	63600	E	Writing/Language—Research	2	2	6	2.97	0.48	1.56	49.89	17.75
	63600	F	Writing/Language—Language Conventions	1	3	6	3.70	0.58	1.35	61.54	15.78
	63600	G	Listening	4	2	8	4.78	0.58	1.89	59.56	19.23

Table 8-47a 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	63140	A	Reading—Key Ideas and Details	8	2	12	6.89	0.57	3.34	57.52	25.83
	63140	B	Reading—Craft & Structure/Integration of Knowledge & Ideas	3	2	7	3.52	0.50	1.71	50.51	18.14
	63140	C	Reading—Vocabulary Use	1	3	5	3.52	0.71	1.29	69.72	20.40
	63140	D	Writing/Language—Text Types and Purposes	1	3	12	5.72	0.59	2.43	47.97	16.56
	63140	E	Writing/Language—Research	4	1	6	3.34	0.57	1.69	55.67	21.72
	63140	F	Writing/Language—Language Conventions	4	1	6	4.14	0.70	1.39	68.53	18.08
	63140	G	Listening	4	2	8	4.67	0.59	2.00	58.40	20.30
8	63248	A	Reading—Key Ideas and Details	5	4	12	7.02	0.60	2.93	58.44	21.83
	63248	B	Reading—Craft & Structure/Integration of Knowledge & Ideas	6	0	6	3.54	0.59	1.60	58.97	21.69
	63248	C	Reading—Vocabulary Use	2	2	6	3.73	0.62	1.73	61.92	24.84
	63248	D	Writing/Language—Text Types and Purposes	2	2	12	6.30	0.61	2.61	52.83	18.20
	63248	E	Writing/Language—Research	5	1	6	3.36	0.56	1.66	56.12	21.91
	63248	F	Writing/Language—Language Conventions	4	1	6	3.46	0.59	1.41	58.06	18.34
	63248	G	Listening	4	2	8	4.84	0.62	2.17	60.31	22.19

Table 8-47b 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	63194	Listening	3	2	7	4.47	0.65	1.95	63.06	22.84
	63194	Reading	13	5	22	11.85	0.54	4.95	54.00	21.41
	63194	Writing	10	4	24	12.69	0.60	4.42	53.01	17.35
4	64354	Listening	4	2	8	4.35	0.57	1.89	54.21	18.06
	64354	Reading	15	5	24	13.59	0.57	5.55	56.64	22.37
	64354	Writing	6	7	24	12.84	0.60	4.10	53.55	15.84
5	64903	Listening	4	2	8	5.30	0.68	2.03	65.81	21.61
	64903	Reading	18	3	24	13.48	0.56	4.97	56.31	19.66
	64903	Writing	9	5	24	12.80	0.63	4.48	53.44	17.42
6	63600	Listening	4	2	8	4.78	0.58	1.89	59.56	19.23
	63600	Reading	10	8	25	14.94	0.60	5.06	59.77	19.48
	63600	Writing	7	6	24	12.39	0.57	4.14	51.78	15.64
7	63140	Listening	4	2	8	4.67	0.59	2.00	58.40	20.30
	63140	Reading	12	7	24	13.94	0.58	5.41	58.06	21.74
	63140	Writing	9	5	24	13.20	0.62	4.48	55.08	17.38
8	63248	Listening	4	2	8	4.84	0.62	2.17	60.31	22.19
	63248	Reading	13	6	24	14.28	0.60	5.52	59.47	22.22
	63248	Writing	11	4	24	13.13	0.59	4.73	54.93	18.58

Table 8-48 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	63314	A	Operations and Algebraic Thinking	6	3	9	4.97	0.55	2.31	55.24	23.00
	63314	B	Number and Operations in Base Ten	6	2	8	4.67	0.59	2.20	58.41	24.41
	63314	C	Number and Operations—Fractions	6	2	8	4.85	0.61	2.07	60.67	22.07
	63314	D	Measurement and Data	7	3	10	4.60	0.46	2.40	46.34	21.50
	63314	E	Geometry	5	2	7	4.36	0.62	1.96	62.07	23.76
4	64462	A	Operations and Algebraic Thinking	8	2	10	5.11	0.51	2.25	51.20	19.16
	64462	B	Number and Operations in Base Ten	5	4	9	5.00	0.56	2.27	55.42	22.29
	64462	C	Number and Operations—Fractions	8	2	10	3.88	0.39	2.94	39.48	26.96
	64462	D	Measurement and Data	9	1	10	4.98	0.50	2.45	49.89	21.67
	64462	E	Geometry	7	0	7	3.91	0.56	1.90	55.60	21.12
5	65021	A	Operations and Algebraic Thinking	5	4	9	4.69	0.52	2.52	51.83	25.57
	65021	B	Number and Operations in Base Ten	5	4	9	4.39	0.49	2.30	48.70	22.73
	65021	C	Number and Operations—Fractions	6	3	9	3.48	0.39	2.33	39.08	22.82
	65021	D	Measurement and Data	7	3	10	4.39	0.44	2.18	43.95	18.03
	65021	E	Geometry	5	4	9	3.83	0.43	2.48	42.67	24.23

Table 8-48 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	63669	E	Geometry	5	2	7	3.12	0.45	1.91	44.80	22.96
	63669	F	Ratios and Proportional Relationships	2	5	7	2.54	0.36	1.85	36.54	22.87
	63669	G	The Number System	7	4	11	5.59	0.51	2.72	50.64	22.72
	63669	H	Expressions and Equations	8	3	11	4.96	0.45	2.71	45.13	22.01
	63669	I	Statistics and Probability	8	2	10	4.74	0.48	2.31	47.22	19.45
7	63218	E	Geometry	7	3	10	3.74	0.38	2.18	37.61	18.12
	63218	F	Ratios and Proportional Relationships	6	2	8	3.79	0.48	2.25	47.13	25.36
	63218	G	The Number System	5	2	7	2.76	0.39	1.92	39.48	23.42
	63218	H	Expressions and Equations	8	2	10	3.77	0.38	2.43	37.93	21.50
	63218	I	Statistics and Probability	9	2	11	4.88	0.45	2.65	44.25	21.20
8	63318	E	Geometry	7	3	10	4.38	0.44	2.51	43.79	21.84
	63318	G	The Number System	5	3	8	3.09	0.39	2.08	38.98	21.48
	63318	H	Expressions and Equations	8	2	10	4.39	0.44	2.49	44.17	22.00
	63318	I	Statistics and Probability	7	1	8	3.58	0.45	1.94	44.93	20.14
	63318	J	Functions	8	2	10	4.80	0.48	2.48	47.89	22.05

Table 8-49 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	64448	AB	Science Connections & Nature of Science	7	0	7	5.10	0.73	1.68	72.84	20.32
	64448	C	Science Inquiry	9	0	9	5.85	0.65	2.14	65.36	20.80
	64448	D	Physical Science	5	0	5	3.34	0.67	1.09	67.00	13.55
	64448	E	Earth and Space Science	4	1	5	2.86	0.57	1.28	57.76	17.83
	64448	F	Life & Environmental Science	6	0	6	4.48	0.75	1.29	74.64	16.77
	64448	GH	Science Applications & Personal Social Perspectives	8	0	8	6.07	0.76	1.84	75.70	20.09
8	63272	AB	Science Connections & Nature of Science	7	0	7	5.31	0.76	1.66	75.80	20.46
	63272	C	Science Inquiry	9	0	9	6.68	0.74	2.20	74.27	22.09
	63272	D	Physical Science	5	0	5	3.87	0.77	1.21	76.78	18.79
	63272	E	Earth and Space Science	5	0	5	3.09	0.62	1.24	62.10	16.50
	63272	F	Life & Environmental Science	6	0	6	3.76	0.63	1.52	63.25	19.46
	63272	GH	Science Applications & Personal Social Perspectives	8	0	8	5.35	0.67	1.60	67.11	16.87

Table 8-50 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	64456	A	Geography	8	1	9	5.96	0.66	2.12	66.36	20.41
	64456	B	History	7	2	9	6.06	0.67	2.21	67.60	21.88
	64456	C	Political Science and Citizenship	6	1	7	4.97	0.71	1.58	70.86	17.73
	64456	D	Economics	6	0	6	3.39	0.57	1.64	57.30	22.54
	64456	E	The Behavioral Sciences	7	0	7	5.59	0.80	1.64	79.31	20.56
8	63230	A	Geography	9	1	10	7.01	0.70	2.43	70.07	22.07
	63230	B	History	11	1	12	8.30	0.69	2.66	69.23	20.25
	63230	C	Political Science and Citizenship	5	1	6	3.69	0.62	1.54	61.94	19.90
	63230	D	Economics	6	0	6	3.91	0.65	1.66	65.47	23.02
	63230	E	The Behavioral Sciences	5	1	6	4.34	0.73	1.50	72.11	21.22
10	62630	A	Geography	9	1	10	6.07	0.61	2.48	60.78	21.75
	62630	B	History	13	0	13	8.18	0.64	3.10	63.14	21.69
	62630	C	Political Science and Citizenship	9	1	10	6.46	0.65	2.63	64.78	23.44
	62630	D	Economics	7	1	8	5.22	0.66	1.97	65.03	20.96
	62630	E	The Behavioral Sciences	8	1	9	5.42	0.61	2.16	60.43	20.67

Table 8-51 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	39	0	39	0	35
	2	40	70	40	62	36	58
	3	71	92	63	84	59	86
	4	93	100	85	100	87	100
Reading—Craft & Structure	1	0	30	0	37	0	36
	2	31	47	38	65	37	57
	3	48	74	66	92	58	83
	4	75	100	93	100	84	100
Reading—Vocabulary Use	1	0	31	0	27	0	47
	2	32	60	28	54	48	63
	3	61	90	55	87	64	86
	4	91	100	88	100	87	100
Writing/Language—Text Types and Purposes	1	0	33	0	33	0	31
	2	34	48	34	45	32	45
	3	49	63	46	62	46	64
	4	64	100	63	100	65	100
Writing/Language—Research	1	0	48	0	40	0	38
	2	49	72	41	60	39	63
	3	73	88	61	82	64	87
	4	89	100	83	100	88	100
Writing/Language—Language Conventions	1	0	38	0	60	0	53
	2	39	65	61	75	54	74
	3	66	87	76	87	75	89
	4	88	100	88	100	90	100
Listening	1	0	44	0	41	0	46
	2	45	73	42	57	47	73
	3	74	91	58	77	74	91
	4	92	100	78	100	92	100
Reading	1	0	34	0	36	0	37
	2	35	60	37	61	38	59
	3	61	85	62	87	60	85
	4	86	100	88	100	86	100
Writing	1	0	38	0	41	0	38
	2	39	58	42	56	39	57
	3	59	75	57	73	58	76
	4	76	100	74	100	77	100

Table 8-51 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	45	0	31	0	41
	2	46	64	32	63	42	68
	3	65	80	64	90	69	85
	4	81	100	91	100	86	100
Reading—Craft & Structure	1	0	40	0	35	0	41
	2	41	65	36	52	42	69
	3	66	84	53	73	70	85
	4	85	100	74	100	86	100
Reading—Vocabulary Use	1	0	43	0	54	0	42
	2	44	68	55	76	43	74
	3	69	87	77	90	75	92
	4	88	100	91	100	93	100
Writing/Language—Text Types and Purposes	1	0	35	0	35	0	39
	2	36	51	36	49	40	59
	3	52	65	50	66	60	74
	4	66	100	67	100	75	100
Writing/Language—Research	1	0	35	0	35	0	39
	2	36	52	36	59	40	64
	3	53	71	60	83	65	84
	4	72	100	84	100	85	100
Writing/Language—Language Conventions	1	0	49	0	54	0	45
	2	50	65	55	73	46	64
	3	66	79	74	88	65	79
	4	80	100	89	100	80	100
Listening	1	0	44	0	39	0	43
	2	45	66	40	62	44	70
	3	67	80	63	83	71	86
	4	81	100	84	100	87	100
Reading	1	0	42	0	37	0	41
	2	43	65	38	62	42	70
	3	66	83	63	85	71	86
	4	84	100	86	100	87	100
Writing	1	0	38	0	40	0	41
	2	39	55	41	57	42	61
	3	56	70	58	76	62	78
	4	71	100	77	100	79	100

Table 8-52 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	31	0	32	0	29
	2	32	55	33	52	30	54
	3	56	84	53	74	55	85
	4	85	100	75	100	86	100
Number and Operations in Base Ten	1	0	32	0	32	0	29
	2	33	59	33	59	30	49
	3	60	88	60	81	50	78
	4	89	100	82	100	79	100
Number and Operations—Fractions	1	0	37	0	13	0	18
	2	38	60	14	35	19	35
	3	61	88	36	78	36	72
	4	89	100	79	100	73	100
Measurement and Data	1	0	24	0	28	0	29
	2	25	45	29	51	30	41
	3	46	73	52	78	42	68
	4	74	100	79	100	69	100
Geometry	1	0	36	0	34	0	21
	2	37	65	35	57	22	42
	3	66	89	58	82	43	76
	4	90	100	83	100	77	100

Table 8-52 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	24	0	25	0	27
	2	25	43	26	36	28	49
	3	44	85	37	75	50	77
	4	86	100	76	100	78	100
Ratios and Proportional Relationships*	1	0	17	0	29		
	2	18	35	30	55		
	3	36	76	56	88		
	4	77	100	89	100		
The Number System	1	0	32	0	22	0	22
	2	33	54	23	42	23	42
	3	55	84	43	82	43	73
	4	85	100	83	100	74	100
Expressions and Equations	1	0	25	0	22	0	25
	2	26	45	23	38	26	48
	3	46	83	39	81	49	80
	4	84	100	82	100	81	100
Statistics and Probability	1	0	32	0	29	0	29
	2	33	47	30	47	30	49
	3	48	78	48	84	50	76
	4	79	100	85	100	77	100
Functions**	1					0	31
	2					32	55
	3					56	80
	4					81	100

* Content standard in grades 6 and 7 only.

** Content standard in grade 8 only.

Table 8-53 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	54
	2	50	77	55	82
	3	78	92	83	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	53	0	59
	2	54	67	60	82
	3	68	79	83	93
	4	80	100	94	100
Earth and Space Science	1	0	37	0	46
	2	38	57	47	63
	3	58	74	64	77
	4	75	100	78	100
Life & Environmental Science	1	0	56	0	42
	2	57	76	43	64
	3	77	90	65	83
	4	91	100	84	100
Science Applications & Social and Personal Perspectives	1	0	52	0	50
	2	53	80	51	69
	3	81	94	70	82
	4	95	100	83	100

Table 8-54 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
Geography	1	0	50	0	53	0	46
	2	51	67	54	73	47	63
	3	68	83	74	89	64	81
	4	84	100	90	100	82	100
History	1	0	48	0	53	0	47
	2	49	68	54	71	48	66
	3	69	86	72	86	67	84
	4	87	100	87	100	85	100
Political Science and Citizenship	1	0	58	0	44	0	48
	2	59	73	45	61	49	69
	3	74	84	62	79	70	87
	4	85	100	80	100	88	100
Economics	1	0	37	0	46	0	51
	2	38	54	47	68	52	68
	3	55	76	69	86	69	84
	4	77	100	87	100	85	100
The Behavioral Sciences	1	0	66	0	57	0	46
	2	67	85	58	76	47	63
	3	86	95	77	89	64	79
	4	96	100	90	100	80	100

Table 8-55 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
	2018	63194	556.70	46.66
4	2016	62609	582.71	49.41
	2017	64423	585.26	52.44
	2018	64354	580.90	51.81
5	2016	62300	599.62	51.11
	2017	62995	603.24	51.00
	2018	64903	600.78	48.35
6	2016	62728	610.36	52.16
	2017	62754	614.59	49.82
	2018	63600	609.61	50.18
7	2016	62084	623.84	54.85
	2017	63091	626.80	59.14
	2018	63140	627.43	56.56
8	2016	61486	637.23	57.27
	2017	62109	637.69	61.61
	2018	63248	630.98	59.94

Table 8-56 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
	2018	63314	555.94	50.87
4	2016	62674	573.45	56.15
	2017	64533	574.33	54.92
	2018	64462	576.76	52.99
5	2016	62368	599.57	50.19
	2017	63152	599.73	51.00
	2018	65021	598.82	56.65
6	2016	62772	612.67	53.00
	2017	62847	612.93	54.81
	2018	63669	611.97	57.64
7	2016	62144	627.49	57.40
	2017	63200	627.48	58.65
	2018	63218	622.82	65.55
8	2016	61551	640.79	57.54
	2017	62175	641.11	59.36
	2018	63318	644.24	60.78

Table 8-57 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
	2018	64448	399.03	53.13
8	2016	61471	597.92	52.54
	2017	62113	594.12	51.25
	2018	63272	595.66	52.26

Table 8-58 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
	2018	64456	398.23	53.72
8	2016	61496	598.06	51.68
	2017	62079	597.60	54.26
	2018	63230	599.17	53.25
10	2016	63991	698.51	53.74
	2017	63764	696.92	56.56
	2018	62630	695.70	58.24

Table 8-59 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
	2018	63194	22.78	37.47	32.58	7.17	39.75
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
	2018	64354	24.04	32.06	35.72	8.19	43.91
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
	2018	64903	21.53	34.30	37.40	6.77	44.17
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
	2018	63600	22.06	35.08	32.73	10.12	42.86
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
	2018	63140	21.29	33.57	35.72	9.43	45.15
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
	2018	63248	24.66	38.01	27.93	9.40	37.33

Table 8-60 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
	2018	63314	18.68	31.48	38.47	11.37	49.83
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
	2018	64462	18.37	37.17	32.71	11.74	44.46
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
	2018	65021	24.73	29.32	35.05	10.90	45.95
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
	2018	63669	24.78	31.27	37.78	6.18	43.96
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
	2018	63218	31.36	29.67	34.33	4.64	38.97
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
	2018	63318	27.95	35.44	28.71	7.90	36.61

Table 8-61 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
	2018	64448	15.24	34.07	34.43	16.26	50.69
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
	2018	63272	17.18	33.96	34.16	14.70	48.86

Table 8-62 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
	2018	64456	22.14	24.20	31.69	21.97	53.66
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
	2018	63230	22.84	24.95	31.85	20.36	52.21
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
	2018	62630	28.19	23.61	28.01	20.18	48.20

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 2018 Wisconsin Forward Exam was estimated in four ways:

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

The present chapter addresses American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (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. (p. 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. (p. 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. (p. 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. (p. 45)

Standard 2.13 The standard error of measurement, both overall and conditional (if reported), should be provided in units of each reported score. (p. 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. (p. 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. (p. 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 standard errors of measurement (CSEMs) 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 2018 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 CSEM 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, English Language Arts (ELA) TDA items were scored by the artificial intelligence (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: for example, they provide estimates of reliability for student scores in ELA or Mathematics. 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, which is 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 multiple-choice (MC) and constructed-response (CR) items. Cronbach's alpha (α) 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, σ_x^2 = the total score variance, and σ_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, English language proficiency, and accommodation use.

As indicated in Table 9-1, reliability was highest in Mathematics and Social Studies. As shown in the "Total" column, reliability ranges from 0.88 to 0.90 across grades for ELA, from 0.91 to 0.92 across grades for Mathematics, 0.88 for both grades in Science, and from 0.89 to 0.92 across grades for Social Studies. All reliability coefficients would ideally 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 ELA grade 3, 6, and 7 assessments, Science assessments, or the Social Studies grade 4 assessment would not 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, 6, and 7 would need to be increased from 53, 57, and 56 points to 59, 70, and 62 score points, respectively. For the current Science assessments in grades 4 and 8, the increase would need to be from 40 to 49 score points. For the current Social Studies assessment in grade 4, the increase would need to be from 38 to 42 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 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 8-27) 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 0.05 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 and American Indian students.

The differences between economically disadvantaged and not economically disadvantaged students were within 0.04 of one another for all grades and content areas. The differences between disabled and not disabled students were within 0.09 of one another for all grades and content areas. The greatest differences were between fully English proficient and limited English proficient students and between students using and not using testing accommodations, with consistently lower reliability among limited English proficient students and students using testing accommodations. In fact, the test reliability coefficients for limited English proficient students were lower than for other subgroups for most grades and content areas. The reliability coefficients for students using testing accommodations in Mathematics were much lower than the reliability coefficients for students using testing accommodations in ELA. The reliability coefficients were not computed for students using testing accommodations in Science and Social Studies, because the number of students using testing accommodations in these subject areas was less than 50. 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 proficient student test scores and Mathematics scores for students using testing accommodations, it was observed that the variance of these scores was often lower than the variance of the scores for other groups. The limited English proficient student groups and students using testing accommodations in Mathematics appear to be more homogeneous on the ability being measured by the test, leading to lower test reliability for these groups of students.

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 an individual's obtained score can be expected to fall within two SEMs 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 and the Social Studies grade 10 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 and Social Studies grade 10 test have more raw score points and relatively larger raw score SDs than other content areas. For every grade and content area, the CSEM for individual scale scores is provided in the scoring tables previously discussed in Part 6 (Tables 6-8 through 6-24).

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.

Table 9-3 shows these reliability coefficients by content standard and domain. The last column presents the reliability for the total test per grade for each content area (with all content

standards or domains) for all examinees. It is clear that the reliability per content standard or 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 for ELA per domain ranged from 7 or 8 in Listening, 22 to 25 in Reading, and 22 to 25 in Writing.. The number of score points ranged from 4 to 12 per standard for ELA, from 7 to 11 per standard for Mathematics, from 5 to 9 per standard for Science, and from 6 to 13 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).

As shown in Table 9-3, the reliability ranges by content standard/domain were as follows:

- For ELA, reliability indices by content standard or domain ranged from 0.33 (for standard C in grade 5) to 0.85 (for the Reading domain in grade 4).
- For Mathematics, reliability indices by content standard ranged from 0.61 (for standard I in grade 8) to 0.81 (for standard C in grade 4).
- For Science, reliability indices by content standard ranged from 0.29 (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 0.50 (for standard C in grade 8) to 0.75 (for standards B and 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 the SEM, the CSEM expresses the degree of measurement error in scale score units and is conditioned on the ability of the student. 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):

$$\text{CSEM}(\theta_i) = \frac{1}{\sqrt{I(\theta_i)}},$$

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 item response theory (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 6-8 through 6-24) 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 CSEM tend to be higher at the low and high ends of the scale score range. The CSEM 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 CSEM 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 4×4 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 ₁₁	P ₂₁	P ₃₁	P ₄₁	P. ₁
Level 2	P ₁₂	P ₂₂	P ₃₂	P ₄₂	P. ₂
Level 3	P ₁₃	P ₂₃	P ₃₃	P ₄₃	P. ₃
Level 4	P ₁₄	P ₂₄	P ₃₄	P ₄₄	P. ₄
Sum	P _{1.}	P _{2.}	P _{3.}	P _{4.}	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).

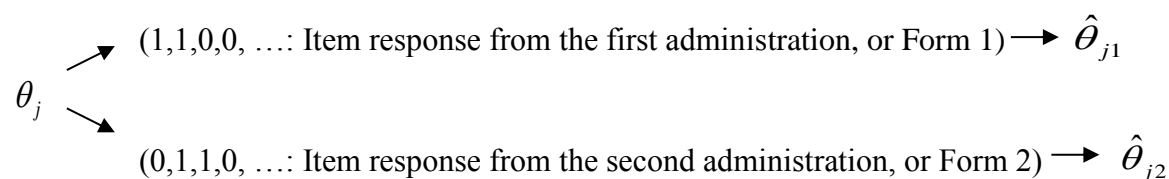
9.2.1 Kolen and Kim’s Method for Pattern Scoring

As stated in Part 6, when 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, θ_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 θ_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. Because the values displayed have been rounded to two decimal places, this sum is not always identical to the sum of the values shown in the table.

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 0.30, kappa (*k*) is 0.62, 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, for ELA, 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. For Mathematics, the classification consistency was greater than or equal to 0.89 and the classification accuracy was greater than or equal to 0.92 for all subgroups across all grades. For Science, the classification consistency was greater than or equal to 0.86 and the classification accuracy was greater than or equal to 0.89 for all subgroups across both grades. For Social Studies, the classification consistency was greater than or equal to 0.86 and the classification accuracy was greater than or equal to 0.91 for all subgroups across all grades. The kappa values were greater than or equal to 0.57 for all subgroups in ELA, greater than or equal to 0.66 for all subgroups in Mathematics, greater than or equal to 0.62 for all subgroups in Science, and greater than or equal to 0.59 for all subgroups in Social Studies. The lowest kappa values were observed for the limited English proficiency subgroups in ELA, Science, and Social Studies and for students using testing accommodations in Mathematics. This is consistent with the trend of the test reliability coefficients, which were found to be lower for these groups of students compared to other subgroups. Because the number of students using testing accommodations in Science and Social Studies was less than 50, the indices for classification consistency and classification accuracy were not computed for students using testing accommodations in these subject areas. The indices for classification consistency and classification accuracy for students using testing accommodations in ELA should be interpreted with caution because of the low number of students using accommodations in this content area.

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 there are N responses, each of which is scored twice. The two scores of response n are denoted 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

$$\bar{X}_{..} = \frac{1}{N} \sum_{n=1}^N (X_{n1} + X_{n2})/2$$

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, a square table of counts may be created 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 an 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, ρ_{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 does take this into consideration. In general, k will have values equal to or less 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 in which the scores of the two raters differed by more than one raw score point. For example, as shown in Table 9-22, for the grade 3 TDA item, the exact agreement, adjacent agreement, and discrepant agreement rates are 77.71%, 21.83%, and 0.47%, respectively. “Mean” reflects the item mean score from the second reads (by human scorers). “No. 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 exact rater agreement percentages were at acceptable levels for all TDA items and ranged from 66.66% in grade 8 to 78.43% in grade 6. The combined exact and adjacent agreement percentages were approximately 99% in all grades. The intraclass correlation coefficients ranged from 0.83 in grade 3 to 0.88 in grade 6. The weighted kappa ranged from 0.66 in grade 3 to 0.77 in grade 6, indicating good rater agreement for all TDA items.

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 an acceptable 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		Accommodations		
			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	Students with Accommodations	Students without Accommodations
English Language Arts	3	0.89	0.89	0.89	0.88	0.86	0.87	0.89	0.85	0.89	0.89	0.84	0.87	0.89	0.88	0.88	0.92	0.89
	4	0.90	0.90	0.90	0.89	0.86	0.88	0.90	0.87	0.90	0.90	0.85	0.88	0.89	0.88	0.89	0.92	0.90
	5	0.90	0.90	0.90	0.89	0.86	0.88	0.90	0.86	0.89	0.90	0.81	0.87	0.89	0.88	0.88	0.90	0.90
	6	0.88	0.87	0.88	0.86	0.85	0.86	0.88	0.86	0.88	0.88	0.78	0.85	0.86	0.86	0.86	0.88	0.88
	7	0.89	0.89	0.89	0.88	0.88	0.88	0.89	0.87	0.89	0.89	0.81	0.85	0.88	0.88	0.88	0.90	0.89
	8	0.90	0.90	0.90	0.89	0.87	0.89	0.90	0.89	0.90	0.90	0.80	0.86	0.89	0.89	0.89	0.92	0.90
Mathematics	3	0.92	0.91	0.92	0.91	0.87	0.89	0.92	0.90	0.92	0.92	0.89	0.91	0.91	0.90	0.91	0.82	0.92
	4	0.92	0.91	0.92	0.91	0.83	0.88	0.93	0.87	0.91	0.92	0.87	0.89	0.91	0.89	0.91	0.75	0.92
	5	0.92	0.91	0.92	0.91	0.85	0.89	0.93	0.88	0.92	0.92	0.82	0.89	0.92	0.89	0.91	0.77	0.92
	6	0.92	0.91	0.92	0.91	0.83	0.87	0.93	0.86	0.91	0.92	0.77	0.86	0.91	0.88	0.91	0.69	0.91
	7	0.91	0.91	0.91	0.91	0.81	0.87	0.93	0.85	0.91	0.91	0.73	0.82	0.91	0.87	0.91	0.58	0.91
	8	0.91	0.90	0.91	0.90	0.81	0.87	0.93	0.87	0.90	0.91	0.73	0.82	0.91	0.87	0.91	0.59	0.91
Science	4	0.88	0.87	0.88	0.85	0.85	0.85	0.88	0.86	0.88	0.88	0.83	0.88	0.87	0.87	0.85	-	0.88
	8	0.88	0.87	0.89	0.86	0.86	0.87	0.88	0.87	0.88	0.88	0.81	0.87	0.87	0.88	0.86	-	0.88
Social Studies	4	0.89	0.89	0.90	0.87	0.87	0.88	0.89	0.87	0.89	0.89	0.86	0.90	0.88	0.88	0.86	-	0.89
	8	0.91	0.90	0.91	0.89	0.88	0.89	0.90	0.89	0.91	0.90	0.83	0.88	0.89	0.90	0.89	-	0.91
	10	0.92	0.91	0.93	0.91	0.89	0.91	0.91	0.90	0.92	0.92	0.82	0.89	0.91	0.91	0.91	-	0.92

Note: The reliability coefficients were not computed for students using testing accommodations in Science and Social Studies because the number of students using testing accommodations in these subject areas was less than 50.

Table 9-2 Standard Error of Measurement for Total Group and Subgroups

Content	Grade	Total	Gender		Race/Ethnicity					ELP		Disability		SES		Accommodations		
			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	Students with Accommodations	Students without Accommodations
English Language Arts	3	3.34	3.33	3.33	3.29	3.37	3.41	3.34	3.42	3.35	3.32	3.42	3.36	3.32	3.39	3.26	3.24	3.34
	4	3.28	3.27	3.29	3.23	3.37	3.36	3.31	3.33	3.31	3.27	3.38	3.33	3.26	3.36	3.20	3.38	3.28
	5	3.31	3.28	3.32	3.25	3.42	3.39	3.32	3.40	3.32	3.30	3.45	3.39	3.28	3.39	3.22	3.46	3.31
	6	3.43	3.40	3.44	3.38	3.53	3.50	3.43	3.46	3.46	3.42	3.52	3.47	3.40	3.50	3.35	3.53	3.43
	7	3.48	3.45	3.48	3.44	3.50	3.53	3.47	3.52	3.50	3.47	3.50	3.43	3.45	3.52	3.42	3.55	3.48
	8	3.50	3.46	3.52	3.44	3.57	3.57	3.48	3.55	3.54	3.49	3.54	3.47	3.46	3.57	3.41	3.51	3.50
Mathematics	3	2.69	2.70	2.68	2.66	2.75	2.77	2.66	2.76	2.72	2.68	2.77	2.74	2.68	2.76	2.63	2.75	2.69
	4	2.86	2.86	2.84	2.86	2.81	2.87	2.81	2.87	2.86	2.86	2.86	2.84	2.86	2.87	2.84	2.80	2.86
	5	2.88	2.89	2.86	2.90	2.72	2.84	2.86	2.83	2.86	2.88	2.80	2.73	2.89	2.84	2.90	2.63	2.88
	6	2.85	2.85	2.84	2.86	2.74	2.83	2.81	2.81	2.83	2.85	2.77	2.72	2.86	2.83	2.85	2.64	2.85
	7	2.89	2.89	2.89	2.93	2.69	2.82	2.85	2.83	2.87	2.90	2.68	2.69	2.91	2.82	2.93	2.60	2.90
	8	2.92	2.93	2.90	2.94	2.78	2.89	2.88	2.85	2.92	2.92	2.78	2.77	2.93	2.89	2.93	2.70	2.92
Science	4	2.54	2.54	2.53	2.43	2.82	2.71	2.58	2.71	2.56	2.51	2.78	2.73	2.51	2.70	2.39	-	2.54
	8	2.52	2.52	2.53	2.42	2.86	2.73	2.53	2.73	2.57	2.50	2.91	2.85	2.47	2.72	2.39	-	2.52
Social Studies	4	2.50	2.49	2.50	2.41	2.75	2.66	2.54	2.67	2.51	2.48	2.72	2.68	2.47	2.66	2.36	-	2.50
	8	2.56	2.55	2.56	2.47	2.81	2.73	2.53	2.75	2.60	2.54	2.87	2.81	2.52	2.72	2.44	-	2.56
	10	3.03	3.05	3.01	2.96	3.25	3.17	3.05	3.20	3.05	3.02	3.31	3.22	3.01	3.19	2.94	-	3.03

Note: The SEMs were not computed for students using testing accommodations in Science and Social Studies because the number of students using testing accommodations in these subject areas was less than 50.

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	0.69	0.46	0.54	0.49	0.53	0.57	0.59	0.80	0.77	0.89
4	0.70	0.66	0.62	0.49	0.58	0.42	0.48	0.85	0.74	0.90
5	0.71	0.57	0.33	0.47	0.54	0.52	0.68	0.81	0.75	0.90
6	0.62	0.60	0.53	0.48	0.34	0.41	0.55	0.81	0.68	0.88
7	0.78	0.40	0.50	0.42	0.49	0.52	0.60	0.83	0.73	0.89
8	0.69	0.54	0.61	0.47	0.56	0.53	0.58	0.83	0.76	0.90

Mathematics

Grade	Alpha per Content Standard										
	A	B	C	D	E	F	G	H	I	J	Total
3	0.72	0.72	0.69	0.69	0.68						0.92
4	0.65	0.70	0.81	0.71	0.65						0.92
5	0.76	0.69	0.71	0.62	0.74						0.92
6					0.66	0.69	0.75	0.73	0.66		0.92
7					0.63	0.72	0.67	0.70	0.70		0.91
8					0.70		0.67	0.70	0.61	0.71	0.91

Science

Grade	Alpha per Content Standard						
	A/B	C	D	E	F	G/H	Total
4	0.63	0.66	0.29	0.40	0.50	0.67	0.88
8	0.64	0.74	0.52	0.34	0.50	0.53	0.88

Social Studies

Grade	Alpha per Content Standard					
	A	B	C	D	E	Total
4	0.64	0.69	0.54	0.59	0.69	0.89
8	0.74	0.72	0.50	0.63	0.63	0.91
10	0.69	0.75	0.75	0.63	0.64	0.92

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	1.48	1.40	0.86	1.55	1.06	1.08	1.25	2.23	2.12	3.34
4	1.56	1.13	0.93	1.59	1.00	0.98	1.36	2.15	2.08	3.28
5	1.52	1.28	0.84	1.68	1.13	1.03	1.15	2.16	2.24	3.31
6	1.41	1.42	0.93	1.67	1.27	1.04	1.27	2.21	2.34	3.43
7	1.58	1.32	0.91	1.85	1.21	0.97	1.27	2.26	2.34	3.48
8	1.64	1.08	1.07	1.91	1.10	0.97	1.40	2.25	2.30	3.50

Mathematics

Grade	SEM per Content Standard										
	A	B	C	D	E	F	G	H	I	J	Total
3	1.23	1.17	1.15	1.33	1.11						2.69
4	1.33	1.24	1.26	1.32	1.13						2.86
5	1.23	1.28	1.25	1.34	1.27						2.88
6					1.12	1.04	1.36	1.41	1.35		2.85
7					1.32	1.18	1.11	1.33	1.46		2.89
8					1.37		1.20	1.36	1.20	1.34	2.92

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	2.54
8	1.00	1.12	0.84	1.01	1.07	1.10	2.52

Social Studies

Grade	SEM per Content Standard					
	A	B	C	D	E	Total
4	1.27	1.22	1.07	1.04	0.92	2.50
8	1.23	1.41	1.09	1.01	0.91	2.56
10	1.39	1.56	1.31	1.19	1.29	3.03

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
Below Basic	0.18	0.05	0.00	0.00	0.23
Basic	0.05	0.26	0.05	0.00	0.36
Proficient	0.00	0.06	0.24	0.03	0.33
Advanced	0.00	0.00	0.03	0.05	0.08
Sum	0.23	0.37	0.32	0.09	

Indexes for Classification Consistency and Classification Accuracy

Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Classification Consistency (P)	0.90	0.89	0.94	0.73
Probability of Chance	0.64	0.52	0.85	0.30
Kappa (k)	0.73	0.77	0.61	0.62
Classification Accuracy	0.93	0.92	0.96	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
Below Basic	0.20	0.04	0.00	0.00	0.24
Basic	0.04	0.22	0.05	0.00	0.31
Proficient	0.00	0.06	0.26	0.03	0.35
Advanced	0.00	0.00	0.03	0.06	0.10
Sum	0.24	0.32	0.35	0.10	

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	0.74
Probability of Chance	0.63	0.51	0.83	0.29
Kappa (k)	0.76	0.78	0.62	0.64
Classification Accuracy	0.93	0.93	0.96	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
Below Basic	0.17	0.04	0.00	0.00	0.22
Basic	0.04	0.25	0.05	0.00	0.34
Proficient	0.00	0.05	0.28	0.03	0.36
Advanced	0.00	0.00	0.03	0.05	0.08
Sum	0.22	0.34	0.36	0.08	

Indexes for Classification Consistency and Classification Accuracy

Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Classification Consistency (P)	0.92	0.90	0.95	0.76
Probability of Chance	0.66	0.51	0.85	0.30
Kappa (k)	0.75	0.80	0.63	0.66
Classification Accuracy	0.94	0.92	0.96	0.82

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
Below Basic	0.18	0.05	0.00	0.00	0.23
Basic	0.04	0.24	0.06	0.00	0.35
Proficient	0.00	0.05	0.21	0.04	0.30
Advanced	0.00	0.00	0.04	0.08	0.12
Sum	0.22	0.34	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	0.88	0.92	0.71
Probability of Chance	0.65	0.51	0.79	0.28
Kappa (k)	0.74	0.76	0.60	0.60
Classification Accuracy	0.94	0.91	0.94	0.79

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	0.17	0.05	0.00	0.00	0.22
Basic	0.04	0.23	0.05	0.00	0.32
Proficient	0.00	0.05	0.26	0.03	0.35
Advanced	0.00	0.00	0.03	0.08	0.11
Sum	0.21	0.33	0.35	0.11	

Indexes for Classification Consistency and Classification Accuracy

Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Classification Consistency (P)	0.91	0.89	0.94	0.73
Probability of Chance	0.66	0.50	0.80	0.29
Kappa (k)	0.73	0.78	0.67	0.63
Classification Accuracy	0.94	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	0.21	0.04	0.00	0.00	0.25
Basic	0.04	0.28	0.05	0.00	0.37
Proficient	0.00	0.06	0.18	0.03	0.27
Advanced	0.00	0.00	0.04	0.07	0.11
Sum	0.25	0.37	0.27	0.10	

Indexes for Classification Consistency and Classification Accuracy

Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Classification Consistency (P)	0.92	0.89	0.93	0.74
Probability of Chance	0.63	0.53	0.81	0.29
Kappa (k)	0.79	0.76	0.64	0.64
Classification Accuracy	0.94	0.92	0.95	0.81

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	0.07	0.02	0.00	0.10
Proficient	0.00	0.02	0.24	0.06	0.32
Advanced	0.00	0.00	0.06	0.47	0.53
Sum	0.05	0.10	0.32	0.53	

Indexes for Classification Consistency and Classification Accuracy

Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Classification Consistency (P)	0.98	0.96	0.89	0.82
Probability of Chance	0.91	0.74	0.50	0.40
Kappa (k)	0.79	0.84	0.77	0.71
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	0.07	0.02	0.00	0.00	0.09
Basic	0.02	0.15	0.04	0.00	0.21
Proficient	0.00	0.04	0.25	0.05	0.34
Advanced	0.00	0.00	0.04	0.32	0.36
Sum	0.09	0.22	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	0.92	0.91	0.78
Probability of Chance	0.83	0.58	0.54	0.30
Kappa (k)	0.71	0.81	0.80	0.69
Classification Accuracy	0.97	0.94	0.93	0.84

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	0.15	0.05	0.00	0.23
Proficient	0.00	0.04	0.28	0.03	0.35
Advanced	0.00	0.00	0.03	0.16	0.19
Sum	0.22	0.24	0.35	0.19	

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	0.65	0.50	0.69	0.26
Kappa (k)	0.77	0.83	0.80	0.69
Classification Accuracy	0.94	0.94	0.95	0.84

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	0.06	0.00	0.00	0.28
Basic	0.06	0.20	0.05	0.00	0.31
Proficient	0.00	0.05	0.29	0.02	0.36
Advanced	0.00	0.00	0.02	0.04	0.06
Sum	0.28	0.30	0.36	0.06	

Indexes for Classification Consistency and Classification Accuracy

Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Classification Consistency (P)	0.89	0.90	0.97	0.76
Probability of Chance	0.60	0.51	0.89	0.30
Kappa (k)	0.72	0.80	0.70	0.65
Classification Accuracy	0.92	0.93	0.98	0.83

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	0.36	0.07	0.00	0.00	0.43
Basic	0.08	0.20	0.04	0.00	0.32
Proficient	0.00	0.04	0.18	0.01	0.23
Advanced	0.00	0.00	0.01	0.01	0.02
Sum	0.44	0.31	0.23	0.02	

Indexes for Classification Consistency and Classification Accuracy

Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Classification Consistency (P)	0.85	0.91	0.99	0.75
Probability of Chance	0.51	0.62	0.96	0.34
Kappa (k)	0.69	0.77	0.65	0.62
Classification Accuracy	0.89	0.94	0.99	0.82

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	0.46	0.08	0.00	0.00	0.54
Basic	0.08	0.21	0.03	0.00	0.33
Proficient	0.00	0.03	0.08	0.00	0.11
Advanced	0.00	0.00	0.01	0.01	0.02
Sum	0.55	0.32	0.12	0.02	

Indexes for Classification Consistency and Classification Accuracy

Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Classification Consistency (P)	0.83	0.94	0.99	0.77
Probability of Chance	0.50	0.77	0.96	0.41
Kappa (k)	0.67	0.76	0.72	0.60
Classification Accuracy	0.88	0.96	0.99	0.84

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	0.09	0.03	0.00	0.00	0.12
Basic	0.03	0.23	0.06	0.00	0.33
Proficient	0.00	0.06	0.23	0.06	0.35
Advanced	0.00	0.00	0.06	0.13	0.20
Sum	0.13	0.33	0.35	0.20	

Indexes for Classification Consistency and Classification Accuracy

Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Classification Consistency (P)	0.94	0.87	0.88	0.69
Probability of Chance	0.78	0.50	0.68	0.28
Kappa (k)	0.71	0.74	0.61	0.56
Classification Accuracy	0.95	0.91	0.91	0.77

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	0.09	0.03	0.00	0.00	0.12
Basic	0.03	0.22	0.06	0.00	0.31
Proficient	0.00	0.07	0.24	0.07	0.38
Advanced	0.00	0.00	0.06	0.13	0.19
Sum	0.12	0.32	0.36	0.20	

Indexes for Classification Consistency and Classification Accuracy

Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Classification Consistency (P)	0.94	0.86	0.86	0.67
Probability of Chance	0.79	0.51	0.68	0.29
Kappa (k)	0.74	0.72	0.56	0.54
Classification Accuracy	0.96	0.90	0.90	0.76

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	0.20	0.07	0.32
Advanced	0.00	0.00	0.06	0.19	0.25
Sum	0.18	0.24	0.32	0.26	

Indexes for Classification Consistency and Classification Accuracy

Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Classification Consistency (P)	0.92	0.88	0.87	0.67
Probability of Chance	0.71	0.51	0.62	0.26
Kappa (k)	0.74	0.74	0.65	0.55
Classification Accuracy	0.95	0.91	0.90	0.76

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	0.19	0.04	0.00	0.00	0.23
Basic	0.04	0.16	0.06	0.00	0.26
Proficient	0.00	0.07	0.19	0.05	0.31
Advanced	0.00	0.00	0.05	0.14	0.20
Sum	0.23	0.27	0.30	0.20	

Indexes for Classification Consistency and Classification Accuracy

Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Classification Consistency (P)	0.92	0.87	0.89	0.68
Probability of Chance	0.64	0.50	0.68	0.26
Kappa (k)	0.77	0.73	0.66	0.57
Classification Accuracy	0.94	0.91	0.92	0.77

Table 9-21 Classification Consistency and Classification Accuracy for Social Studies Grade 10

Contingency Table with All Cut Scores

Performance Level	Below Basic	Basic	Proficient	Advanced	Sum
Below Basic	0.20	0.04	0.00	0.00	0.24
Basic	0.04	0.15	0.06	0.00	0.25
Proficient	0.00	0.05	0.19	0.05	0.30
Advanced	0.00	0.00	0.04	0.17	0.21
Sum	0.24	0.24	0.29	0.22	

Indexes for Classification Consistency and Classification Accuracy

Indexes	Cut 1	Cut 2	Cut 3	All Cuts
Classification Consistency (P)	0.91	0.89	0.90	0.71
Probability of Chance	0.63	0.50	0.66	0.25
Kappa (k)	0.76	0.78	0.72	0.61
Classification Accuracy	0.94	0.92	0.93	0.79

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	1	4	77.71	21.83	0.47	0.83	0.66	1.44	11600	5910	2145	299	16	3230
4	1	4	77.46	21.48	1.06	0.84	0.67	1.45	10647	5224	2260	295	44	2824
5	1	4	78.23	21.01	0.76	0.85	0.70	1.47	14524	4856	2211	296	24	7137
6	1	4	78.43	20.99	0.58	0.88	0.77	1.63	10113	3764	2533	601	38	3177
7	1	4	72.94	26.02	1.05	0.86	0.73	1.81	8938	2689	2840	612	63	2734
8	1	4	66.66	32.06	1.28	0.86	0.73	2.09	11746	2314	2987	1051	148	5246

Note: The sum of the modes of agreement and codes may not equal exactly 100% due to rounding.

Note: TDA item scores presented in this table reflect a 1–4-point scoring rubric (before application of a weight of 2).

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* (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. (p. 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 that either supports or challenges 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 internal structure of the test, evidence of validity based on relationship between test scores and 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 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 2018 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 test development served as critical guides throughout the 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 use of test scores 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, EBSR, 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' interpretations and variations in the 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 the 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 area 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 2018 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 2018 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 in this part, as well as the way the standard is addressed.

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. (p. 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. (p. 64)

DRC conducted DIF studies following the operational administration of the Wisconsin Forward Exam. Items are often evaluated for possible DIF in the field test phase of 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 are first conducted after the items are field tested. Items flagged for DIF during Spring 2017 field test analysis were reviewed again by DRC content experts for potential bias and were avoided during the selection of the Spring 2018 operational test forms. Only items deemed to be free of bias were included in the selection of the Spring 2018 forms. An additional DIF analysis was performed on the Spring 2018 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. (p. 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. (p. 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. Therefore, 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 that 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, English language proficiency, and accommodation use groups. The reference and focal groups are as follows:

- **Gender**—reference group: male students; focal group: female students
- **Race/Ethnicity**—reference group: White students; focal groups African-American, Asian, Hispanic, American Indian students
- **Socioeconomic status**—reference group: not economically disadvantaged students; focal group: economically disadvantaged students
- **Disability status**—reference group: students without disabilities; focal group: students with disabilities
- **English language proficiency**—reference group: fully English proficient students; focal group: limited English proficient students
- **Accommodation use**—reference group: students not using testing accommodations; focal group: students using testing accommodations

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 the chi-square.

In addition to the MH chi-square statistic, the delta statistic (MH-D DIF) was computed for all items. The delta statistic was developed by Educational Testing Service (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 (χ_{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 α is assumed to be constant across the k matched levels. Then the χ_{MH}^2 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 (Δ_{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} = the 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 the NAEP:

- Moderate DIF: If the MH statistic is significant ($p < 0.05$) and $|\text{ES}|$ is between 0.17 and 0.25
- Large DIF: If the MH statistic is significant ($p < 0.05$) and $|\text{ES}| \geq 0.25$

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 10-9 show the DIF results for all subgroups of students.

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 10-9 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 results are presented in Table 10-6, socioeconomic status DIF results are shown in Table 10-7, the DIF results based on disability status are presented in Table 10-8, and the DIF results for accommodation use are presented in Table 10-9.

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 MH SMD statistics and

the Mantel-Haenszel statistic (Δ_{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 (Δ_{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 and 1 item was flagged for large (C flag) DIF in ELA. Of these items, 6 were flagged in favor of the focal group (females) and 4 were flagged against the focal group. Seven items were flagged for moderate DIF in Mathematics (2 items in favor and 5 items against female students). Three items were flagged for moderate DIF in Science (1 item in favor and 2 items against female students). In addition, 8 items were flagged for moderate DIF in Social Studies (3 items in favor and 5 items against female students). Overall, 12 items were flagged in favor of the focal group (females) and 16 items were flagged against the focal group across all grades and content areas. Of all items flagged for gender DIF, only one displayed large DIF (in grade 7 ELA) and 27 items displayed moderate DIF.

The other DIF results in Tables 10-2 through 10-9 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 10-8, the item type flagged most often was a TDA item. TDA items were flagged against students with disabilities in all grades.

The Spring 2018 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 Spring 2017 were reviewed by DRC content experts for potential content-related bias. Only items deemed to be free of bias were included in the selection of the Spring 2018 forms. Items flagged for DIF after the Spring 2018 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 2018 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 large 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 2018 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 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 10-10 shows the correlations between main test domains for ELA, and Tables 10-11 through 10-14 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 between two content standards, 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 10-10, the correlations between the ELA main test domains of Reading, Writing, and Listening are moderate to high and range from 0.56 to 0.76 across all grades. The lowest correlations were observed between Listening and Writing domains, while the highest correlations were observed between Reading and Writing domains. The correlations between ELA content standards (see Table 10-11) are typically moderate for all grades and all standard pairs and range from 0.35 to 0.68. It should be noted that the number of items in content standards was smaller than the number of items in ELA domains, resulting in lower correlations at the standard level compared to the correlations at the ELA domain level.

As indicated in Table 10-12, the correlations between Mathematics content standards are moderate to high and range from 0.52 to 0.72. The correlations between Science content standards range from 0.35 to 0.69 (see Table 10-13), and the correlations between Social Studies

content standards range from 0.53 to 0.75 (as shown in Table 10-14). 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 suggests 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. A 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 high level of reading ability, the Mathematics test would be measuring not only mathematics ability but also reading ability. 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 a principal component 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 10-15 show that the eigenvalue of the first factor is at least five times as large as the eigenvalue of the second factor.

As can be seen in Table 10-15, the ratios of the first two eigenvalues range from 5.53 to 8.40. The eigenvalues are proportional to the amount of common variance explained by each component, indicating that the variance explained by the first component alone is approximately 5 to 8 times greater than the variance explained by the second component. The eigenvalue ratios range from 6.23 to 7.59 in ELA, from 6.07 to 7.65 in Mathematics, from 5.53 to 5.92 in Science, and from 6.56 to 8.40 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 relationship between the Wisconsin Forward Exam scores and 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 the 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 2018 were computed and examined for the total student population and by subgroup. Table 10-16 shows the correlations between the content area scores for the total population of Wisconsin students. These correlations ranged from 0.72 (between ELA and Mathematics in grade 7, between Mathematics and Science in grade 4, and between Mathematics and Social Studies in grades 4 and 8) to 0.82 (between ELA and Social Studies in grades 4 and 8).

Tables 10-17 through 10-21 show correlation coefficients between the content area scores by gender, ethnicity, English language proficiency status, socioeconomic status, and disability status, respectively. As seen in Table 10-17, the correlations between the content area scores for male or female groups ranged from 0.71 to 0.83 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 0.57 to 0.83 (see Table 10-18). 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 10-19, the correlations between the content area scores by English proficiency status ranged from 0.45 to 0.81. 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 10-20. These correlations ranged from 0.65 to 0.80 across all grades and pairs of correlated scores. In all grade levels, the correlations between each pair of scores were, in most cases, comparable for the groups of students considered economically disadvantaged and not economically disadvantaged. The correlations between the content area scores by student disability status are shown in Table 10-

21. These correlations ranged from 0.53 to 0.81 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. The correlation coefficients between the content area scores were not computed by accommodation use because the accommodation use status is not consistent across content areas for the same students (for example, students who used accommodations in one content area did not necessarily use accommodations in another content area).

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 one another, 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 socioeconomic status), were also computed. Partial correlations allow for the evaluation of the relationship of 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 10-22. These correlations ranged from 0.61 (between Mathematics and Social Studies in grade 8) to 0.76 (between ELA and Social Studies in grades 4 and 8, and between Social Studies and Science 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 the 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 content 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 (Reading and Mathematics) and Spring 2015 (Science). The Wisconsin Forward Exam state assessment results are compared to the latest available NAEP results 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 10-23. 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 10% or less for any performance level in both grades and in all three content areas. The exceptions were the percentages of students classified in the *Advanced* level for Science, where the differences were over 15% in grade 4 and over 12% 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 grade 4, Mathematics grade 4, and Science (both grades) compared to the corresponding NAEP Reading, Mathematics, and Science assessment. Higher proportions of students were classified in the *Proficient* or above category on the NAEP Reading and Mathematics assessments in grade 8 compared to the Wisconsin Forward Exam in Reading and Mathematics for that grade level. All differences were 10% or less. The same pattern of impact data was observed for students classified as *Basic* or above.

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 student ability. Various activities may take place, such as a student copying from another student's paper, a student receiving inappropriate assistance before or during testing, or a student's 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 MC items on the Spring 2018 Wisconsin Forward Exam in ELA, Mathematics, Science, and Social Studies that were changed to correct responses. We refer to these answer changes as wrong-to-right answer changes. Inordinate numbers of wrong-to-right answer changes in a specifically identifiable testing administration group may indicate inappropriate student behavior or intervention by an educator during the testing session.

The second study evaluates the time spent on the test and individual test items by students. These analyses serve to inform of any events in which students (within one school) spent a 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 abilities.

Standardized test administration, extensive training of the test scorers and artificial intelligence (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. (p. 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. (p. 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 evaluation 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 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	3	1	TDA	.18		B
	3	17	MC	-.09	-1.41	B-
	4	1	TDA	.16		B
	5	15	MC	-.07	-1.15	B-
	5	37	MC	-.08	-1.14	B-
	6	1	TDA	.21		B
	7	1	TDA	.25		C
	7	10	MC	-.09	-1.12	B-
	8	1	TDA	.19		B
	8	9	MC	.07	1.42	B
Math	3	4	MC	-.07	-1.23	B-
	3	10	SA	-.09	-1.31	B-
	3	31	MC	.08	1.04	B
	4	7	MC	-.10	-1.30	B-
	5	9	MC	-.09	-1.17	B-
	6	24	TE	.08	1.15	B
	7	2	MC	-.07	-1.06	B-
Science	4	22	MC	-.03	-1.07	B-
	4	27	MC	-.09	-1.12	B-
	8	20	MC	.08	1.12	B
Social Studies	4	20	MC	.04	1.11	B
	4	26	TE	-.11	-1.29	B-
	8	19	MC	-.08	-1.05	B-
	8	30	MC	.03	1.34	B
	10	4	MC	-.09	-1.06	B-
	10	10	MC	-.10	-1.19	B-
	10	40	MC	.05	1.10	B
10	42	MC	-.10	-1.44	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	37	MC	-.07	-1.13	B-
	6	1	TDA	-.16		B-
	8	1	TDA	-.27		C-
Math	6	24	TE	-.08	-.81	B-
	7	31	MC	-.07	-1.06	B-
	8	16	TE	-.10	-1.76	B-
Science	8	6	MC	-.08	-1.15	B-
	8	8	MC	-.07	-1.04	B-
	8	18	MC	-.08	-1.03	B-
Social Studies	4	4	MC	.10	1.16	B
	4	5	MC	.09	1.21	B
	4	6	TE	-.08	-.96	B-
	4	11	MC	.12	1.31	B
	4	19	MC	-.09	-1.38	B-
	8	15	MC	-.08	-1.16	B-
	8	17	TE	-.09	-1.81	B-
	8	22	TE	-.09	-.95	B-
	8	36	MC	-.11	-1.59	C-
10	9	TE	-.10	-1.47	B-	

Table 10-3 Items Flagged for DIF by Race/Ethnicity, Focal Group: Hispanic

Content	Grade	Item Number	Item Type	MH SMD Statistic	MH Delta Statistic	DIF Flag
ELA	5	5	MC	-.07	-1.04	B-
	5	16	MC	-.08	-1.05	B-
	6	17	EBSR	-.16		B-
Science	8	8	MC	-.07	-1.21	B-
Social Studies	4	2	MC	-.08	-1.32	B-
	8	18	MC	.10	1.22	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	5	MC	-.10	-1.78	C-
	5	12	MC	.13	1.51	C
	5	16	MC	-.07	-1.08	B-
	5	19	MC	-.09	-1.12	B-
	5	31	MC	-.09	-1.03	B-
	5	37	MC	-.10	-1.38	B-
	6	17	EBSR	-.16		B-
	6	18	MC	-.07	-1.05	B-
	7	7	MC	-.07	-1.39	B-
	8	17	MC	-.07	-1.85	C-
Math	5	20	TE	.10	1.40	B
	6	18	MC	-.09	-1.18	B-
	6	34	MC	-.10	-1.33	B-
	7	29	MC	.08	1.18	B
	7	31	MC	-.06	-1.09	B-
	7	42	MC	-.08	-1.10	B-
	8	7	MC	.09	1.02	B
	8	17	MC	-.07	-1.02	B-
Science	8	6	MC	-.05	-1.23	B-
	8	8	MC	-.09	-1.85	C-
Social Studies	4	2	MC	-.11	-2.08	C-
	4	20	MC	.06	1.48	B
	8	4	MC	-.06	-1.16	B-
	8	10	MC	-.07	-1.15	B-
	8	20	MC	.07	1.36	B
	8	30	MC	.03	1.37	B
	8	36	MC	-.08	-1.30	B-
	8	39	MC	.05	1.29	B
	10	2	MC	-.08	-1.46	B-
	10	18	MC	-.08	-1.07	B-
	10	20	MC	.10	1.16	B
10	50	MC	.07	1.00	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	4	1	TDA	-.14		B-
	6	1	TDA	-.16		B-
	8	1	TDA	-.25		C-
Science	8	1	MC	-.06	-1.00	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	5	5	MC	-.18	-2.15	C-
	5	16	MC	-.11	-1.32	B-
	6	17	EBSR	-.22		C-
	7	3	TE	-.14		B-
	8	6	TE	-.12		B-
	8	17	MC	-.11	-1.39	B-
	8	24	EBSR	.19		B
Math	5	20	TE	.09	1.18	B
	7	31	MC	-.07	-1.03	B-
Science	8	8	MC	-.11	-1.36	B-
Social Studies	4	2	MC	-.14	-2.00	C-
	8	10	MC	-.09	-1.07	B-
	8	30	MC	.06	1.11	B
	10	2	MC	-.12	-1.38	B-

Table 10-7 Items Flagged for DIF by Socioeconomic Status, Focal Group: Socioeconomically Disadvantaged Students

Content	Grade	Item Number	Item Type	MH SMD Statistic	MH Delta Statistic	DIF Flag
ELA	8	1	TDA	-.20		B-

Table 10-8 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	3	1	TDA	-.17		B-
	4	1	TDA	-.21		C-
	5	1	TDA	-.25		C-
	5	13	TE	-.15		B-
	5	37	MC	.10	1.37	B
	6	1	TDA	-.31		C-
	7	1	TDA	-.35		C-
	8	1	TDA	-.36		C-
	8	6	TE	-.16		C-
Math	7	1	MC	.06	1.19	B
Science	4	17	MC	-.10	-1.44	B-
	4	21	MC	-.08	-1.14	B-
Social Studies	10	7	TE	-.06	-.84	B-

Table 10-9 Items Flagged for DIF by Accommodation Use, Focal Group: Students Using Testing Accommodations

Content	Grade	Item Number	Item Type	MH SMD Statistic	MH Delta Statistic	DIF Flag
Math	3	3	MC	-.10	-1.17	B-
	3	4	MC	-.13	-1.57	C-
	4	12	TE	-.11	-1.20	B-
	4	41	MC	.05	1.08	B
	5	1	SA	-.09	-1.26	B-
	7	1	MC	.07	1.34	B

Note: DIF analysis by accommodation use was not performed on ELA, Science, and Social Studies data due to insufficient number of students using testing accommodations in these content areas.

Table 10-10 Correlations among English Language Arts Test Domains

Grade	ELA Domain	Listening	Reading
3	Reading	0.64	
	Writing	0.61	0.74
4	Reading	0.61	
	Writing	0.56	0.75
5	Reading	0.67	
	Writing	0.64	0.74
6	Reading	0.64	
	Writing	0.59	0.71
7	Reading	0.64	
	Writing	0.62	0.74
8	Reading	0.66	
	Writing	0.63	0.76

Table 10-11 Correlations among English Language Arts Standards

Grade	Standard Code	A	B	C	D	E	F
3	B	0.58					
	C	0.58	0.48				
	D	0.55	0.45	0.46			
	E	0.57	0.46	0.48	0.50		
	F	0.59	0.48	0.52	0.52	0.53	
	G	0.61	0.48	0.50	0.49	0.52	0.51
4	B	0.68					
	C	0.66	0.63				
	D	0.57	0.54	0.51			
	E	0.60	0.57	0.54	0.51		
	F	0.50	0.46	0.44	0.44	0.46	
	G	0.56	0.52	0.52	0.47	0.48	0.41
5	B	0.64					
	C	0.50	0.43				
	D	0.58	0.52	0.37			
	E	0.60	0.53	0.40	0.52		
	F	0.53	0.47	0.35	0.49	0.49	
	G	0.63	0.56	0.45	0.53	0.55	0.50
6	B	0.63					
	C	0.57	0.58				
	D	0.54	0.54	0.50			
	E	0.44	0.45	0.42	0.42		
	F	0.47	0.45	0.43	0.45	0.36	
	G	0.56	0.56	0.53	0.52	0.43	0.42
7	B	0.58					
	C	0.58	0.44				
	D	0.58	0.43	0.47			
	E	0.58	0.43	0.45	0.46		
	F	0.56	0.41	0.48	0.49	0.50	
	G	0.62	0.47	0.48	0.50	0.53	0.50
8	B	0.65					
	C	0.67	0.62				
	D	0.59	0.53	0.57			
	E	0.58	0.53	0.56	0.52		
	F	0.53	0.48	0.53	0.53	0.50	
	G	0.60	0.54	0.59	0.55	0.53	0.49

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 10-12 Correlations among Mathematics Standards

Grade	Standard Code	A	B	C	D	E	F	G	H	I
3	B	0.71								
	C	0.65	0.66							
	D	0.70	0.70	0.66						
	E	0.65	0.65	0.65	0.66					
4	B	0.66								
	C	0.66	0.66							
	D	0.64	0.66	0.71						
	E	0.52	0.54	0.58	0.60					
5	B	0.72								
	C	0.69	0.69							
	D	0.60	0.61	0.64						
	E	0.71	0.68	0.67	0.62					
6	F					0.64				
	G					0.67	0.69			
	H					0.66	0.68	0.71		
	I					0.59	0.59	0.63	0.62	
7	F					0.61				
	G					0.60	0.67			
	H					0.63	0.70	0.68		
	I					0.61	0.69	0.64	0.68	
8	G					0.60				
	H					0.65		0.65		
	I					0.62		0.56	0.64	
	J					0.67		0.60	0.68	0.66

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 10-13 Correlations among Science Standards

Grade	Standard Code	A/B	C	D	E	F
4	C	0.65				
	D	0.41	0.43			
	E	0.45	0.49	0.35		
	F	0.55	0.56	0.40	0.44	
	G/H	0.65	0.66	0.44	0.48	0.59
8	C	0.69				
	D	0.57	0.58			
	E	0.47	0.48	0.43		
	F	0.55	0.57	0.47	0.41	
	G/H	0.60	0.63	0.51	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 10-14 Correlations among Social Studies Standards

Grade	Standard Code	A	B	C	D
4	B	0.67			
	C	0.54	0.58		
	D	0.58	0.62	0.53	
	E	0.62	0.65	0.58	0.56
8	B	0.72			
	C	0.59	0.60		
	D	0.68	0.67	0.55	
	E	0.67	0.67	0.56	0.61
10	B	0.72			
	C	0.72	0.75		
	D	0.65	0.69	0.67	
	E	0.68	0.71	0.71	0.64

Note: Standard Codes are as follows: A = Geography; B = History; C = Political Science and Citizenship; D = Economics; E = The Behavioral Sciences

Table 10-15 Principal Components Analysis

Content Area	Grade	First Eigenvalue	Second Eigenvalue	Ratio of First Two Eigenvalues
ELA	3	7.917	1.213	6.525
	4	8.476	1.209	7.011
	5	8.454	1.274	6.636
	6	7.314	1.173	6.233
	7	8.257	1.210	6.825
	8	8.938	1.178	7.587
Mathematics	3	9.903	1.485	6.668
	4	10.085	1.586	6.360
	5	10.458	1.367	7.649
	6	10.200	1.679	6.075
	7	9.959	1.375	7.242
	8	9.564	1.402	6.820
Science	4	7.525	1.360	5.531
	8	7.951	1.344	5.915
Social Studies	4	8.028	1.224	6.557
	8	8.938	1.198	7.458
	10	10.534	1.254	8.403

Table 10-16 Correlations between Content Area Scale Scores

Grade	ELA & Mathematics	ELA & Science	ELA & Social Studies	Mathematics & Science	Mathematics & Social Studies	Science & Social Studies
3	0.76					
4	0.75	0.81	0.82	0.72	0.72	0.81
5	0.73					
6	0.77					
7	0.72					
8	0.74	0.80	0.82	0.73	0.72	0.81

Table 10-17 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	0.77					
	Male	0.77					
4	Female	0.76	0.82	0.82	0.73	0.72	0.81
	Male	0.76	0.81	0.82	0.72	0.72	0.81
5	Female	0.73					
	Male	0.74					
6	Female	0.77					
	Male	0.77					
7	Female	0.73					
	Male	0.73					
8	Female	0.74	0.81	0.83	0.73	0.72	0.81
	Male	0.75	0.80	0.82	0.72	0.71	0.81

Table 10-18 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	0.73					
	African-American	0.67					
	Hispanic	0.70					
	Asian	0.78					
	American Indian	0.70					
	Two or More	0.76					
4	White	0.72	0.77	0.79	0.68	0.68	0.78
	African-American	0.65	0.74	0.74	0.63	0.62	0.76
	Hispanic	0.70	0.79	0.79	0.68	0.68	0.81
	Asian	0.76	0.82	0.83	0.75	0.75	0.83
	American Indian	0.69	0.79	0.78	0.66	0.64	0.80
	Two or More	0.75	0.81	0.81	0.72	0.73	0.81
5	White	0.71					
	African-American	0.61					
	Hispanic	0.67					
	Asian	0.74					
	American Indian	0.63					
	Two or More	0.72					
6	White	0.74					
	African-American	0.67					
	Hispanic	0.71					
	Asian	0.77					
	American Indian	0.71					
	Two or More	0.76					
7	White	0.70					
	African-American	0.57					
	Hispanic	0.64					
	Asian	0.75					
	American Indian	0.65					
	Two or More	0.72					
8	White	0.73	0.77	0.79	0.70	0.69	0.78
	African-American	0.58	0.74	0.77	0.57	0.57	0.76
	Hispanic	0.68	0.78	0.80	0.66	0.66	0.79
	Asian	0.76	0.82	0.83	0.76	0.73	0.82
	American Indian	0.65	0.80	0.79	0.63	0.62	0.80
	Two or More	0.72	0.77	0.80	0.70	0.70	0.80

Table 10-19 Correlations between Content Area Scale Scores by English Proficiency Status

Grade	Demographic Group	ELA & Math	ELA & Science	ELA & Social Studies	Math & Science	Math & Social Studies	Science & Social Studies
3	Fully English Proficient	0.76					
	Limited English Proficient	0.68					
4	Fully English Proficient	0.75	0.80	0.81	0.72	0.72	0.81
	Limited English Proficient	0.67	0.75	0.76	0.65	0.65	0.79
5	Fully English Proficient	0.73					
	Limited English Proficient	0.58					
6	Fully English Proficient	0.76					
	Limited English Proficient	0.58					
7	Fully English Proficient	0.72					
	Limited English Proficient	0.45					
8	Fully English Proficient	0.74	0.79	0.81	0.72	0.71	0.80
	Limited English Proficient	0.51	0.69	0.68	0.51	0.49	0.69

Table 10-20 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	Not Economically Disadvantaged	0.73					
	Economically Disadvantaged	0.72					
4	Not Economically Disadvantaged	0.72	0.77	0.78	0.68	0.67	0.77
	Economically Disadvantaged	0.71	0.79	0.79	0.69	0.68	0.80
5	Not Economically Disadvantaged	0.71					
	Economically Disadvantaged	0.68					
6	Not Economically Disadvantaged	0.74					
	Economically Disadvantaged	0.72					
7	Not Economically Disadvantaged	0.71					
	Economically Disadvantaged	0.65					
8	Not Economically Disadvantaged	0.72	0.76	0.79	0.70	0.69	0.77
	Economically Disadvantaged	0.68	0.78	0.80	0.67	0.66	0.80

Table 10-21 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	Not Disabled	0.75					
	Disabled	0.72					
4	Not Disabled	0.74	0.80	0.81	0.71	0.71	0.80
	Disabled	0.71	0.78	0.77	0.68	0.68	0.80
5	Not Disabled	0.72					
	Disabled	0.64					
6	Not Disabled	0.75					
	Disabled	0.66					
7	Not Disabled	0.71					
	Disabled	0.53					
8	Not Disabled	0.73	0.78	0.80	0.71	0.70	0.79
	Disabled	0.56	0.73	0.75	0.57	0.56	0.76

Table 10-22 Partial Correlations between Content Area Scale Scores

Grade	ELA & Mathematics	ELA & Science	ELA & Social Studies	Mathematics & Science	Mathematics & Social Studies	Science & Social Studies
3	0.69					
4	0.67	0.75	0.76	0.63	0.63	0.76
5	0.63					
6	0.67					
7	0.62					
8	0.65	0.72	0.76	0.62	0.61	0.73

Table 10-23 Comparison of Most Recent Wisconsin NAEP and Spring 2018 Wisconsin Forward Exam Impact Data

Content	Grade	Wisconsin NAEP Percentages of Students							Wisconsin Forward Exam Spring 2018 Percentages of Students					
		NAEP Year	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
Reading/ ELA	4	2017	34	31	27	8	35	66	24.04	32.06	35.72	8.19	43.91	75.96
Reading/ ELA	8	2017	21	40	35	4	39	79	24.66	38.01	27.93	9.40	37.33	75.34
Math	4	2017	21	37	33	9	42	79	18.37	37.17	32.71	11.74	44.46	81.63
Math	8	2017	24	37	27	12	39	76	27.95	35.44	28.71	7.90	36.61	72.05
Science	4	2015	21	38	40	1	41	79	15.24	34.07	34.43	16.26	50.69	84.76
Science	8	2015	25	35	38	2	40	75	17.18	33.96	34.16	14.70	48.86	82.82

Note: NAEP assessed student knowledge and skills in Reading, while Wisconsin Forward Exam assessed student knowledge and skills in ELA, which included Reading, Listening, and Writing.

Note: NAEP data are from <https://nces.ed.gov/nationsreportcard/assessments>.

Part 11: Summary Recommendations

Results and key findings of the Spring 2018 Wisconsin Forward Exam administration are presented throughout the body of this report. This last section of the report presents some recommendations for DPI consideration.

The 2018 Wisconsin Forward Exam administration was the third administration of the assessment. For three consecutive years, the assessment results were reported on the same scales and students were classified into the proficiency levels using the same cut scores, allowing for longitudinal tracking of student performance. Using the same scales and the same cut scores for Wisconsin assessments allows for monitoring student growth across administration years.

Following the Spring 2016 through 2018 field testing 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 good agreement with scores by trained human scorers.

From the psychometric perspective, it was noticed that the ELA grade 7 and 8 tests are of comparable difficulty, as indicated by the test characteristic curves (refer to Part 6 of this report). In order to achieve better ordinality of the ELA assessments' overall difficulty across grade levels, a few easier items could be added to the grade 7 test or a few more difficult items could be added to the grade 8 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.

In Mathematics grades 4 through 8, more than 2% of students received the lowest obtainable scale score (LOSS). These percentages at the LOSS indicate that the Mathematics assessments were difficult for some students. The response patterns of students at the LOSS in Mathematics indicated that these students typically answered very few MC items and none of the non-MC items. As explained in detail in Part 6 of this report, for these students to receive a scale score above the LOSS, they would need to correctly answer more items, including some non-MC items. Therefore, it is recommended that some easier non-MC items be included in the future forms of Mathematics tests.

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. (December 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 No. 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. (April 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. (April 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 data [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 No. 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 No. 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, NY: 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

Summer 2016 Item Review Training Slides

Wisconsin Forward Exam Item Review

Madison, WI
August 2016




Purpose of Meeting

- Provide overview of the Wisconsin Forward Exam
- Provide item review training
- Review items for potential placement on Wisconsin Forward Exam



Wisconsin Graduates are College and Career **READY**



ALL STUDENTS IN WISCONSIN GRADUATE FROM HIGH SCHOOL ACADEMICALLY PREPARED AND SOCIALLY AND EMOTIONALLY COMPETENT BY POSSESSING AND DEMONSTRATING...

Knowledge
Proficiency in academic content

Skills
Application of knowledge through skills such as critical thinking, communication, collaboration, and creativity

Habits
Behaviors such as perseverance, responsibility, adaptability, and leadership

These proficiencies and attributes come from rigorous, rich, and well-rounded public school experiences.

DRC
DATA RECOGNITION CORPORATION

WISCONSIN DEPARTMENT OF PUBLIC INSTRUCTION
They Enrich, Push, Inspire Superintendents

WISCONSIN DEPARTMENT OF PUBLIC INSTRUCTION

Wisconsin's Definition of College and Career Readiness

Wisconsin Forward Exam

- Grades 3–8 for English Language Arts and Mathematics
- Grades 4, 8, and 10 for Social Studies
 - Science Grades 4 and 8 (Review in October)
- All items written are aligned to Wisconsin Academic Standards

DRC
DATA RECOGNITION CORPORATION

WISCONSIN DEPARTMENT OF PUBLIC INSTRUCTION

Security and Confidentiality

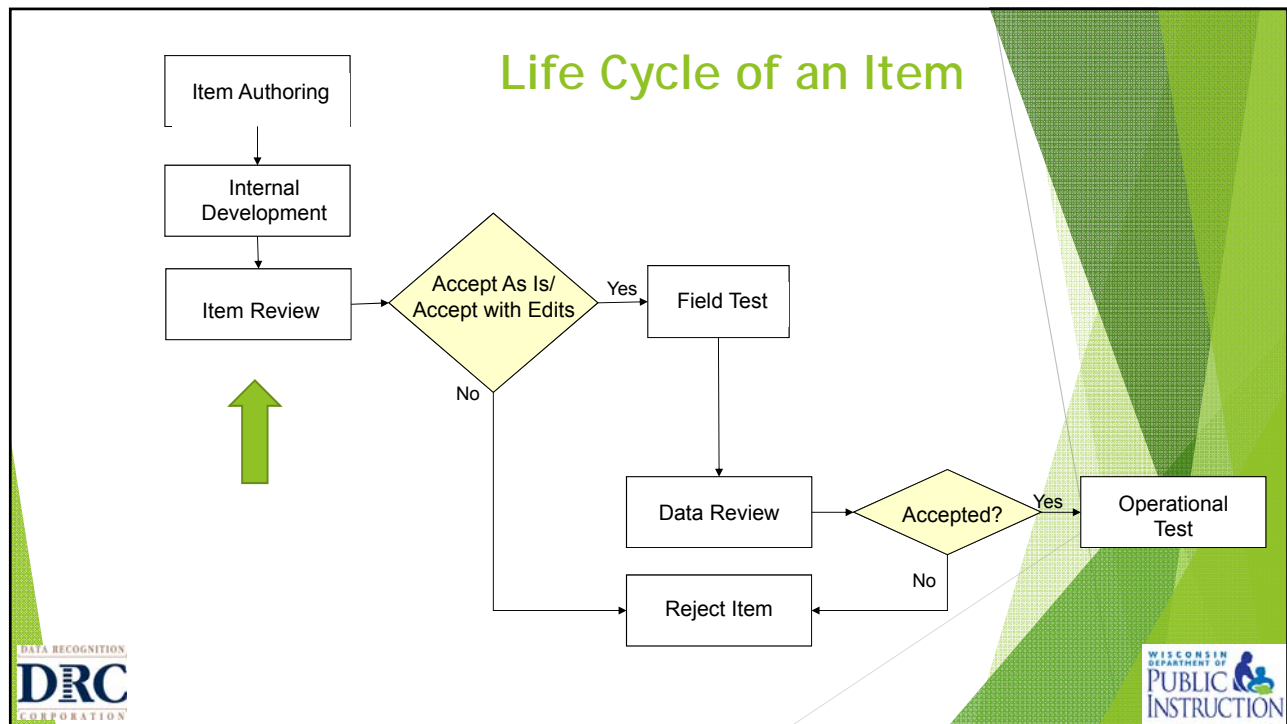
- Critical Importance of Security
 - Security/Nondisclosure Agreement
 - Security of passage and item content
 - Note-taking policy
 - Cell phone and personal computer use
 - Communication following the meeting



Item Review Process-NEW!

- Reviews will be completed online using the same testing engine students use-INSIGHT
 - Allows for reviewer interaction with item functionality, particularly technology-enhanced items
 - Facilitator will provide specific directions for logging in to begin reviews





Item Types

- Selected Response (SR)
 - Multiple Choice (MC)
 - Enhanced Selected Response (ESR)
 - Evidence-Based Selected Response (EBSR)
- Scorable Equation/Numeric (SEQ)
- Text Dependent Analysis (TDA)
- Technology Enhanced (TE)

Selected-Response Item Type- Multiple Choice (MC)

- All MC items have 4 answer choices
 - 3 distractors and 1 correct answer
- Used in all content areas
- Can be linked to a passage or stimuli or used as a “stand-alone MC”
- May have graphs, tables, or other information to support the stem



MC Sample

An increase in smartphone sales would most likely result in

- (a) lower salaries for workers at smartphone companies
- (b) fewer resources being used by the smartphone industry
- (c) higher earnings for stockholders in smartphone companies
- (d) higher unemployment rates for workers in the smartphone industry



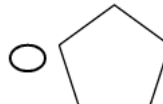
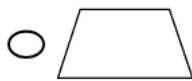
Selected-Response Item Type- ESR

- Varying combinations of multiple choice, multiple response, completion or short answer
- Explores authentic problem-solving skills
- Multi-part, auto scored



ESR Sample

Select all the shapes that are quadrilaterals but **not** rectangles.



Selected Response Item Type-EBSR

- 2-part item
 - Part A-Accuracy portion; single correct answer
 - Part B-Evidence portion; one or more correct answers based upon Part A
- 2-point item; student may get 0, 1, or 2 points



EBSR Sample

This question has two parts. First, answer part A. Then, answer part B.

Part A

What is the main way the passage “Public Transportation, Not for Everyone” supports the claim that taking public transportation may be problematic for some people?

Part B

Which sentence from the passage **best** supports your answer in part A?



Scorable Equation/Numeric Item Type-SEQ

- Used in Mathematics Items
- Grade-level specific keypad that allows for more guided input of student responses

The image shows two keypad interfaces side-by-side. The left keypad is labeled "Grades 3-5 'numeric' keypad" and features a standard numeric keypad with buttons for digits 0-9, a decimal point, and a fraction button. The right keypad is labeled "Grades 6-8 'numeric' keypad (with fraction button)" and includes an additional button for a negative sign (-). Both keypads are part of a larger interface with navigation arrows and a question mark icon.

SEQ Sample

A rectangular section of a kitchen wall will be tiled.  What is the area, in square feet, of the section of wall that will be tiled?

The image shows the Grade 5 keypad interface from the previous slide. A green arrow points to the input field above the keypad, labeled "Student Response Area". Another green arrow points to the keypad itself, labeled "Grade 5 Keypad".

Text Dependent Analysis (TDA)

- Used in ELA assessment
- Based on a passage
- Used for both literature and informational texts
- Basic writing skills used while inferring and synthesizing information from the passage
- Scored using a holistic scoring guide
- Character counter feature



TDA Sample

Both passages focus on creatures from two different species helping each other. Write a response explaining how both passages show ways in which people and animals help each other. Use evidence from **both** passages to support your response.

A large, empty rectangular box with a thin black border, intended for the student's written response. The box is positioned below the prompt text.

0/5000



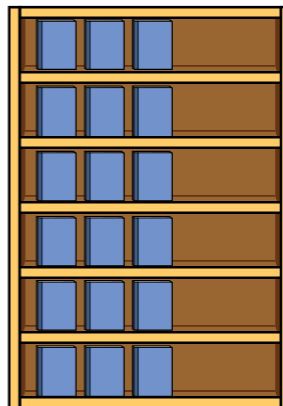
Technology Enhanced (TE)

- TE items present in all content areas
- Interactive
- Wide Variety: clock input, angle draw, drop down list, matching, graphing, highlighting text, drag and drop



TE Sample Item

Clayton puts 18 books in a bookcase. He puts the same number of books on each shelf. Move groups of books to the bookcase to show how Clayton could arrange them.



Webb's Depth-of-Knowledge (DOK) Levels



Definition of DOK

the degree or complexity of knowledge that the content curriculum standards and expectations require

- Includes four levels, from lowest (basic recall) to highest (extended thinking)
- Focuses on how well the students need to know the content before they can respond to a given item
- Used by item writers to gauge the *cognitive level* of item, does not correlate to the *difficulty* of the item



DOK Levels

DOK 1 Recall and Reproduction

DOK 2 Skills and Concepts

DOK 3 Strategic Thinking and Reasoning

DOK 4 Extended Thinking

(rarely on standardized assessments — more “project-like” or on performance assessments)



DOK 1: Recall and Reproduction

- Students demonstrate a rote response, use a well-known formula, or follow a simple procedure.
- A “simple” procedure is well defined and typically involves only one step.

Key Words: identify, recall, recognize facts, use, measure, solve a one-step problem



DOK 2: Skills and Concepts

- Students make some decisions regarding how to approach the question or problem.
- This level requires deeper knowledge than just giving a definition, such as explaining how or why; it may involve two or more steps.

Key Words: explain, categorize, use context clues, select a procedure, compare/contrast



DOK 2-(cont.)

Activities may include:

- Making observations/collecting information
- Classifying/comparing information
- Organizing/displaying data or information in tables and graphs

Note: Some action verbs, such as “explain,” “describe,” or “interpret,” could be classified at different DOK levels, depending on the complexity of the action.



DOK 3: Strategic Thinking and Reasoning

- Students demonstrate deep understanding through planning, using evidence, and exhibiting higher levels of cognitive reasoning.

Key Words: connect ideas, explain thinking, cite evidence, analyze, apply a concept,



DOK 3-(cont.)

Activities may include the following:

- Use concepts to solve non-routine problems
- Describe how word choice, point of view or bias, may help the readers' interpretation of text
- Apply a concept in a new context
- Cite evidence and develop a logical argument for concepts
- Compare information within or across data sets



DOK 4: Extended Thinking

- Students demonstrate an integrated use of higher order thinking processes such as critical and creative and productive thinking, reflection, and adjustment of plans.

Key words: analyze, synthesize, examine and explain, describe and illustrate common themes



DOK 4- cont.

- Higher order thinking skills

Activities may include the following:

- Developing generalizations
- Analyzing abstract themes
- Evaluating relevancy, accuracy, and completeness of information from multiple sources

Key words: analyze, synthesize, examine and explain, describe and illustrate common themes



Sample of Difficult DOK 1 Item

Which of these events occurred in the twentieth century?

- (a) Germans began migrating to Wisconsin.
- (b) Madison was chosen to be the capital of Wisconsin.
- (c) Wisconsin senator Joseph McCarthy became a controversial national figure.
- (d) Father Jacques Marquette and Louis Joliet traced the Wisconsin and Mississippi Rivers.

- DOK used by item writers to gauge the *cognitive level* of item, does not correlate to the *difficulty* of the item



Sample of an Easy DOK 3 Item

The area model below is used to solve a multiplication problem.

20×10	3×10
20×7	3×7

$$200 + 30 + 140 + 21 = 391$$

Fill in the multiplication problem that goes with this area model.

$$\boxed{} \times \boxed{} = 391$$

- DOK used by item writers to gauge the *cognitive level* of item, does not correlate to the *difficulty* of the item



Item Review Process

- Reviews will be completed in groups and individually
- Items will be reviewed for:
 - Standard alignment
 - Grade-level appropriateness
 - Correct Key(s)
 - Rigor-level alignment
 - DOK level and estimated item difficulty
 - Bias and sensitivity concerns



Item Review Tally Sheet

Session Number	Sequence #	Item ID	Passage Title	Standard	Item Type	Key(s)	DOK	Bias/Sensitivity Comments	Accept (A); /Accept with Revisions A (AR); Dissenting View (DV)	Comments



Evaluating an Item: Grade 3 Writing

WBTE Preview 697075 // Albert Einstein

Question 2 Item ID ?

Read the book titles. Choose the **two** book titles that have errors in capitalization.

Who Was Walt Disney?

The Twin Toddlers Turn Two

Harold and the Purple Crayon

Science Experiments For Kids

The Mystery of the Lost Backpack

The Best Day I ever Had in my Life

Review/End Test Pause Flag Options Back Next

Step 1: Standard Alignment

After reading item ask yourself:

Does the standard listed match the state standard?

- Each member will have copy of standards
- Match item to appropriate standard as noted on item rating sheet
- Indicate agreement of alignment on item rating sheet or recommend new standard

Step 2: Check the Keys

- Is the key (or keys) listed correct?
 - If yes, move on to step 3
 - If no, discuss with committee and note new key(s)



Step 3: Confirm the DOK Levels

- Is the DOK level listed correct?
 - If yes, move on to step 4.
 - If no, mark your thinking and discuss with committee.



Step 4: Check for Bias and Sensitivity

- Stereotyping
- Gender
- Regional or geographical
- Ethnic or cultural
- Socioeconomic class
- Persons with a disability
- Ageism
- Religious



Also Keep in Mind...Technical Design

- Does the item meet requirements for technical quality?
 - Stem: Complete question/problem; does not clue correct answer(s)
 - Correct answer(s): clear and accurate
 - Distractors (or incorrect options): may contain common misperceptions or processes
 - Graphics/visuals: compliment and support item



Be mindful of Principles of Universal Design

- Items should respect the diversity of the assessment population.
- Items should have a clear format for text.
- Items should measure what is intended.
- Stimuli and items should have clear pictures and graphics.



Principles of Universal Design (cont.)

- Items should have concise and readable text.
- Items should be written to provide for a test that will have an overall appearance that is clean and organized.



Everything in Moderation



Steps 5 and 6: Mark Comments

In document, mark column noting the following:

- **Accept- "A"**
 - Item is OK as is
- **Accept with Revisions- "AR"**
 - Accept but apply recommended edits
- **Dissenting View- "DV"**
 - Item contains major flaws; do not recommend placement on assessment
- **Additional comments as needed**

Session Number	Sequence #	Item ID	Passage Title	Standard	Item Type	Key(s)	DOK	Bias/Sensitivity Comments	Accept (A); /Accept with Revisions A (AR); Dissenting View (DV)	Comments

Main Question to Ask During Review

- Does the item provide for an optimal standard assessment of all students?

When to Edit an Item

- If the subject matter is above grade level or out of scope for the standard/course.
- If there is an opportunity to make the item/passage/stimulus easier for students to understand.
- If the topic or language is inappropriate, controversial, or inflammatory.



What if I Disagree with the Committee?

- Speak up! It's possible that another committee member has the same concern or you may have noticed something that other committee members have not.
- Record your dissenting view on the item review tracking sheet. Discussion by all is encouraged however, if you choose not to share your opinion, your facilitator can voice your concern for you.
- DRC and DPI will reconcile any major disagreements/concerns noted on tracking sheet following the meeting.



Item Review Process: Summary

- Standard Alignment
- Key(s)
- DOK Levels
- Grade-level Appropriateness
- Bias and Sensitivity



Roles & Responsibilities

- ▶ Participants
 - ▶ Item Review
- ▶ DRC Facilitators
 - ▶ Lead the group through the agenda
 - ▶ Encourage interaction
 - ▶ Lead discussions
 - ▶ Collect secure materials
- ▶ DPI and DRC
 - ▶ Answers to questions



Roles & Responsibilities

- Educators
 - Invest yourself in the process
 - Share your opinions
 - Listen to your colleagues



Questions?



Appendix B

Spring 2017 Field Test Data Review Training Slides

Wisconsin Forward Exam Item Data Review

Wisconsin Department of Public Instruction
&
Data Recognition Corporation

1

Purpose

- Establish a robust pool of items for use in new test development to ensure proper representation.
 - Content standards
 - Test design
- General statistical guidelines are presented.
 - No item flags are created equal
 - Guidelines vs. hard-and-fast rules
 - Item content needs to be considered as well.
 - Approving an item does not guarantee its appearance on a future test, but rather maximizes the size of the pool for item selection during test development.

2

Key Objectives

- Review item development process.
- Review and understand item card layout.
- Understand and interpret item statistics.
- Review item cards for a few Wisconsin items with different statistics.
- Apply knowledge of item statistics to evaluate the remaining items.

3

Some Definitions

- **Item pool**: Set of items on a given test scale that are available for operational test construction
- **Item statistics**: Statistical values generated during data analysis after item administration (more detail later)
- **Field tested items**: Items that have been embedded among operational items to gather item statistics before placing them on the operational tests
- **Operational items**: Items that have already been used in an operational test administration

4

Some More Definitions

- **Item type:** Refers to the format of the item, e.g.
 - Multiple-choice (MC)
 - Technology-enhanced (TE)
 - Multi-select (MS)
 - Short answers (SA)
 - Evidence-based selected response (EBSR)
 - Text Dependent Analysis (TDA)

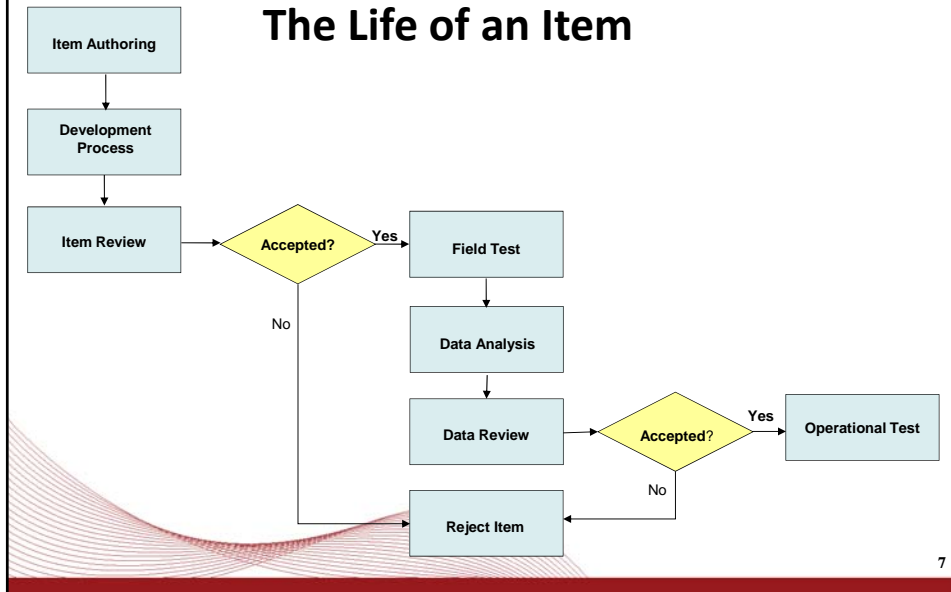
5

...and a few More Definitions

- **Item Scoring:** refers to the score range (not item type)
 - **Dichotomous:** Item is scored as 0 or 1
 - **Polytomous:** Item has a range of possible scores from 0 to greater than 1

6

The Life of an Item



7

Sample Item Card

1. The distances Esteban runs during the first 7 days form a pattern. The pattern starts with 1.2 miles. The table below shows the mileage he runs each day.

Day	Miles Run
Monday	1.2
Tuesday	1.6
Wednesday	2.0
Thursday	2.4
Friday	2.8
Saturday	3.2
Sunday	3.6

Item Stem

Content Area

Item Grade

Content Standard

Max Points

Key

Item Options

What is the rule of the pattern displayed in the table?

- The pattern increases by 4 miles every day.
- The pattern is skip counting by 4.
- The pattern increases by $\frac{4}{10}$ of a mile every day.
- The pattern increases by 1.4 miles each day.

Administration(s)									
Form Name	Use Function	Seq	Period	Year	Session	Calc	Model/Ext	Grade	
GS MA1	PT	51	Spring	2017	3	No	SPL/SPL	5	
Item ID									
Item ID	N	P-Val	Mean	Item Total Corr					
2000	0.44		0.45						
Fit Statistics									
Outfit I	Infit I	Outfit MeanQ	Infit MeanQ	Chi-sq	Deg Free	Item Fit	FI		
0.000						60.43			
IRT Statistics									
Label	Final	Final S.E.	Preliminary	Preliminary S.E.	Depth				
Slope	2.26								
Location	0.49								
Asymptote	0.21								
Distractor/Step Specific									
Label	Percent	Corr	Avg Mean	>Step Mean					
A	0.34	-0.19							
B	0.16	-0.19							
C	0.44	0.45							
D	0.13	-0.20							
OMTS	0.00								
DIF Analysis									
Category	Stat Code	Item Value	N - Ref	N - Focal					
ACC	A	0.16	24993	2382					
MALE/FEMALE	B-	-1.22	13263	13863					
WHITE/AMN			17974	102					
WHITE/ASIAN	A	-0.36	19525	622					
WHITE/BLACK	A-	-0.68	19444	5806					
WHITE/HISPANIC	A-	-0.68	19467	3059					
WHITE/MULTI	A	-0.35	19444	982					

Admin Info

Classical Stats

Item Fit

Distractor or Score Point Stats

DIF Index

1. The distances Esteban runs during the first 7 days form a pattern. The pattern starts with 1.2 miles. The table below shows the mileage he runs each day.

Day	Miles Run
Monday	1.2
Tuesday	1.6
Wednesday	2.0
Thursday	2.4
Friday	2.8
Saturday	3.2
Sunday	3.6

What is the rule of the pattern displayed in the table?

A. The pattern increases by 4 miles every day.
 B. The pattern is skip counting by 4.
 C. The pattern increases by $\frac{4}{10}$ of a mile every day.
 D. The pattern increases by 1.4 miles each day.

Item ID
Grade
Standard
Key(s)

Item ID
851887
Content Area
Mathematics
Passage ID
122959
Passage Title
Training
Grade
5
Standards
MLS2016: 5.RA.A.2
Item Type
Multiple Choice
Points
1
Key
C
Calculator
No
Previous Use

9

Administration(s)

Form Name	Use Function	Seq	Period	Year	Session	Calc	Model/Ext	Grade
GG MA1	FT	51	Spring	2017	3	No	3PL/SPL	5

Traditional Statistics

N	P-Val	Mean	Item Total Corr
29080	0.44		0.45

Fit Statistics

Outfit I	Infit I	Outfit MnSq	Infit MnSq	Chi-sq	Deg Free	Item Fit	FR
						60.43	

IRT Statistics

Label	Final	Final S.E.	Preliminary	Preliminary S.E.	Depth
Slope	2.26				
Location	0.49				
Asymptote	0.21				

Distractor/Step Specific

Label	Percent	Corr	Avg Meas	>Step Meas
A	0.24	-0.19		
B	0.18	-0.19		
C	0.44	0.45		
D	0.13	-0.20		
OMITS	0.00			

DIF Analysis

Category	Bias Code	Num Value	N - Ref	N - Focal
ACC	A	0.18	24593	2282
MALE/FEMALE	B-	-1.22	15233	13863
WHITE/AMN			17574	102
WHITE/ASIAN	A	-0.36	19525	822
WHITE/BLACK	A-	-0.88	19444	5806
WHITE/HISPANIC	A-	-0.68	19487	2058
WHITE/MULTI	A	-0.35	19444	982

Admin Info
Classical Stats
Distractor or Score Point Stats
DIF Index

10

Item Statistics



Difficulty

P-Value

- Proportion of students who answered item correctly (or mean as the percent of maximum points possible for polytomously score items)

- Ranges from 0.0 to 1.0

Mean for polytomously scored items

- Average score obtained by the students

- Ranges from 0 to max points possible

11

11

Item Statistics



Discrimination

Item Total Correlation

- Correlation of examinee raw scores on a single item with their raw scores on all remaining test items (-1.0 to +1.0)

- Measures item's ability to **differentiate** between **high** and **low** achievers

12

Item Statistics



Distractors/
Score Points

- **Proportions** and **correlations** for incorrect response options for MC items
- **Proportions** and **correlations** for each score point

13

More Item Statistics



DIF

- **Differential Item Functioning (DIF)**
 - Statistical analysis to determine if test items are *potentially* unfair or inappropriate for subgroups of interest (e.g., gender, ethnicity, and using testing accommodations)

14

Item Difficulty



- **“P-Val”** for MC items
 - Proportion of students who answered an item correctly (or mean as a percent of maximum points possible for polytomously scored items)
 - 0.0 means all students answered incorrectly
 - 1.0 means all students answered correctly
- **“Mean”** for polytomously scored items
 - Average score obtained by the students

Dichotomously Scored Item

Traditional Statistics

N	P-Val	Mean	Item Total Corr
4349	0.73		0.49

Polytomously Scored Item

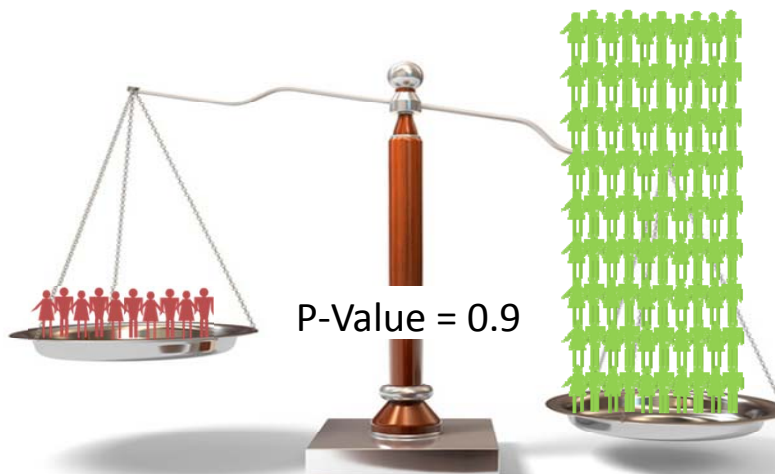
Traditional Statistics

N	P-Val	Mean	Item Total Corr
2008	0.25	1.01	0.65

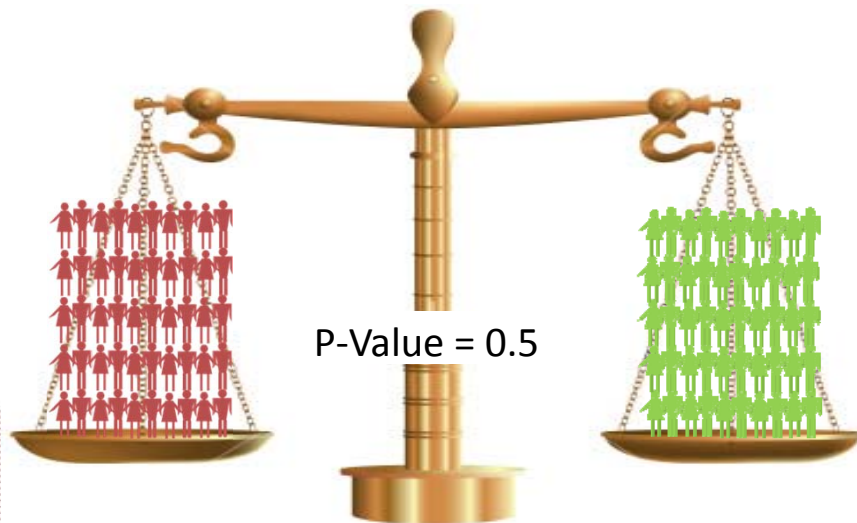
15

15

Visualizing P-Values



Visualizing P-Values



17

P-Values on Item Card

- (*) indicates key
- Other answer options (distractors)
 - Proportions selecting each distractor are also displayed.
 - MULTS means multiple marks.
 - OMITS means student omitted.

Traditional Statistics

N	P-Val	Mean	Item Total Corr
14016	0.78		0.48

Distractor/Step Specific

Label	Proportion	Corr	Avg Meas	Step Meas
A*	0.78	0.48		
B	0.06	-0.29		
C	0.11	-0.25		
D	0.04	-0.23		
MULTS	0.00			
OMITS	0.00			

18

Polytomously Scored Items

- ELA only
 - 2 point EBSR some TE
 - 4 point TDA (operational items)

Polytomously Scored Items

- The mean gives a general idea of item difficulty but can sometimes be deceptive.

Proportion of students			
Score 0	Score 1	Score 2	Item Mean
0.40	0.20	0.40	1
0.15	0.70	0.15	1
0.33	0.33	0.33	1

- Use the score point proportions to determine if the distribution is reasonable.
- We want some students in all score-point categories.
 - If there are no students in every category, the score levels will be collapsed and item parameters will not be estimated for the category with no or very few students.

Guidelines for Polytomously Scored Items

- For a 2-point item, a mean of 1.8 and above may be too easy, and a mean of 0.2 and below may be too difficult.

21

Item Cards for Polytomously Scored Items

Traditional Statistics

N	P-Val	Mean	Item Total Corr
1101	0.25	1.00	0.57

Distractor/Step Specific

Label	Proportion	Corr	Avg Meas	Step Meas
0	0.50	-0.51		
1	0.22	0.04		
2	0.12	0.26		
3	0.11	0.29		
4	0.05	0.28		
BL	0.01			

- **Mean**
 - Average student score on that item
- **Item Total Corr**
 - Item-Total Correlation
- **Proportion**
 - Percent of students receiving a certain score point

22

P-Values: Summary

Theoretical Range

P-Value: 0 to 1

- 0 = no students answered item correctly
- 1 = all students answered item correctly
- Lower values = more difficult (hard)
- Higher values = less difficult (easy)

23

23

P-Values: Summary (cont.)

Targeted Range

- P-Value: 0.20 to 0.90
- Items below 0.20 may be approved if content is appropriate.
- Items above 0.90 may be approved if content is appropriate.

24

P-Values: Summary (cont.)

Content Consideration

- We need to build tests with a wide range of p-values (generally 0.20 – 0.90) in order to effectively place students into the four performance categories.
 - Hard items to distinguish between Proficient/Advanced
 - Easy items to distinguish between Below Basic/Basic
- Why did most students answer this item correctly or incorrectly?
- Are there any reasons other than item difficulty to support a decision to ACCEPT or REJECT this item?

25

Item Discrimination

Discrimination

- Measures item's ability to differentiate between high and low performers
- Item-Total Test Correlation: Correlation of examinee raw scores on a single item with their raw scores on all remaining test items (-1.0 to +1.0)
 - Positive—high achievers outperformed low achievers (targeted).
 - Negative—low achievers outperformed high achievers (unexpected).
 - Around zero—high and low achievers performed about the same on an item (not desired).

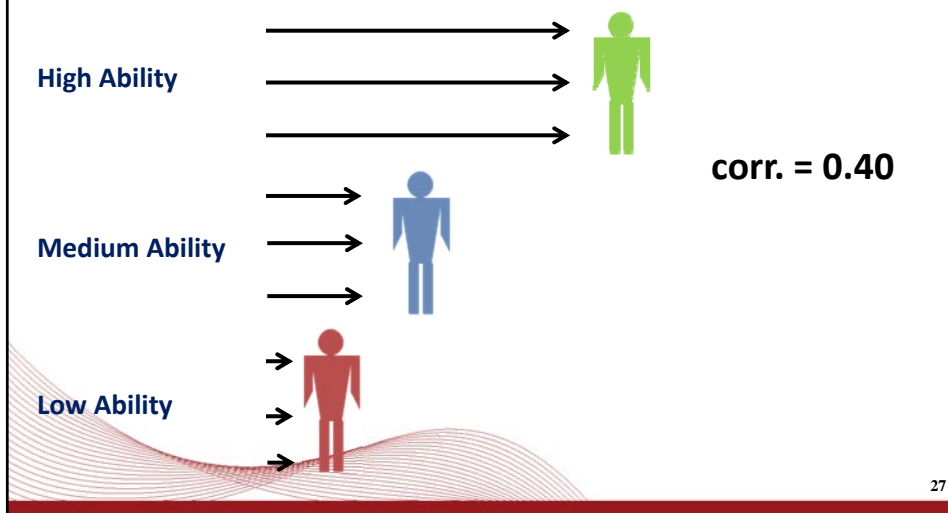
Traditional Statistics

N	P-Val	Mean	Item Total Corr
4349	0.73		0.49

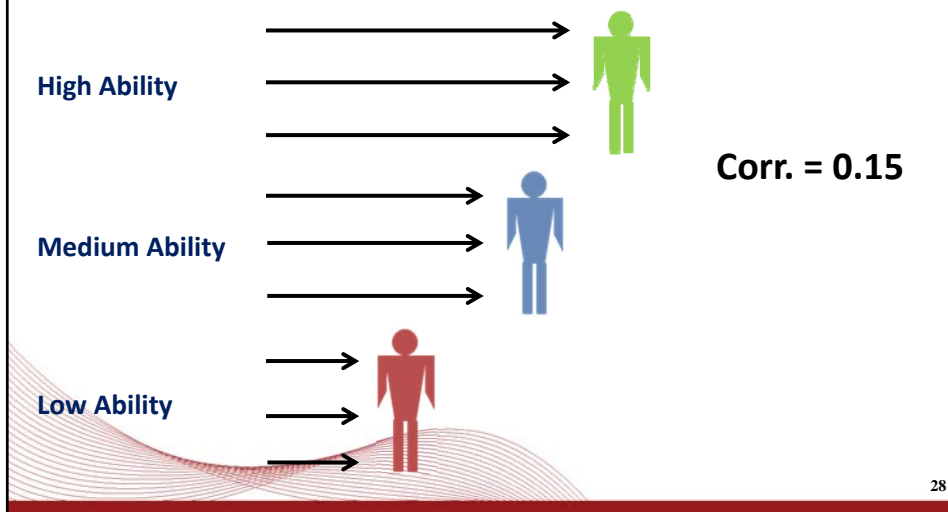
26

26

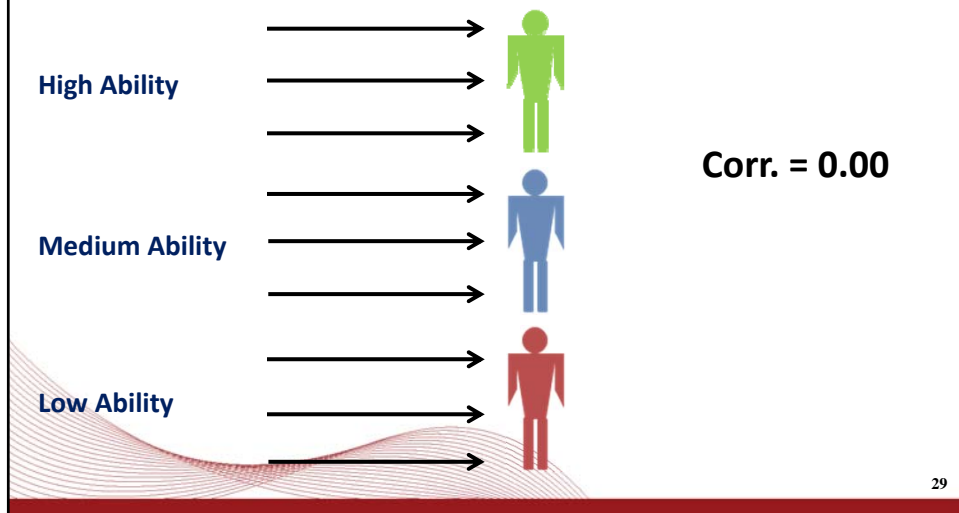
Visualizing Item-Total Test Correlation



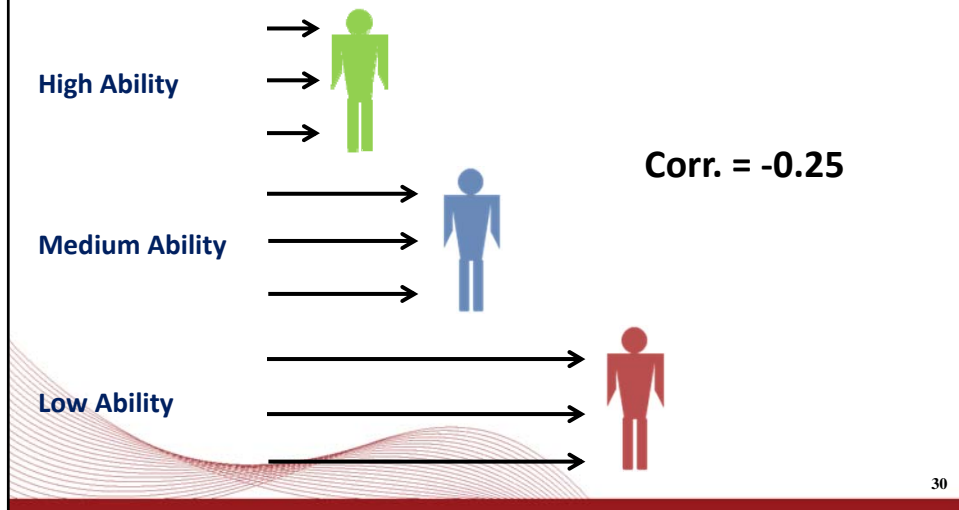
Visualizing Item-Total Test Correlation



Visualizing Item-Total Test Correlation



Visualizing Item-Total Test Correlation



Item-Total Test Correlation on Item Card

- (*) indicates key

Traditional Statistics

N	P-Val	Mean	Item Total Corr
14016	0.78		0.48

Distractor/Step Specific

Label	Proportion	Corr	Avg Meas	Step Meas
A*	0.78	0.48		
B	0.06	-0.29		
C	0.11	-0.25		
D	0.04	-0.23		
MULTS	0.00			
OMITS	0.00			

31

Item-Total Test Correlation: Summary



- Ranges from -1 to +1
- -1 = “perfect” negative relationship
- 0 = no linear relationship
- +1 = “perfect” positive relationship

32

Item-Total Test Correlation: Summary



Targeted Range

- at or above 0.15
- Smaller sometimes is okay, depending on difficulty.
- Items with negative or around 0.0 item total correlations are very poor items that often should be rejected.

33

Item-Total Test Correlation: Summary



Content Consideration

- 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)?
- Are there any other reasons other than item discrimination to support your decision on ACCEPTING or REJECTING this item?

34

Distractor-Specific Analysis (MC Items)

Distractor/Step Specific

Label	Proportion	Corr	Avg Meas	Threshold
A	0.05	-0.22		
B	0.10	-0.26		
C	0.12	-0.28		
D*	0.73	0.49		
MULTS	0.00			
OMITS	0.00			

Guideline

•MC items:

- Correlations for the distractors should be negative.
- Correlations for the distractors should never be higher than correlation for the correct answer.
- Proportion of distractor < proportion of key

35

35

Distractor-Specific Analysis (MC Items)

Content Consideration

- Is the correlation of selecting any incorrect option greater than 0? If yes, why does this option distract more high achievers than low achievers?
- Is the proportion of selecting any incorrect option greater than the proportion of selecting the key? If yes, why?

36

Score Point-Specific Analysis

Distractor/Step Specific

Label	Percent	Corr	Avg Meas	>Step Meas
0	0.39	-0.46		
1	0.09	-0.09		
2	0.52	0.51		
BL	0.00			

Guideline

- Non-MC items:
 - Correlations for the score 0 expected to be negative
 - Correlation for higher scores should be positive
- Proportion for each each score point ≥ 0.05 – desirable property

Content Consideration

Non-MC items

- Is the proportion to a score point < 0.05 ? If yes, is there a reason that explains why so few students received this score point?
- Is the pattern of item score correlation as expected?

37

37

Differential Item Functioning

DIF

- Procedure used to identify items that function differently for particular groups of students (e.g., gender, ethnicity, and disability status, SES status, and LEP status).
- Hypothesis is that test takers with similar knowledge or ability should perform in similar ways on a test item.
- Items are flagged if they do not behave the same in different groups of students, after controlling for student ability.

Procedure

- Compares “focal” vs. “reference” groups.
- Reference groups: Males, Whites, students without disabilities, students who are not SES-disadvantaged, and students who are fully English proficient
- Focal groups: Females, non-White ethnic groups, students with disabilities, SES-disadvantaged students, and LEP students

38

38

Differential Item Functioning

Guideline

- Each item is assigned a bias code of A, B, or C.
 - A – minor DIF (no DIF)
 - B – moderate DIF
 - C – Large DIF

DIF signs: “-” favors Reference group; ‘+’ favors Focal group.

- Only items with C (i.e., large) DIF require review. Items with C DIF may be acceptable if no potential bias causes the differential item functioning.

Content Consideration

- 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?
 - Disability status, SES status, or LEP status?

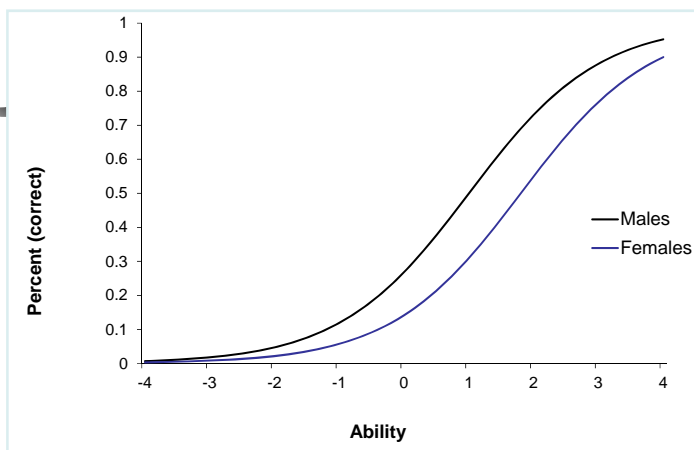
An Index for DIF

DIF Analysis

Category	Bias Code	Num Value	N - Ref	N - Focal
DISAB	A	-0.02	56644	7107
ECODISAD	A-	-0.04	37424	26416
LEP	A-	-0.05	58028	5718
MALEFEMALE	A-	-0.05	32508	31334
WHITEAMIN	A	0.01	41397	766
WHITEASIAN	A-	-0.09	42293	2557
WHITEBLACK	A	-0.01	42225	7116
WHITEHISPANIC	A	-0.02	42290	8531
WHITEMULTI	A	-0.02	42282	2566

Reference Group/Focal Group

Visualizing DIF (Gender)



41

41

DIF: Summary

- All biased items should show DIF, but **Not** all items with DIF will be biased.
 - The smaller sample sizes of the minority ethnicity groups causes many false positives.
 - DIF not computed if focal group $N < 200$.
 - You **must** be able to provide a reason for the bias to call the item biased.



42

Summary of Item Flags



- P-value less than 0.20 or higher than 0.90
- Item-total test correlation < 0.15
 - Negative or close to 0 item-total test correlation is a very serious flag, especially when combined with a positive correlation for a distractor for MC items.
- Positive pt. biserial correlation for a distractor
 - Especially if pt. biserial for a distractor is higher than pt. biserial for the correct option
- Fewer than 5% of students at each score point for non-MC items
 - No students at any of the score points leads to collapsed levels.
- Large DIF (C +/-)
- Omit rates > 5% (not used in this data review; no items were flagged)

43

Roles, Responsibilities, Questions



- DPI
 - Review Item Data.
 - Accept, Revise for re-field testing or reject items.
- DRC
 - Facilitate Data Review.
 - Answer DPI questions.
- Questions?

44

Appendix C

Spring 2018 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	Depth of Knowledge	Standard	Domain
3	1	1	TDA	OP	4	3	3.W.2	Writing
3	2	2	MC	OP	1	2	3.W.1c	Writing
3	2	3	MC	OP	1	2	3.W.3b	Writing
3	2	4	TE	OP	2	2	3.W.1a	Writing
3	2	5	MC	OP	1	2	3.W.8	Writing
3	2	6	MC	OP	1	2	3.W.8	Writing
3	2	7	MC	OP	1	2	3.W.8	Writing
3	2	8	TE	OP	2	2	3.W.8	Writing
3	2	9	MC	OP	1	2	3.W.8	Writing
3	2	10	MC	OP	1	2	3.L.1g	Writing
3	2	11	MC	OP	1	2	3.L.2d	Writing
3	2	12	MC	OP	1	2	3.L.2a	Writing
3	2	13	MC	OP	1	2	3.L.1d	Writing
3	2	14	TE	OP	2	2	3.L.1h	Writing
3	3	15	MC	OP	1	1	3.SL.3	Listening
3	3	16	TE	OP	2	2	3.SL.3	Listening
3	3	17	MC	OP	1	2	3.SL.3	Listening
3	3	18	EBSR	OP	2	3	3.SL.2	Listening
3	3	19	MC	OP	1	1	3.SL.3	Listening
3	4	20	MC	OP	1	3	3.L.4	Reading
3	4	21	TE	OP	1	2	3.L.4	Reading
3	4	22	MC	OP	1	1	3.RI.1	Reading
3	4	23	MC	OP	1	2	3.RI.2	Reading
3	4	24	EBSR	OP	2	3	3.RI.6	Reading
3	4	25	MC	OP	1	3	3.RL.7	Reading
3	4	26	MC	OP	1	2	3.RL.4	Reading
3	4	27	MC	OP	1	3	3.RL.6	Reading
3	4	28	TE	OP	2	2	3.RL.2	Reading
3	4	29	MC	OP	1	2	3.L.4	Reading
3	4	30	MC	OP	1	2	3.RI.5	Reading
3	4	31	MC	OP	1	2	3.RI.1	Reading
3	4	32	MC	OP	1	2	3.RI.7	Reading

Note: TDA item is weighted x 2 in computation of student scores.

Table C-1. English Language Arts, Grade 3 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Depth of Knowledge	Standard	Domain
3	4	33	EBSR	OP	2	2	3.RI.2	Reading
3	4	34	TE	OP	2	2	3.RL.1	Reading
3	4	35	MC	OP	1	3	3.RL.5	Reading
3	4	36	MC	OP	1	2	3.RL.3	Reading
3	4	37	MC	OP	1	3	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	Depth of Knowledge	Standard	Domain
4	1	1	TDA	OP	4	3	4.W.2	Writing
4	2	2	MC	OP	1	2	4.W.1a	Writing
4	2	3	TE	OP	1	3	4.W.3e	Writing
4	2	4	MC	OP	1	2	4.W.3c	Writing
4	2	5	MC	OP	1	2	4.W.1d	Writing
4	2	6	TE	OP	1	2	4.W.8	Writing
4	2	7	TE	OP	2	2	4.W.8	Writing
4	2	8	MC	OP	1	2	4.W.8	Writing
4	2	9	TE	OP	2	2	4.W.8	Writing
4	2	10	MC	OP	1	2	4.L.2b	Writing
4	2	11	TE	OP	2	2	4.L.1b	Writing
4	2	12	MC	OP	1	1	4.L.2a	Writing
4	2	13	TE	OP	2	2	4.L.1c	Writing
4	3	14	MC	OP	1	2	4.SL.3	Listening
4	3	15	MC	OP	1	1	4.SL.2	Listening
4	3	16	EBSR	OP	2	2	4.SL.3	Listening
4	3	17	EBSR	OP	2	2	4.SL.2	Listening
4	3	18	MC	OP	1	2	4.SL.2	Listening
4	3	19	MC	OP	1	2	4.SL.3	Listening
4	4	20	MC	OP	1	2	4.L.4	Reading
4	4	21	MC	OP	1	2	4.RL.3	Reading
4	4	22	TE	OP	2	2	4.RL.3	Reading
4	4	23	MC	OP	1	3	4.RL.2	Reading
4	4	24	TE	OP	2	2	4.RI.1	Reading
4	4	25	MC	OP	1	2	4.RI.5	Reading
4	4	26	MC	OP	1	2	4.RI.5	Reading
4	4	27	MC	OP	1	2	4.RI.5	Reading
4	4	28	MC	OP	1	2	4.L.5	Reading
4	4	29	TE	OP	1	2	4.L.4	Reading

Note: TDA item is weighted x 2 in computation of student scores.

Table C-2. English Language Arts, Grade 4 Test Map (cont.)

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Depth of Knowledge	Standard	Domain
4	4	30	TE	OP	2	2	4.RL.3	Reading
4	4	31	MC	OP	1	2	4.RL.6	Reading
4	4	32	MC	OP	1	2	4.RI.5	Reading
4	4	33	MC	OP	1	2	4.L.5	Reading
4	4	34	MC	OP	1	1	4.RI.1	Reading
4	4	35	MC	OP	1	2	4.RI.8	Reading
4	4	36	MC	OP	1	1	4.L.5	Reading
4	4	37	TE	OP	2	2	4.RL.1	Reading
4	4	38	MC	OP	1	2	4.RL.2	Reading
4	4	39	MC	OP	1	2	4.RL.6	Reading

Table C-3. English Language Arts, Grade 5 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Depth of Knowledge	Standard	Domain
5	1	1	TDA	OP	4	3	5.W.2	Writing
5	2	2	MC	OP	1	2	5.W.1c	Writing
5	2	3	MC	OP	1	2	5.W.2b	Writing
5	2	4	TE	OP	1	2	5.W.3a	Writing
5	2	5	MC	OP	1	2	5.W.1c	Writing
5	2	6	TE	OP	2	2	5.W.8	Writing
5	2	7	TE	OP	2	2	5.W.8	Writing
5	2	8	MC	OP	1	2	5.W.8	Writing
5	2	9	MC	OP	1	2	5.W.5	Writing
5	2	10	MC	OP	1	2	5.L.2	Writing
5	2	11	MC	OP	1	2	5.L.2b	Writing
5	2	12	MC	OP	1	2	5.L.3a	Writing
5	2	13	TE	OP	2	2	5.L.1b	Writing
5	2	14	MC	OP	1	1	5.L.2b	Writing
5	3	15	MC	OP	1	3	5.SL.3	Listening
5	3	16	MC	OP	1	2	5.SL.2	Listening
5	3	17	TE	OP	2	1	5.SL.3	Listening
5	3	18	TE	OP	2	3	5.SL.3	Listening
5	3	19	MC	OP	1	2	5.SL.2	Listening
5	3	20	MC	OP	1	2	5.SL.2	Listening
5	4	21	MC	OP	1	2	5.RL.5	Reading
5	4	22	TE	OP	2	2	5.RL.3	Reading
5	4	23	MC	OP	1	2	5.RL.2	Reading
5	4	24	MC	OP	1	3	5.RL.6	Reading
5	4	25	TE	OP	2	2	5.L.4	Reading
5	4	26	MC	OP	1	1	5.RI.3	Reading
5	4	27	MC	OP	1	2	5.RI.1	Reading

Note: TDA item is weighted x 2 in computation of student scores.

Table C-3. English Language Arts, Grade 5 Test Map (cont.)

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Depth of Knowledge	Standard	Domain
5	4	28	MC	OP	1	2	5.RI.1	Reading
5	4	29	MC	OP	1	2	5.RL.2	Reading
5	4	30	MC	OP	1	2	5.RL.5	Reading
5	4	31	MC	OP	1	2	5.L.5	Reading
5	4	32	MC	OP	1	2	5.RL.1	Reading
5	4	33	MC	OP	1	3	5.RL.6	Reading
5	4	34	TE	OP	2	2	5.RI.1	Reading
5	4	35	MC	OP	1	3	5.RI.8	Reading
5	4	36	MC	OP	1	2	5.RI.1	Reading
5	4	37	MC	OP	1	3	5.L.4	Reading
5	4	38	MC	OP	1	3	5.RL.5	Reading
5	4	39	MC	OP	1	2	5.RL.1	Reading
5	4	40	MC	OP	1	3	5.RL.6	Reading
5	4	41	MC	OP	1	2	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	Depth of Knowledge	Standard	Domain
6	1	1	TDA	OP	4	3	6.W.9	Writing
6	2	2	MC	OP	1	2	6.W.2e	Writing
6	2	3	MC	OP	1	2	6.W.3b	Writing
6	2	4	MC	OP	1	2	6.W.1e	Writing
6	2	5	MC	OP	1	2	6.W.2b	Writing
6	2	6	TE	OP	2	2	6.L.1c	Writing
6	2	7	TE	OP	2	2	6.L.1d	Writing
6	2	8	MC	OP	1	2	6.L.2a	Writing
6	2	9	TE	OP	1	2	6.L.3a	Writing
6	2	10	MC	OP	1	2	6.W.8	Writing
6	2	11	TE	OP	2	2	6.W.8	Writing
6	2	12	EBSR	OP	2	2	6.W.8	Writing
6	2	13	MC	OP	1	2	6.W.8	Writing
6	3	14	MC	OP	1	2	6.SL.2	Listening
6	3	15	MC	OP	1	2	6.SL.2	Listening
6	3	16	TE	OP	2	1	6.SL.2	Listening
6	3	17	EBSR	OP	2	3	6.SL.3	Listening
6	3	18	MC	OP	1	2	6.SL.3	Listening
6	3	19	MC	OP	1	3	6.SL.2	Listening
6	4	20	MC	OP	1	3	6.RL.5	Reading
6	4	21	TE	OP	2	2	6.RL.1	Reading
6	4	22	MC	OP	1	2	6.RL.4	Reading
6	4	23	TE	OP	2	2	6.RL.1	Reading
6	4	24	MC	OP	1	2	6.RI.8	Reading
6	4	25	MC	OP	1	2	6.RI.4	Reading
6	4	26	TE	OP	2	2	6.RI.1	Reading
6	4	27	MC	OP	1	3	6.RI.7	Reading
6	4	28	TE	OP	2	2	6.RI.2	Reading
6	4	29	MC	OP	1	1	6.L.5	Reading

Note: TDA item is weighted x 2 in computation of student scores.

Table C-4. English Language Arts, Grade 6 Test Map (cont.)

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Depth of Knowledge	Standard	Domain
6	4	30	MC	OP	1	3	6.RL.4	Reading
6	4	31	MC	OP	1	2	6.RL.6	Reading
6	4	32	TE	OP	2	3	6.RL.2	Reading
6	4	33	MC	OP	1	2	6.RI.4	Reading
6	4	34	TE	OP	2	2	6.RI.6	Reading
6	4	35	TE	OP	1	2	6.RI.3	Reading
6	4	36	MC	OP	1	2	6.RI.5	Reading
6	4	37	EBSR	OP	2	3	6.RI.6	Reading

Table C-5. English Language Arts, Grade 7 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Depth of Knowledge	Standard	Domain
7	1	1	TDA	OP	4	3	7.W.9	Writing
7	2	2	TE	OP	1	3	7.W.3e	Writing
7	2	3	TE	OP	2	2	7.W.3c	Writing
7	2	4	MC	OP	1	2	7.W.1e	Writing
7	2	5	MC	OP	1	2	7.L.2a	Writing
7	2	6	TE	OP	2	2	7.L.3a	Writing
7	2	7	MC	OP	1	2	7.L.1c	Writing
7	2	8	MC	OP	1	2	7.L.2	Writing
7	2	9	MC	OP	1	2	7.L.1b	Writing
7	2	10	MC	OP	1	2	7.W.8	Writing
7	2	11	MC	OP	1	2	7.W.8	Writing
7	2	12	EBSR	OP	2	3	7.W.8	Writing
7	2	13	MC	OP	1	2	7.W.5	Writing
7	2	14	MC	OP	1	2	7.W.8	Writing
7	3	15	TE	OP	2	2	7.SL.2	Listening
7	3	16	MC	OP	1	2	7.SL.2	Listening
7	3	17	MC	OP	1	3	7.SL.3	Listening
7	3	18	MC	OP	1	2	7.SL.3	Listening
7	3	19	MC	OP	1	2	7.SL.2	Listening
7	3	20	EBSR	OP	2	3	7.SL.3	Listening
7	4	21	TE	OP	1	2	7.RL.4	Reading
7	4	22	MC	OP	1	2	7.RL.3	Reading
7	4	23	MC	OP	1	2	7.L.4	Reading
7	4	24	EBSR	OP	2	2	7.RL.6	Reading
7	4	25	MC	OP	1	3	7.RL.2	Reading
7	4	26	MC	OP	1	3	7.RI.6	Reading
7	4	27	MC	OP	1	2	7.RI.1	Reading
7	4	28	TE	OP	1	2	7.RI.4	Reading

Note: TDA item is weighted x 2 in computation of student scores.

Table C-5. English Language Arts, Grade 7 Test Map (cont.)

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Depth of Knowledge	Standard	Domain
7	4	29	TE	OP	2	2	7.RI.1	Reading
7	4	30	MC	OP	1	2	7.RI.6	Reading
7	4	31	TE	OP	2	2	7.RI.5	Reading
7	4	32	EBSR	OP	2	2	7.RL.1	Reading
7	4	33	MC	OP	1	2	7.RL.3	Reading
7	4	34	MC	OP	1	2	7.RL.3	Reading
7	4	35	MC	OP	1	2	7.RL.2	Reading
7	4	36	MC	OP	1	2	7.RI.5	Reading
7	4	37	TE	OP	2	2	7.RI.4	Reading
7	4	38	MC	OP	1	2	7.RI.3	Reading
7	4	39	MC	OP	1	2	7.RI.2	Reading

Table C-6. English Language Arts, Grade 8 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Depth of Knowledge	Standard	Domain
8	1	1	TDA	OP	4	3	8.W.2	Writing
8	2	2	MC	OP	1	2	8.W.2b	Writing
8	2	3	TE	OP	2	2	8.W.1c	Writing
8	2	4	MC	OP	1	2	8.W.3d	Writing
8	2	5	MC	OP	1	2	8.L.2b	Writing
8	2	6	TE	OP	2	2	8.L.2c	Writing
8	2	7	MC	OP	1	2	8.L.2a	Writing
8	2	8	MC	OP	1	2	8.L.2b	Writing
8	2	9	MC	OP	1	2	8.L.2c	Writing
8	2	10	MC	OP	1	2	8.W.8	Writing
8	2	11	TE	OP	1	2	8.W.8	Writing
8	2	12	MC	OP	1	3	8.W.8	Writing
8	2	13	MC	OP	1	2	8.W.5	Writing
8	2	14	MC	OP	1	2	8.W.8	Writing
8	2	15	MC	OP	1	2	8.W.8	Writing
8	3	16	MC	OP	1	2	8.SL.2	Listening
8	3	17	MC	OP	1	2	8.SL.2	Listening
8	3	18	EBSR	OP	2	3	8.SL.3	Listening
8	3	19	EBSR	OP	2	3	8.SL.3	Listening
8	3	20	MC	OP	1	2	8.SL.3	Listening
8	3	21	MC	OP	1	2	8.SL.2	Listening
8	4	22	TE	OP	1	2	8.RI.1	Reading
8	4	23	MC	OP	1	2	8.RI.5	Reading
8	4	24	EBSR	OP	2	3	8.RI.2	Reading
8	4	25	MC	OP	1	3	8.RI.3	Reading
8	4	26	MC	OP	1	3	8.RI.8	Reading
8	4	27	MC	OP	1	2	8.RL.4	Reading
8	4	28	MC	OP	1	3	8.RL.3	Reading

Note: TDA item is weighted x 2 in computation of student scores.

Table C-6. English Language Arts, Grade 8 Test Map (cont.)

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Depth of Knowledge	Standard	Domain
8	4	29	TE	OP	2	2	8.RL.4	Reading
8	4	30	MC	OP	1	2	8.RL.3	Reading
8	4	31	MC	OP	1	3	8.RL.6	Reading
8	4	32	EBSR	OP	2	3	8.RL.2	Reading
8	4	33	MC	OP	1	3	8.RI.5	Reading
8	4	34	EBSR	OP	2	2	8.L.4	Reading
8	4	35	MC	OP	1	2	8.RI.5	Reading
8	4	36	MC	OP	1	2	8.RI.3	Reading
8	4	37	MC	OP	1	2	8.L.4	Reading
8	4	38	TE	OP	2	2	8.RL.2	Reading
8	4	39	MC	OP	1	3	8.RL.3	Reading
8	4	40	MC	OP	1	3	8.RL.6	Reading

Appendix D

Spring 2018 Mathematics Operational Test Maps

Table D-1 Mathematics Grade 3 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Depth of Knowledge	Standard	Domain
3	1	1	MC	OP	1	2	3.NBT.1	NBT
3	1	2	MC	OP	1	2	3.NF.1	NF
3	1	3	MC	OP	1	2	3.OA.1	OA
3	1	4	MC	OP	1	2	3.MD.1	MD
3	1	5	MC	OP	1	3	3.G.1	G
3	1	6	TE	OP	1	2	3.NBT.1	NBT
3	1	7	SA	OP	1	1	3.OA.4	OA
3	1	8	MC	OP	1	2	3.MD.7	MD
3	1	9	MC	OP	1	1	3.NF.2	NF
3	1	10	SA	OP	1	1	3.NBT.1	NBT
3	1	11	MC	OP	1	2	3.OA.6	OA
3	1	12	SA	OP	1	2	3.MD.1	MD
3	1	13	MC	OP	1	1	3.NBT.2	NBT
3	1	14	MC	OP	1	1	3.MD.5	MD
3	1	15	MC	OP	1	2	3.NF.3	NF
3	1	16	SA	OP	1	1	3.NF.3b	NF
3	1	17	MC	OP	1	2	3.OA.8	OA
3	1	18	MC	OP	1	2	3.G.2	G
3	1	19	MC	OP	1	2	3.NBT.3	NBT
3	1	20	TE	OP	1	3	3.NF.3	NF
3	1	21	MC	OP	1	2	3.G.2	G
3	2	22	MC	OP	1	2	3.MD.8	MD
3	2	23	MC	OP	1	2	3.NBT.1	NBT
3	2	24	MC	OP	1	2	3.OA.2	OA
3	2	25	MC	OP	1	1	3.G.1	G
3	2	26	SA	OP	1	2	3.OA.7	OA
3	2	27	TE	OP	1	2	3.G.1	G
3	2	28	MC	OP	1	2	3.NF.2	NF
3	2	29	MC	OP	1	3	3.NBT.2	NBT
3	2	30	MC	OP	1	1	3.MD.2	MD
3	2	31	MC	OP	1	2	3.G.2	G
3	2	32	MC	OP	1	1	3.OA.5	OA
3	2	33	SA	OP	1	1	3.MD.7b	MD
3	2	34	MC	OP	1	3	3.NF.3d	NF
3	2	35	MC	OP	1	2	3.NF.2	NF

Table D-1 Mathematics Grade 3 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Depth of Knowledge	Standard	Domain
3	2	36	TE	OP	1	3	3.OA.3	OA
3	2	37	MC	OP	1	2	3.OA.7	OA
3	2	38	SA	OP	1	1	3.G.2	G
3	2	39	MC	OP	1	2	3.NBT.2	NBT
3	2	40	MC	OP	1	2	3.MD.6	MD
3	2	41	MC	OP	1	2	3.MD.3	MD
3	2	42	TE	OP	1	2	3.MD.4	MD

Domain Names: OA= Operations and Algebraic Thinking; NBT= Number and Operations in Base Ten; NF= Number and Operations – Fractions; MD= Measurement and Data; G=Geometry

Table D-2 Mathematics Grade 4 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Depth of Knowledge	Standard	Domain
4	1	1	MC	OP	1	3	4.NBT.2	NBT
4	1	2	MC	OP	1	1	4.OA.1	OA
4	1	3	MC	OP	1	2	4.G.1	G
4	1	4	SA	OP	1	3	4.OA.3	OA
4	1	5	MC	OP	1	2	4.NBT.3	NBT
4	1	6	TE	OP	1	2	4.NBT.5	NBT
4	1	7	MC	OP	1	2	4.OA.2	OA
4	1	8	MC	OP	1	1	4.G.1	G
4	1	9	MC	OP	1	2	4.OA.3	OA
4	1	10	MC	OP	1	2	4.G.2	G
4	1	11	MC	OP	1	1	4.OA.4	OA
4	1	12	TE	OP	1	2	4.NBT.1	NBT
4	1	13	MC	OP	1	2	4.NF.3	NF
4	1	14	MC	OP	1	2	4.MD.4	MD
4	1	15	MC	OP	1	2	4.NBT.5	NBT
4	1	16	SA	OP	1	1	4.NF.6	NF
4	1	17	MC	OP	1	1	4.MD.1	MD
4	1	18	MC	OP	1	2	4.NF.4	NF
4	1	19	MC	OP	1	2	4.OA.5	OA
4	1	20	TE	OP	1	2	4.NF.6	NF
4	1	21	MC	OP	1	1	4.MD.6	MD
4	1	22	MC	OP	1	2	4.G.3	G
4	1	23	MC	OP	1	2	4.NF.7	NF
4	2	24	MC	OP	1	2	4.NBT.2	NBT
4	2	25	MC	OP	1	2	4.OA.1	OA
4	2	26	MC	OP	1	2	4.G.1	G
4	2	27	MC	OP	1	2	4.MD.2	MD
4	2	28	MC	OP	1	1	4.NF.1	NF
4	2	29	TE	OP	1	2	4.NBT.4	NBT
4	2	30	SA	OP	1	1	4.OA.4	OA
4	2	31	MC	OP	1	2	4.MD.3	MD
4	2	32	MC	OP	1	2	4.OA.3	OA
4	2	33	MC	OP	1	1	4.NF.2	NF
4	2	34	MC	OP	1	2	4.OA.2	OA

Table D-2 Mathematics Grade 4 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Depth of Knowledge	Standard	Domain
4	2	35	MC	OP	1	2	4.MD.7	MD
4	2	36	TE	OP	1	2	4.MD.3	MD
4	2	37	MC	OP	1	3	4.G.3	G
4	2	38	MC	OP	1	2	4.NBT.6	NBT
4	2	39	MC	OP	1	1	4.MD.5	MD
4	2	40	MC	OP	1	1	4.G.2	G
4	2	41	MC	OP	1	2	4.NF.5	NF
4	2	42	MC	OP	1	3	4.NF.1	NF
4	2	43	SA	OP	1	1	4.NBT.4	NBT
4	2	44	MC	OP	1	2	4.NF.4c	NF
4	2	45	MC	OP	1	2	4.MD.7	MD
4	2	46	MC	OP	1	1	4.MD.6	MD

Domain Names: OA= Operations and Algebraic Thinking; NBT= Number and Operations in Base Ten; NF= Number and Operations – Fractions; MD= Measurement and Data; G=Geometry

Table D-3 Mathematics Grade 5 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Depth of Knowledge	Standard	Domain
5	1	1	SA	OP	1	1	5.OA.1	OA
5	1	2	MC	OP	1	2	5.NF.1	NF
5	1	3	MC	OP	1	2	5.MD.1	MD
5	1	4	MC	OP	1	2	5.G.1	G
5	1	5	MC	OP	1	3	5.NF.5	NF
5	1	6	TE	OP	1	2	5.NF.2	NF
5	1	7	TE	OP	1	2	5.G.2	G
5	1	8	MC	OP	1	2	5.MD.2	MD
5	1	9	MC	OP	1	1	5.NBT.4	NBT
5	1	10	SA	OP	1	2	5.MD.4	MD
5	1	11	MC	OP	1	2	5.OA.2	OA
5	1	12	SA	OP	1	1	5.NBT.1	NBT
5	1	13	MC	OP	1	2	5.MD.3	MD
5	1	14	MC	OP	1	2	5.OA.2	OA
5	1	15	TE	OP	1	2	5.OA.3	OA
5	1	16	SA	OP	1	2	5.G.4	G
5	1	17	MC	OP	1	1	5.NF.6	NF
5	1	18	MC	OP	1	2	5.G.1	G
5	1	19	MC	OP	1	2	5.OA.1	OA
5	1	20	TE	OP	1	2	5.NBT.5	NBT
5	1	21	MC	OP	1	3	5.NBT.6	NBT
5	1	22	SA	OP	1	2	5.NF.7	NF
5	1	23	MC	OP	1	2	5.NBT.1	NBT
5	2	24	MC	OP	1	1	5.NBT.2	NBT
5	2	25	SA	OP	1	3	5.OA.1	OA
5	2	26	MC	OP	1	1	5.G.3	G
5	2	27	MC	OP	1	2	5.G.2	G
5	2	28	MC	OP	1	2	5.MD.2	MD
5	2	29	TE	OP	1	2	5.OA.1	OA
5	2	30	SA	OP	1	2	5.G.2	G
5	2	31	SA	OP	1	2	5.MD.1	MD
5	2	32	MC	OP	1	1	5.OA.2	OA
5	2	33	SA	OP	1	1	5.NBT.5	NBT
5	2	34	SA	OP	1	1	5.G.1	G

Table D-3 Mathematics Grade 5 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Depth of Knowledge	Standard	Domain
5	2	35	TE	OP	1	2	5.MD.2	MD
5	2	36	MC	OP	1	2	5.NF.5	NF
5	2	37	TE	OP	1	2	5.NBT.3	NBT
5	2	38	MC	OP	1	2	5.G.2	G
5	2	39	MC	OP	1	2	5.NF.3	NF
5	2	40	MC	OP	1	2	5.NBT.7	NBT
5	2	41	MC	OP	1	2	5.MD.5	MD
5	2	42	MC	OP	1	2	5.MD.3	MD
5	2	43	MC	OP	1	2	5.OA.2	OA
5	2	44	SA	OP	1	1	5.NF.4b	NF
5	2	45	MC	OP	1	3	5.NF.2	NF
5	2	46	MC	OP	1	2	5.MD.4	MD

Domain Names: OA= Operations and Algebraic Thinking; NBT= Number and Operations in Base Ten; NF= Number and Operations – Fractions; MD= Measurement and Data; G=Geometry

Table D-4 Mathematics Grade 6 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Depth of Knowledge	Standard	Domain
6	1	1	MC	OP	1	2	6.EE.1	EE
6	1	2	SA	OP	1	1	6.NS.2	NS
6	1	3	MC	OP	1	2	6.EE.4	EE
6	1	4	TE	OP	1	3	6.RP.3	RP
6	1	5	MC	OP	1	1	6.RP.3	RP
6	1	6	SA	OP	1	2	6.EE.2c	EE
6	1	7	MC	OP	1	2	6.NS.3	NS
6	1	8	MC	OP	1	1	6.NS.3	NS
6	1	9	SA	OP	1	1	6.EE.1	EE
6	1	10	MC	OP	1	2	6.NS.2	NS
6	1	11	SA	OP	1	2	6.RP.3c	RP
6	1	12	MC	OP	1	2	6.EE.2	EE
6	1	13	SA	OP	1	2	6.RP.1	RP
6	1	14	MC	OP	1	2	6.RP.2	RP
6	1	15	TE	OP	1	2	6.RP.3a	RP
6	1	16	SA	OP	1	1	6.RP.2	RP
6	2	17	MC	OP	1	2	6.SP.1	SP
6	2	18	MC	OP	1	1	6.NS.5	NS
6	2	19	SA	OP	1	1	6.G.1	G
6	2	20	MC	OP	1	2	6.G.1	G
6	2	21	MC	OP	1	1	6.SP.3	SP
6	2	22	MC	OP	1	2	6.EE.7	EE
6	2	23	MC	OP	1	2	6.NS.6	NS
6	2	24	TE	OP	1	1	6.SP.4	SP
6	2	25	MC	OP	1	1	6.SP.1	SP
6	2	26	MC	OP	1	2	6.G.3	G
6	2	27	SA	OP	1	2	6.NS.6b	NS
6	2	28	MC	OP	1	2	6.EE.8	EE
6	2	29	MC	OP	1	2	6.EE.9	EE
6	2	30	TE	OP	1	2	6.NS.6c	NS
6	2	31	SA	OP	1	2	6.EE.5	EE
6	2	32	MC	OP	1	2	6.NS.8	NS
6	2	33	MC	OP	1	2	6.SP.3	SP
6	2	34	MC	OP	1	2	6.SP.4	SP

Table D-4 Mathematics Grade 6 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Depth of Knowledge	Standard	Domain
6	2	35	MC	OP	1	2	6.EE.6	EE
6	2	36	MC	OP	1	2	6.SP.5	SP
6	2	37	TE	OP	1	1	6.NS.6	NS
6	2	38	MC	OP	1	2	6.EE.5	EE
6	2	39	MC	OP	1	2	6.SP.2	SP
6	2	40	MC	OP	1	2	6.NS.8	NS
6	2	41	SA	OP	1	2	6.SP.5c	SP
6	2	42	MC	OP	1	2	6.G.2	G
6	2	43	MC	OP	1	2	6.G.4	G
6	2	44	TE	OP	1	2	6.G.3	G
6	2	45	MC	OP	1	2	6.G.3	G
6	2	46	MC	OP	1	2	6.SP.5	SP

Domain Names: G=Geometry; EE=Expressions and Equations; NS=The Number System; SP=Statistics and Probability; RP= Ratios and Proportional Relationships

Table D-5 Mathematics Grade 7 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Depth of Knowledge	Standard	Domain
7	1	1	MC	OP	1	2	7.EE.2	EE
7	1	2	MC	OP	1	2	7.NS.3	NS
7	1	3	MC	OP	1	2	7.EE.2	EE
7	1	4	SA	OP	1	1	7.NS.2d	NS
7	1	5	MC	OP	1	2	7.EE.1	EE
7	1	6	MC	OP	1	2	7.NS.2	NS
7	1	7	MC	OP	1	2	7.EE.1	EE
7	1	8	MC	OP	1	2	7.NS.1	NS
7	1	9	TE	OP	1	1	7.NS.2	NS
7	1	10	MC	OP	1	2	7.NS.3	NS
7	1	11	MC	OP	1	2	7.NS.1	NS
7	2	12	MC	OP	1	2	7.EE.3	EE
7	2	13	MC	OP	1	2	7.RP.2	RP
7	2	14	SA	OP	1	2	7.RP.1	RP
7	2	15	MC	OP	1	2	7.G.3	G
7	2	16	MC	OP	1	2	7.EE.4	EE
7	2	17	MC	OP	1	2	7.G.1	G
7	2	18	MC	OP	1	2	7.RP.3	RP
7	2	19	TE	OP	1	2	7.RP.2d	RP
7	2	20	MC	OP	1	2	7.G.4	G
7	2	21	SA	OP	1	1	7.SP.7a	SP
7	2	22	MC	OP	1	2	7.G.4	G
7	2	23	MC	OP	1	2	7.SP.1	SP
7	2	24	TE	OP	1	2	7.EE.4	EE
7	2	25	MC	OP	1	2	7.SP.5	SP
7	2	26	SA	OP	1	2	7.G.1	G
7	2	27	MC	OP	1	3	7.SP.2	SP
7	2	28	MC	OP	1	2	7.SP.1	SP
7	2	29	MC	OP	1	2	7.EE.4a	EE
7	2	30	MC	OP	1	3	7.RP.2a	RP
7	2	31	MC	OP	1	2	7.RP.1	RP
7	2	32	MC	OP	1	2	7.G.3	G
7	2	33	SA	OP	1	2	7.EE.4	EE
7	2	34	MC	OP	1	2	7.RP.3	RP

Table D-5 Mathematics Grade 7 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Depth of Knowledge	Standard	Domain
7	2	35	MC	OP	1	2	7.RP.3	RP
7	2	36	MC	OP	1	2	7.G.6	G
7	2	37	MC	OP	1	2	7.SP.6	SP
7	2	38	MC	OP	1	2	7.SP.3	SP
7	2	39	MC	OP	1	2	7.G.4	G
7	2	40	TE	OP	1	2	7.SP.6	SP
7	2	41	SA	OP	1	1	7.G.4	G
7	2	42	MC	OP	1	2	7.EE.3	EE
7	2	43	MC	OP	1	2	7.SP.2	SP
7	2	44	TE	OP	1	2	7.G.1	G
7	2	45	MC	OP	1	2	7.SP.2	SP
7	2	46	MC	OP	1	2	7.SP.3	SP

Domain Names: G=Geometry; EE=Expressions and Equations; NS=The Number System; SP=Statistics and Probability; RP= Ratios and Proportional Relationships

Table D-6 Mathematics Grade 8 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Depth of Knowledge	Standard	Domain
8	1	1	MC	OP	1	2	8.EE.1	EE
8	1	2	SA	OP	1	2	8.NS.2	NS
8	1	3	MC	OP	1	1	8.EE.2	EE
8	1	4	MC	OP	1	1	8.NS.1	NS
8	1	5	SA	OP	1	1	8.EE.1	EE
8	1	6	MC	OP	1	3	8.NS.1	NS
8	1	7	MC	OP	1	2	8.NS.1	NS
8	1	8	MC	OP	1	1	8.EE.2	EE
8	1	9	MC	OP	1	1	8.NS.1	NS
8	1	10	MC	OP	1	1	8.NS.1	NS
8	1	11	TE	OP	1	2	8.NS.2	NS
8	1	12	MC	OP	1	2	8.EE.3	EE
8	1	13	SA	OP	1	1	8.NS.2	NS
8	2	14	MC	OP	1	2	8.G.1	G
8	2	15	MC	OP	1	2	8.EE.5	EE
8	2	16	TE	OP	1	3	8.G.5	G
8	2	17	MC	OP	1	2	8.SP.1	SP
8	2	18	MC	OP	1	2	8.F.4	F
8	2	19	MC	OP	1	2	8.G.7	G
8	2	20	MC	OP	1	1	8.G.1	G
8	2	21	MC	OP	1	3	8.F.2	F
8	2	22	MC	OP	1	2	8.SP.2	SP
8	2	23	MC	OP	1	2	8.F.5	F
8	2	24	TE	OP	1	2	8.F.2	F
8	2	25	MC	OP	1	2	8.G.3	G
8	2	26	MC	OP	1	2	8.SP.3	SP
8	2	27	SA	OP	1	2	8.EE.8b	EE
8	2	28	MC	OP	1	2	8.F.4	F
8	2	29	TE	OP	1	2	8.SP.4	SP
8	2	30	MC	OP	1	2	8.SP.3	SP
8	2	31	MC	OP	1	2	8.F.2	F
8	2	32	TE	OP	1	2	8.F.3	F
8	2	33	MC	OP	1	2	8.G.4	G
8	2	34	MC	OP	1	2	8.SP.4	SP

Table D-6 Mathematics Grade 8 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Depth of Knowledge	Standard	Domain
8	2	35	MC	OP	1	2	8.G.5	G
8	2	36	MC	OP	1	2	8.EE.6	EE
8	2	37	TE	OP	1	2	8.G.2	G
8	2	38	MC	OP	1	2	8.F.2	F
8	2	39	MC	OP	1	2	8.G.6	G
8	2	40	MC	OP	1	2	8.F.5	F
8	2	41	MC	OP	1	1	8.F.1	F
8	2	42	MC	OP	1	2	8.EE.7	EE
8	2	43	SA	OP	1	2	8.G.8	G
8	2	44	MC	OP	1	2	8.SP.1	SP
8	2	45	MC	OP	1	2	8.EE.8	EE
8	2	46	MC	OP	1	2	8.SP.3	SP

Domain Names: G=Geometry; EE=Expressions and Equations; NS=The Number System; SP=Statistics and Probability; F= Functions

Appendix E

Spring 2018 Science Operational Test Maps

Table E-1 Science Grade 4 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Depth of Knowledge	Standard	Domain
4	1	1	MC	OP	1	2	B.4.3	Science Connections & Nature of Science
4	1	2	MC	OP	1	1	C.4.4	Science Inquiry
4	1	3	MC	OP	1	1	D.4.4	Physical Science
4	1	4	MC	OP	1	2	F.4.1	Life and Environmental Science
4	1	5	MC	OP	1	3	A.4.2	Science Connections & Nature of Science
4	1	6	MC	OP	1	2	H.4.4	Science Applications & Science in Social and Personal Perspectives
4	1	7	MC	OP	1	3	C.4.2	Science Inquiry
4	1	8	MC	OP	1	1	E.4.8	Earth Science
4	1	9	MC	OP	1	2	C.4.5	Science Inquiry
4	1	10	MC	OP	1	2	C.4.6	Science Inquiry
4	1	11	TE (text highlight)	OP	1	2	E.4.8	Earth Science
4	1	12	MC	OP	1	2	H.4.1	Science Applications & Science in Social and Personal Perspectives
4	1	13	MC	OP	1	2	G.4.3	Science Applications & Science in Social and Personal Perspectives
4	1	14	MC	OP	1	3	C.4.7	Science Inquiry
4	1	15	MC	OP	1	1	D.4.3	Physical Science
4	1	16	MC	OP	1	3	F.4.2	Life and Environmental Science
4	1	17	MC	OP	1	2	A.4.4	Science Connections & Nature of Science
4	1	18	MC	OP	1	2	E.4.6	Earth Science
4	1	19	MC	OP	1	1	G.4.5	Science Applications & Science in Social and Personal Perspectives
4	1	20	MC	OP	1	1	F.4.2	Life and Environmental Science

Table E-1 Science Grade 4 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Depth of Knowledge	Standard	Domain
4	1	21	MC	OP	1	2	A.4.3	Science Connections & Nature of Science
4	2	22	MC	OP	1	2	F.4.4	Life and Environmental Science
4	2	23	MC	OP	1	2	H.4.2	Science Applications & Science in Social and Personal Perspectives
4	2	24	MC	OP	1	1	D.4.4	Physical Science
4	2	25	MC	OP	1	2	D.4.8	Physical Science
4	2	26	MC	OP	1	2	B.4.1	Science Connections & Nature of Science
4	2	27	MC	OP	1	1	C.4.5	Science Inquiry
4	2	28	MC	OP	1	1	F.4.1	Life and Environmental Science
4	2	29	MC	OP	1	1	E.4.5	Earth Science
4	2	30	MC	OP	1	1	C.4.4	Science Inquiry
4	2	31	MC	OP	1	2	C.4.8	Science Inquiry
4	2	32	MC	OP	1	2	F.4.4	Life and Environmental Science
4	2	33	MC	OP	1	1	H.4.1	Science Applications & Science in Social and Personal Perspectives
4	2	34	MC	OP	1	2	B.4.1	Science Connections & Nature of Science
4	2	35	MC	OP	1	2	G.4.1	Science Applications & Science in Social and Personal Perspectives
4	2	36	MC	OP	1	2	E.4.4	Earth Science
4	2	37	MC	OP	1	2	G.4.4	Science Applications & Science in Social and Personal Perspectives
4	2	38	MC	OP	1	1	D.4.8	Physical Science
4	2	39	MC	OP	1	2	C.4.1	Science Inquiry
4	2	40	MC	OP	1	2	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	Depth of Knowledge	Standard	Domain
8	1	1	MC	OP	1	2	B.8.3	Science Connections & Nature of Science
8	1	2	MC	OP	1	2	G.8.6	Science Applications & Science in Social and Personal Perspectives
8	1	3	MC	OP	1	1	A.8.5	Science Connections & Nature of Science
8	1	4	MC	OP	1	2	C.8.3	Science Inquiry
8	1	5	MC	OP	1	2	C.8.4	Science Inquiry
8	1	6	MC	OP	1	2	D.8.8	Physical Science
8	1	7	MC	OP	1	2	G.8.1	Science Applications & Science in Social and Personal Perspectives
8	1	8	MC	OP	1	2	D.8.6	Physical Science
8	1	9	MC	OP	1	2	C.8.6	Science Inquiry
8	1	10	MC	OP	1	2	E.8.3	Earth Science
8	1	11	MC	OP	1	2	F.8.8	Life and Environmental Science
8	1	12	MC	OP	1	2	F.8.9	Life and Environmental Science
8	1	13	MC	OP	1	2	D.8.2	Physical Science
8	1	14	MC	OP	1	2	G.8.7	Science Applications & Science in Social and Personal Perspectives
8	1	15	MC	OP	1	2	E.8.3	Earth Science
8	1	16	MC	OP	1	1	F.8.8	Life and Environmental Science
8	1	17	MC	OP	1	2	E.8.1	Earth Science
8	1	18	MC	OP	1	1	B.8.1	Science Connections & Nature of Science
8	1	19	MC	OP	1	2	C.8.1	Science Inquiry
8	1	20	MC	OP	1	2	C.8.2	Science Inquiry
8	1	21	MC	OP	1	2	D.8.6	Physical Science

Table E-2 Science Grade 8 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Depth of Knowledge	Standard	Domain
8	2	22	MC	OP	1	2	E.8.5	Earth Science
8	2	23	MC	OP	1	2	A.8.6	Science Connections & Nature of Science
8	2	24	MC	OP	1	2	E.8.2	Earth Science
8	2	25	MC	OP	1	1	F.8.1	Life and Environmental Science
8	2	26	MC	OP	1	2	G.8.3	Science Applications & Science in Social and Personal Perspectives
8	2	27	MC	OP	1	2	F.8.8	Life and Environmental Science
8	2	28	MC	OP	1	2	D.8.8	Physical Science
8	2	29	MC	OP	1	2	A.8.3	Science Connections & Nature of Science
8	2	30	MC	OP	1	2	C.8.6	Science Inquiry
8	2	31	MC	OP	1	2	G.8.3	Science Applications & Science in Social and Personal Perspectives
8	2	32	MC	OP	1	2	G.8.4	Science Applications & Science in Social and Personal Perspectives
8	2	33	MC	OP	1	2	G.8.5	Science Applications & Science in Social and Personal Perspectives
8	2	34	MC	OP	1	2	H.8.3	Science Applications & Science in Social and Personal Perspectives
8	2	35	MC	OP	1	2	B.8.6	Science Connections & Nature of Science
8	2	36	MC	OP	1	2	C.8.6	Science Inquiry
8	2	37	MC	OP	1	2	C.8.6	Science Inquiry
8	2	38	MC	OP	1	2	B.8.4	Science Connections & Nature of Science
8	2	39	MC	OP	1	2	C.8.10	Science Inquiry
8	2	40	MC	OP	1	1	F.8.8	Life and Environmental Science

Appendix F

Spring 2018 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	Depth of Knowledge	Standard	Domain
4	1	1	MC	OP	1	2	C.4.5	Civics
4	1	2	MC	OP	1	2	C.4.4	Civics
4	1	3	MC	OP	1	2	C.4.3	Civics
4	1	4	MC	OP	1	N/A	D.4.2	Economics
4	1	5	MC	OP	1	N/A	D.4.1	Economics
4	1	6	Drag and Drop	OP	1	2	B.4.8	History
4	1	7	MC	OP	1	N/A	A.4.9	Geography
4	1	8	MC	OP	1	2	E.4.15	Behavioral Sciences
4	1	9	MC	OP	1	2	B.4.2	History
4	1	10	Multi-Select	OP	1	2	C.4.1	Civics
4	1	11	MC	OP	1	2	C.4.6	Civics
4	1	12	MC	OP	1	N/A	A.4.4	Geography
4	1	13	MC	OP	1	3	D.4.3	Economics
4	1	14	MC	OP	1	2	D.4.2	Economics
4	1	15	MC	OP	1	2	E.4.6	Behavioral Sciences
4	1	16	MC	OP	1	N/A	E.4.3	Behavioral Sciences
4	1	17	MC	OP	1	3	B.4.9	History
4	1	18	MC	OP	1	2	A.4.7	Geography
4	1	19	MC	OP	1	N/A	A.4.4	Geography
4	2	20	MC	OP	1	2	E.4.9	Behavioral Sciences
4	2	21	MC	OP	1	N/A	B.4.7	History
4	2	22	MC	OP	1	2	C.4.1	Civics
4	2	23	MC	OP	1	2	B.4.10	History
4	2	24	MC	OP	1	2	C.4.4	Civics
4	2	25	MC	OP	1	3	B.4.4	History
4	2	26	Drag and Drop	OP	1	1	A.4.2	Geography
4	2	27	MC	OP	1	2	A.4.9	Geography
4	2	28	Drag and Drop	OP	1	2	B.4.2	History
4	2	29	MC	OP	1	2	D.4.7	Economics

Table F-1 Social Studies Grade 4 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Depth of Knowledge	Standard	Domain
4	2	30	MC	OP	1	1	A.4.2	Geography
4	2	31	MC	OP	1	2	A.4.6	Geography
4	2	32	MC	OP	1	2	B.4.7	History
4	2	33	MC	OP	1	2	A.4.4	Geography
4	2	34	MC	OP	1	N/A	E.4.10	Behavioral Sciences
4	2	35	MC	OP	1	N/A	D.4.5	Economics
4	2	36	MC	OP	1	N/A	E.4.15	Behavioral Sciences
4	2	37	MC	OP	1	N/A	E.4.15	Behavioral Sciences
4	2	38	MC	OP	1	2	B.4.6	History

Table F-2 Social Studies Grade 8 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Depth of Knowledge	Standard	Domain
8	1	1	MC	OP	1	2	B.8.1	History
8	1	2	MC	OP	1	2	B.8.1	History
8	1	3	MC	OP	1	2	A.8.7	Geography
8	1	4	MC	OP	1	3	A.8.8	Geography
8	1	5	MC	OP	1	3	B.8.4	History
8	1	6	MC	OP	1	2	B.8.2	History
8	1	7	MC	OP	1	3	A.8.3	Geography
8	1	8	MC	OP	1	3	A.8.1	Geography
8	1	9	MC	OP	1	1	C.8.9	Civics
8	1	10	MC	OP	1	2	D.8.8	Economics
8	1	11	MC	OP	1	2	D.8.9	Economics
8	1	12	Multi-Select	OP	1	2	E.8.4	Behavioral Sciences
8	1	13	MC	OP	1	3	B.8.5	History
8	1	14	MC	OP	1	3	C.8.9	Civics
8	1	15	MC	OP	1	2	A.8.5	Geography
8	1	16	MC	OP	1	3	E.8.9	Behavioral Sciences
8	1	17	Multi-Select	OP	1	2	A.8.11	Geography
8	1	18	MC	OP	1	2	A.8.9	Geography
8	1	19	MC	OP	1	2	A.8.11	Geography
8	1	20	MC	OP	1	2	E.8.3	Behavioral Sciences
8	2	21	MC	OP	1	2	D.8.4	Economics
8	2	22	Drag and Drop	OP	1	2	B.8.12	History
8	2	23	MC	OP	1	3	B.8.2	History
8	2	24	MC	OP	1	3	D.8.3	Economics
8	2	25	MC	OP	1	2	C.8.2	Civics
8	2	26	MC	OP	1	2	E.8.11	Behavioral Sciences
8	2	27	Matching	OP	1	2	C.8.1	Civics
8	2	28	MC	OP	1	3	C.8.8	Civics
8	2	29	MC	OP	1	2	B.8.1	History
8	2	30	MC	OP	1	2	E.8.10	Behavioral Sciences
8	2	31	MC	OP	1	N/A	A.8.8	Geography

Table F-2 Social Studies Grade 8 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Depth of Knowledge	Standard	Domain
8	2	32	MC	OP	1	3	B.8.9	History
8	2	33	MC	OP	1	3	B.8.7	History
8	2	34	MC	OP	1	2	C.8.6	Civics
8	2	35	MC	OP	1	3	B.8.3	History
8	2	36	MC	OP	1	2	D.8.2	Economics
8	2	37	MC	OP	1	2	D.8.7	Economics
8	2	38	MC	OP	1	3	B.8.10	History
8	2	39	MC	OP	1	3	E.8.4	Behavioral Sciences
8	2	40	MC	OP	1	3	A.8.9	Geography

Table F-3 Social Studies Grade 10 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Depth of Knowledge	Standard	Domain
10	1	1	MC	OP	1	N/A	A.10.6	Geography
10	1	2	MC	OP	1	N/A	C.10.16	Civics
10	1	3	MC	OP	1	2	E.10.6	Behavioral Sciences
10	1	4	MC	OP	1	2	A.10.3	Geography
10	1	5	MC	OP	1	2	B.10.7	History
10	1	6	MC	OP	1	N/A	D.10.8	Economics
10	1	7	Matching	OP	1	2	D.10.9	Economics
10	1	8	MC	OP	1	2	E.10.14	Behavioral Sciences
10	1	9	Drag and Drop	OP	1	2	A.10.8	Geography
10	1	10	MC	OP	1	2	D.10.10	Economics
10	1	11	MC	OP	1	2	C.10.10	Civics
10	1	12	MC	OP	1	2	B.10.16	History
10	1	13	MC	OP	1	2	B.10.13	History
10	1	14	MC	OP	1	3	D.10.7	Economics
10	1	15	MC	OP	1	2	E.10.5	Behavioral Sciences
10	1	16	MC	OP	1	2	A.10.6	Geography
10	1	17	MC	OP	1	3	B.10.9	History
10	1	18	MC	OP	1	3	D.10.2	Economics
10	1	19	MC	OP	1	2	A.10.11	Geography
10	1	20	MC	OP	1	3	C.10.4	Civics
10	1	21	MC	OP	1	2	D.10.14	Economics
10	1	22	MC	OP	1	N/A	B.10.6	History
10	1	23	MC	OP	1	N/A	B.10.12	History
10	1	24	MC	OP	1	2	E.10.6	Behavioral Sciences
10	1	25	MC	OP	1	N/A	D.10.1	Economics
10	2	26	MC	OP	1	2	C.10.16	Civics
10	2	27	MC	OP	1	N/A	D.10.4	Economics
10	2	28	MC	OP	1	2	C.10.13	Civics
10	2	29	MC	OP	1	2	B.10.10	History
10	2	30	Matching	OP	1	2	C.10.2	Civics
10	2	31	MC	OP	1	2	E.10.12	Behavioral Sciences
10	2	32	MC	OP	1	2	A.10.12	Geography
10	2	33	MC	OP	1	3	B.10.18	History

Table F-3 Social Studies Grade 10 Test Map

Grade	Session	Item Sequence	Item Type	Item Usage	Max Score Points	Depth of Knowledge	Standard	Domain
10	2	34	MC	OP	1	3	B.10.6	History
10	2	35	MC	OP	1	3	E.10.4	Behavioral Sciences
10	2	36	MC	OP	1	2	C.10.14	Civics
10	2	37	Multi-Select	OP	1	2	E.10.3	Behavioral Sciences
10	2	38	MC	OP	1	2	C.10.6	Civics
10	2	39	MC	OP	1	1	B.10.14	History
10	2	40	MC	OP	1	2	E.10.12	Behavioral Sciences
10	2	41	MC	OP	1	3	A.10.5	Geography
10	2	42	MC	OP	1	N/A	C.10.12	Civics
10	2	43	MC	OP	1	3	B.10.15	History
10	2	44	MC	OP	1	2	A.10.12	Geography
10	2	45	MC	OP	1	2	B.10.3	History
10	2	46	MC	OP	1	2	E.10.17	Behavioral Sciences
10	2	47	MC	OP	1	2	C.10.11	Civics
10	2	48	MC	OP	1	1	A.10.8	Geography
10	2	49	MC	OP	1	N/A	B.10.15	History
10	2	50	MC	OP	1	2	A.10.5	Geography

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	Proportion Omit	Proportion 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	TDA	4	62923	0.34	0.53	0.00	0.10	0.52	0.32	0.06	0.00	-0.36	-0.23	0.34	0.24	0.08
2	MC	1	63045	0.67	0.31	0.00		0.12	0.17	0.67	0.04		-0.17	-0.14	0.31	-0.18
3	MC	1	63004	0.59	0.25	0.00		0.09	0.21	0.11	0.59		-0.10	-0.08	-0.21	0.26
4	TE	2	63020	0.70	0.50	0.00	0.10	0.39	0.50			-0.33	-0.26	0.46		
5	MC	1	63011	0.65	0.32	0.00		0.65	0.16	0.11	0.08		0.32	-0.13	-0.16	-0.20
6	MC	1	62996	0.83	0.45	0.00		0.08	0.82	0.05	0.04		-0.28	0.45	-0.23	-0.21
7	MC	1	62991	0.66	0.37	0.00		0.08	0.08	0.66	0.17		-0.29	-0.19	0.37	-0.11
8	TE	2	62948	0.57	0.45	0.00	0.20	0.44	0.35			-0.29	-0.17	0.43		
9	MC	1	62992	0.60	0.43	0.00		0.16	0.13	0.11	0.60		-0.24	-0.15	-0.23	0.44
10	MC	1	62984	0.54	0.32	0.00		0.13	0.54	0.14	0.19		-0.25	0.32	-0.13	-0.07
11	MC	1	62933	0.49	0.31	0.00		0.49	0.17	0.20	0.13		0.31	-0.16	-0.09	-0.16
12	MC	1	62974	0.39	0.39	0.00		0.39	0.14	0.08	0.39		-0.25	-0.16	-0.05	0.39
13	MC	1	62974	0.72	0.54	0.00		0.09	0.13	0.07	0.71		-0.28	-0.30	-0.26	0.55
14	TE	2	62990	0.67	0.53	0.00	0.15	0.37	0.48			-0.41	-0.19	0.48		
15	MC	1	63046	0.71	0.36	0.00		0.11	0.71	0.10	0.08		-0.16	0.36	-0.19	-0.20
16	TE	2	63005	0.65	0.53	0.00	0.12	0.44	0.43			-0.38	-0.21	0.47		
17	MC	1	62990	0.73	0.46	0.00		0.10	0.72	0.11	0.06		-0.26	0.46	-0.21	-0.25
18	ESR	2	63032	0.58	0.52	0.00	0.37	0.09	0.54			-0.46	-0.14	0.53		
19	MC	1	62995	0.56	0.31	0.00		0.29	0.10	0.04	0.56		-0.10	-0.20	-0.24	0.31
20	MC	1	62929	0.62	0.39	0.00		0.62	0.21	0.09	0.08		0.40	-0.23	-0.18	-0.17

Note: TDA responses that received a condition code were converted to 0 and are reported as a score of 0.

Note: TDA item is weighted x 2 in computation of student scores.

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	Proportion Omit	Proportion 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	1	61741	0.41	0.54	0.02	0.57	0.40				-0.50	0.54			
22	MC	1	62966	0.81	0.31	0.00		0.08	0.81	0.05	0.06		-0.09	0.31	-0.24	-0.18
23	MC	1	62940	0.57	0.31	0.00		0.11	0.16	0.57	0.17		-0.25	-0.08	0.31	-0.12
24	ESR	2	63015	0.43	0.40	0.00	0.45	0.23	0.31			-0.32	-0.07	0.41		
25	MC	1	62925	0.54	0.24	0.00		0.21	0.04	0.21	0.54		-0.07	-0.21	-0.12	0.25
26	MC	1	62755	0.66	0.41	0.01		0.20	0.65	0.09	0.05		-0.23	0.41	-0.15	-0.24
27	MC	1	62658	0.55	0.22	0.01		0.55	0.15	0.18	0.12		0.23	0.05	-0.09	-0.28
28	TE	2	62911	0.67	0.65	0.00	0.22	0.21	0.56			-0.51	-0.25	0.64		
29	MC	1	62909	0.47	0.37	0.00		0.17	0.47	0.22	0.14		-0.20	0.37	-0.09	-0.21
30	MC	1	62734	0.45	0.40	0.01		0.21	0.12	0.44	0.22		-0.09	-0.18	0.40	-0.24
31	MC	1	62615	0.59	0.52	0.01		0.59	0.18	0.15	0.07		0.52	-0.20	-0.29	-0.27
32	MC	1	62900	0.48	0.34	0.00		0.18	0.48	0.07	0.27		-0.06	0.34	-0.23	-0.19
33	ESR	2	62959	0.47	0.54	0.00	0.47	0.12	0.40			-0.47	-0.09	0.55		
34	TE	2	62869	0.68	0.53	0.00	0.11	0.43	0.46			-0.27	-0.37	0.55		
35	MC	1	62871	0.41	0.33	0.00		0.34	0.41	0.08	0.16		-0.02	0.33	-0.27	-0.20
36	MC	1	62890	0.49	0.35	0.00		0.23	0.11	0.16	0.49		-0.14	-0.19	-0.15	0.36
37	MC	1	62876	0.35	0.23	0.00		0.20	0.35	0.22	0.23		-0.06	0.23	-0.25	0.05

Table G-2. Item Statistics, ELA Grade 4

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Proportion Omit	Proportion 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	TDA	4	64143	0.35	0.52	0.00	0.08	0.54	0.30	0.06	0.01	-0.38	-0.23	0.33	0.22	0.13
2	MC	1	64214	0.39	0.22	0.00		0.13	0.37	0.39	0.11		-0.17	0.01	0.22	-0.17
3	TE	1	64178	0.64	0.43	0.00	0.36	0.64				-0.42	0.43			
4	MC	1	64192	0.75	0.36	0.00		0.75	0.18	0.04	0.03		0.36	-0.29	-0.07	-0.18
5	MC	1	64164	0.68	0.35	0.00		0.19	0.68	0.09	0.04		-0.24	0.35	-0.13	-0.14
6	TE	1	64119	0.47	0.45	0.00	0.53	0.47				-0.44	0.45			
7	TE	2	64130	0.63	0.46	0.00	0.12	0.50	0.38			-0.33	-0.17	0.40		
8	MC	1	64151	0.58	0.46	0.00		0.14	0.21	0.58	0.07		-0.25	-0.24	0.46	-0.16
9	TE	2	64163	0.53	0.48	0.00	0.18	0.57	0.24			-0.40	0.00	0.36		
10	MC	1	64160	0.57	0.18	0.00		0.57	0.24	0.09	0.10		0.18	0.03	-0.17	-0.18
11	TE	2	63963	0.70	0.41	0.00	0.06	0.49	0.45			-0.17	-0.31	0.41		
12	MC	1	64163	0.70	0.46	0.00		0.70	0.10	0.10	0.10		0.46	-0.26	-0.21	-0.23
13	TE	2	64180	0.80	0.35	0.00	0.03	0.35	0.62			-0.21	-0.25	0.32		
14	MC	1	64229	0.58	0.28	0.00		0.11	0.19	0.58	0.12		-0.12	-0.09	0.28	-0.21
15	MC	1	64178	0.76	0.34	0.00		0.02	0.19	0.03	0.76		-0.18	-0.23	-0.19	0.34
16	ESR	2	64240	0.33	0.34	0.00	0.56	0.23	0.21			-0.31	0.08	0.29		
17	ESR	2	64232	0.62	0.44	0.00	0.23	0.31	0.46			-0.39	-0.05	0.37		
18	MC	1	64184	0.74	0.39	0.00		0.08	0.74	0.08	0.11		-0.20	0.39	-0.23	-0.18
19	MC	1	64175	0.38	0.29	0.00		0.12	0.24	0.26	0.38		-0.11	-0.14	-0.10	0.29
20	MC	1	64159	0.47	0.43	0.00		0.13	0.22	0.18	0.47		-0.33	-0.08	-0.17	0.43

Note: TDA responses that received a condition code were converted to 0 and are reported as a score of 0.

Note: TDA item is weighted x 2 in computation of student scores.

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	Proportion Omit	Proportion 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	64146	0.78	0.49	0.00		0.09	0.07	0.78	0.06		-0.31	-0.26	0.49	-0.20
22	TE	2	64104	0.58	0.58	0.00	0.38	0.09	0.53			-0.53	-0.10	0.58		
23	MC	1	64124	0.60	0.43	0.00		0.17	0.60	0.14	0.08		-0.25	0.43	-0.23	-0.12
24	TE	2	64076	0.54	0.49	0.00	0.21	0.50	0.29			-0.38	-0.06	0.42		
25	MC	1	64093	0.64	0.51	0.00		0.13	0.63	0.12	0.11		-0.22	0.51	-0.27	-0.25
26	MC	1	64070	0.59	0.41	0.00		0.08	0.12	0.21	0.59		-0.21	-0.27	-0.12	0.41
27	MC	1	64032	0.77	0.42	0.00		0.10	0.76	0.06	0.07		-0.19	0.43	-0.28	-0.20
28	MC	1	64096	0.71	0.51	0.00		0.14	0.70	0.09	0.07		-0.25	0.51	-0.28	-0.25
29	TE	1	63873	0.26	0.49	0.01	0.74	0.26				-0.47	0.49			
30	TE	2	64128	0.68	0.50	0.00	0.10	0.43	0.47			-0.34	-0.25	0.46		
31	MC	1	63984	0.53	0.40	0.00		0.17	0.15	0.53	0.15		-0.10	-0.25	0.40	-0.20
32	MC	1	63702	0.50	0.37	0.01		0.20	0.17	0.50	0.12		-0.16	-0.14	0.37	-0.18
33	MC	1	64052	0.54	0.39	0.00		0.54	0.18	0.12	0.15		0.39	-0.21	-0.24	-0.09
34	MC	1	64080	0.42	0.27	0.00		0.09	0.26	0.23	0.42		-0.17	-0.06	-0.13	0.27
35	MC	1	64062	0.59	0.49	0.00		0.59	0.16	0.15	0.10		0.50	-0.22	-0.25	-0.24
36	MC	1	64063	0.57	0.49	0.00		0.16	0.16	0.11	0.57		-0.24	-0.21	-0.23	0.49
37	TE	2	64080	0.46	0.41	0.00	0.24	0.58	0.17			-0.34	0.07	0.31		
38	MC	1	64080	0.55	0.42	0.00		0.55	0.18	0.15	0.12		0.42	-0.20	-0.23	-0.15
39	MC	1	64093	0.60	0.43	0.00		0.09	0.14	0.60	0.17		-0.22	-0.17	0.44	-0.24

Table G-3. Item Statistics, ELA Grade 5

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Proportion Omit	Proportion 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	TDA	4	64740	0.32	0.57	0.00	0.16	0.48	0.29	0.07	0.01	-0.46	-0.12	0.34	0.24	0.11
2	MC	1	64772	0.66	0.23	0.00		0.05	0.06	0.66	0.23		-0.13	-0.05	0.23	-0.16
3	MC	1	64756	0.70	0.28	0.00		0.07	0.70	0.13	0.09		-0.15	0.28	-0.14	-0.14
4	TE	1	64696	0.50	0.30	0.00	0.50	0.50				-0.29	0.30			
5	MC	1	64746	0.79	0.45	0.00		0.03	0.79	0.02	0.16		-0.14	0.45	-0.18	-0.36
6	TE	2	64722	0.58	0.44	0.00	0.17	0.49	0.33			-0.32	-0.11	0.38		
7	TE	2	64626	0.52	0.51	0.00	0.33	0.29	0.37			-0.35	-0.20	0.54		
8	MC	1	64737	0.63	0.43	0.00		0.24	0.05	0.63	0.07		-0.22	-0.27	0.44	-0.20
9	MC	1	64746	0.66	0.40	0.00		0.17	0.66	0.11	0.06		-0.19	0.40	-0.20	-0.24
10	MC	1	64708	0.51	0.31	0.00		0.14	0.20	0.50	0.15		-0.24	-0.10	0.31	-0.08
11	MC	1	64705	0.76	0.45	0.00		0.10	0.06	0.76	0.08		-0.21	-0.25	0.45	-0.25
12	MC	1	64724	0.58	0.31	0.00		0.03	0.30	0.09	0.58		-0.23	-0.14	-0.18	0.31
13	TE	2	64742	0.71	0.40	0.00	0.12	0.34	0.53			-0.35	-0.10	0.33		
14	MC	1	64740	0.85	0.41	0.00		0.08	0.84	0.04	0.03		-0.29	0.41	-0.21	-0.15
15	MC	1	64772	0.74	0.44	0.00		0.74	0.04	0.13	0.08		0.44	-0.21	-0.21	-0.28
16	MC	1	64740	0.71	0.50	0.00		0.11	0.71	0.08	0.11		-0.25	0.50	-0.21	-0.30
17	TE	2	64741	0.62	0.54	0.00	0.16	0.44	0.39			-0.39	-0.18	0.48		
18	TE	2	64742	0.61	0.45	0.00	0.15	0.49	0.37			-0.35	-0.11	0.38		
19	MC	1	64730	0.63	0.39	0.00		0.63	0.29	0.04	0.04		0.39	-0.24	-0.19	-0.21
20	MC	1	64727	0.77	0.46	0.00		0.77	0.04	0.09	0.09		0.46	-0.25	-0.27	-0.21

Note: TDA responses that received a condition code were converted to 0 and are reported as a score of 0.

Note: TDA item is weighted x 2 in computation of student scores.

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	Proportion Omit	Proportion 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	64705	0.50	0.28	0.00		0.50	0.16	0.27	0.06		0.28	-0.19	-0.11	-0.07
22	TE	2	64718	0.66	0.54	0.00	0.13	0.42	0.45			-0.39	-0.22	0.48		
23	MC	1	64669	0.69	0.40	0.00		0.19	0.06	0.06	0.68		-0.19	-0.24	-0.21	0.40
24	MC	1	64637	0.61	0.42	0.00		0.61	0.12	0.07	0.20		0.42	-0.20	-0.25	-0.18
25	TE	2	64520	0.72	0.17	0.00	0.03	0.50	0.46			-0.15	-0.08	0.14		
26	MC	1	64635	0.58	0.39	0.00		0.15	0.12	0.57	0.16		-0.19	-0.20	0.39	-0.16
27	MC	1	64625	0.64	0.48	0.00		0.15	0.13	0.64	0.08		-0.27	-0.25	0.48	-0.18
28	MC	1	64492	0.50	0.48	0.00		0.13	0.27	0.10	0.50		-0.27	-0.16	-0.26	0.48
29	MC	1	64626	0.67	0.42	0.00		0.67	0.17	0.10	0.06		0.42	-0.22	-0.21	-0.21
30	MC	1	64632	0.58	0.46	0.00		0.13	0.09	0.20	0.58		-0.21	-0.23	-0.23	0.46
31	MC	1	64634	0.58	0.40	0.00		0.22	0.58	0.10	0.09		-0.18	0.41	-0.23	-0.19
32	MC	1	64655	0.51	0.31	0.00		0.14	0.22	0.51	0.12		-0.22	-0.03	0.32	-0.21
33	MC	1	64497	0.73	0.49	0.00		0.09	0.08	0.11	0.72		-0.26	-0.23	-0.26	0.49
34	TE	2	64279	0.42	0.39	0.01	0.34	0.48	0.17			-0.27	-0.02	0.38		
35	MC	1	64571	0.50	0.33	0.00		0.27	0.50	0.14	0.08		-0.05	0.33	-0.18	-0.28
36	MC	1	64572	0.40	0.34	0.00		0.20	0.40	0.17	0.23		-0.11	0.34	-0.19	-0.12
37	MC	1	64599	0.50	0.50	0.00		0.20	0.18	0.12	0.50		-0.20	-0.25	-0.21	0.50
38	MC	1	64584	0.52	0.31	0.00		0.26	0.11	0.11	0.52		0.00	-0.21	-0.27	0.31
39	MC	1	64593	0.50	0.45	0.00		0.50	0.16	0.24	0.10		0.45	-0.17	-0.22	-0.23
40	MC	1	64596	0.54	0.33	0.00		0.13	0.54	0.14	0.19		-0.11	0.33	-0.25	-0.10
41	MC	1	64618	0.39	0.22	0.00		0.14	0.22	0.39	0.25		-0.13	-0.05	0.22	-0.08

Table G-4. Item Statistics, ELA Grade 6

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Proportion Omit	Proportion 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	TDA	4	63374	0.38	0.53	0.00	0.09	0.43	0.35	0.12	0.01	-0.38	-0.24	0.26	0.28	0.11
2	MC	1	63428	0.70	0.38	0.00		0.70	0.18	0.10	0.03		0.38	-0.26	-0.17	-0.14
3	MC	1	63391	0.68	0.24	0.00		0.68	0.07	0.13	0.12		0.24	-0.13	-0.13	-0.11
4	MC	1	63378	0.73	0.46	0.00		0.04	0.13	0.73	0.10		-0.20	-0.29	0.46	-0.22
5	MC	1	63371	0.57	0.30	0.00		0.57	0.25	0.15	0.03		0.30	-0.19	-0.11	-0.16
6	TE	2	63377	0.62	0.27	0.00	0.08	0.61	0.31			-0.29	0.00	0.17		
7	TE	2	63366	0.75	0.48	0.00	0.08	0.33	0.59			-0.35	-0.24	0.43		
8	MC	1	63386	0.46	0.32	0.00		0.46	0.06	0.21	0.27		0.32	-0.22	-0.14	-0.11
9	TE	1	63358	0.51	0.28	0.00	0.49	0.51				-0.28	0.28			
10	MC	1	63345	0.37	0.12	0.00		0.13	0.36	0.25	0.25		-0.20	0.12	-0.13	0.16
11	TE	2	63348	0.61	0.42	0.00	0.14	0.49	0.37			-0.36	-0.06	0.33		
12	ESR	2	63371	0.43	0.36	0.00	0.51	0.10	0.38			-0.28	-0.19	0.41		
13	MC	1	63366	0.52	0.35	0.00		0.03	0.27	0.52	0.17		-0.23	-0.14	0.36	-0.19
14	MC	1	63408	0.56	0.34	0.00		0.07	0.21	0.56	0.16		-0.15	-0.34	0.34	0.02
15	MC	1	63352	0.57	0.33	0.00		0.06	0.57	0.23	0.14		-0.13	0.33	-0.25	-0.08
16	TE	2	63356	0.61	0.39	0.00	0.16	0.47	0.37			-0.28	-0.13	0.35		
17	ESR	2	63397	0.71	0.53	0.00	0.17	0.23	0.59			-0.43	-0.19	0.50		
18	MC	1	63328	0.25	0.18	0.00		0.21	0.14	0.25	0.40		-0.20	-0.07	0.19	0.05
19	MC	1	63336	0.77	0.48	0.00		0.05	0.13	0.06	0.76		-0.21	-0.29	-0.27	0.48
20	MC	1	63329	0.45	0.21	0.00		0.25	0.27	0.03	0.45		-0.08	-0.04	-0.26	0.21

Note: TDA responses that received a condition code were converted to 0 and are reported as a score of 0.

Note: TDA item is weighted x 2 in computation of student scores.

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	Proportion Omit	Proportion 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	63328	0.66	0.46	0.00	0.12	0.45	0.43			-0.37	-0.14	0.39		
22	MC	1	63289	0.51	0.34	0.00		0.08	0.51	0.20	0.21		-0.21	0.34	-0.23	-0.04
23	TE	2	63256	0.59	0.44	0.00	0.14	0.52	0.33			-0.37	-0.06	0.35		
24	MC	1	63266	0.55	0.33	0.00		0.25	0.09	0.55	0.11		-0.14	-0.20	0.33	-0.15
25	MC	1	63261	0.81	0.51	0.00		0.07	0.04	0.08	0.80		-0.28	-0.28	-0.27	0.51
26	TE	2	63289	0.57	0.43	0.00	0.16	0.54	0.30			-0.31	-0.11	0.38		
27	MC	1	63219	0.62	0.45	0.00		0.05	0.22	0.11	0.61		-0.19	-0.16	-0.34	0.45
28	TE	2	63181	0.63	0.39	0.00	0.13	0.48	0.39			-0.30	-0.11	0.33		
29	MC	1	63305	0.37	0.31	0.00		0.13	0.06	0.44	0.37		-0.18	-0.20	-0.08	0.31
30	MC	1	63296	0.74	0.46	0.00		0.74	0.10	0.06	0.10		0.47	-0.28	-0.29	-0.17
31	MC	1	63175	0.74	0.43	0.00		0.09	0.06	0.73	0.11		-0.20	-0.23	0.43	-0.24
32	TE	2	63199	0.54	0.46	0.00	0.20	0.51	0.28			-0.37	-0.03	0.37		
33	MC	1	63254	0.68	0.43	0.00		0.18	0.09	0.68	0.05		-0.18	-0.31	0.43	-0.17
34	TE	2	63252	0.66	0.51	0.00	0.14	0.39	0.46			-0.39	-0.18	0.45		
35	TE	1	63239	0.58	0.39	0.00	0.42	0.58				-0.38	0.39			
36	MC	1	63268	0.51	0.34	0.00		0.21	0.18	0.51	0.09		-0.10	-0.12	0.35	-0.28
37	ESR	2	63338	0.58	0.50	0.00	0.31	0.23	0.46			-0.43	-0.07	0.46		

Table G-5. Item Statistics, ELA Grade 7

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Proportion Omit	Proportion 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	TDA	4	62939	0.42	0.56	0.00	0.09	0.36	0.38	0.14	0.03	-0.36	-0.29	0.21	0.31	0.19
2	TE	1	62930	0.64	0.27	0.00	0.36	0.64				-0.27	0.27			
3	TE	2	62960	0.52	0.38	0.00	0.26	0.44	0.30			-0.29	-0.05	0.34		
4	MC	1	62956	0.72	0.25	0.00		0.08	0.18	0.02	0.72		-0.16	-0.09	-0.22	0.25
5	MC	1	62954	0.75	0.39	0.00		0.06	0.74	0.12	0.08		-0.25	0.39	-0.20	-0.17
6	TE	2	62880	0.66	0.32	0.00	0.07	0.55	0.38			-0.17	-0.21	0.31		
7	MC	1	62889	0.79	0.46	0.00		0.79	0.06	0.06	0.09		0.47	-0.27	-0.26	-0.22
8	MC	1	62938	0.78	0.37	0.00		0.78	0.09	0.08	0.05		0.37	-0.24	-0.27	-0.05
9	MC	1	62937	0.51	0.39	0.00		0.20	0.13	0.51	0.16		-0.13	-0.32	0.39	-0.09
10	MC	1	62944	0.64	0.38	0.00		0.09	0.04	0.64	0.23		-0.25	-0.25	0.38	-0.14
11	MC	1	62926	0.61	0.29	0.00		0.18	0.12	0.61	0.09		-0.11	-0.24	0.29	-0.08
12	ESR	2	62947	0.51	0.48	0.00	0.38	0.22	0.40			-0.42	-0.04	0.45		
13	MC	1	62913	0.47	0.30	0.00		0.18	0.15	0.47	0.20		-0.10	-0.19	0.30	-0.10
14	MC	1	62914	0.61	0.44	0.00		0.07	0.10	0.22	0.60		-0.21	-0.34	-0.14	0.44
15	TE	2	62990	0.77	0.45	0.00	0.03	0.40	0.57			-0.20	-0.37	0.43		
16	MC	1	62911	0.71	0.47	0.00		0.03	0.11	0.71	0.15		-0.24	-0.24	0.47	-0.27
17	MC	1	62900	0.59	0.40	0.00		0.18	0.59	0.15	0.08		-0.19	0.40	-0.21	-0.17
18	MC	1	62936	0.50	0.43	0.00		0.49	0.23	0.12	0.15		0.43	-0.09	-0.24	-0.27
19	MC	1	62917	0.57	0.33	0.00		0.09	0.18	0.17	0.57		-0.10	-0.22	-0.14	0.33
20	ESR	2	62960	0.38	0.43	0.00	0.52	0.19	0.29			-0.35	-0.05	0.43		

Note: TDA responses that received a condition code were converted to 0 and are reported as a score of 0.

Note: TDA item is weighted x 2 in computation of student scores.

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	Proportion Omit	Proportion 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	1	62770	0.74	0.35	0.00	0.26	0.74				-0.34	0.35			
22	MC	1	62861	0.60	0.52	0.00		0.12	0.21	0.08	0.59		-0.30	-0.20	-0.27	0.52
23	MC	1	62878	0.76	0.41	0.00		0.76	0.07	0.05	0.12		0.41	-0.22	-0.26	-0.18
24	ESR	2	62947	0.50	0.38	0.00	0.41	0.17	0.42			-0.30	-0.13	0.40		
25	MC	1	62839	0.50	0.35	0.00		0.23	0.49	0.18	0.09		-0.10	0.35	-0.16	-0.25
26	MC	1	62843	0.48	0.21	0.00		0.31	0.14	0.48	0.06		0.07	-0.24	0.21	-0.21
27	MC	1	62707	0.58	0.48	0.00		0.06	0.16	0.19	0.58		-0.24	-0.18	-0.28	0.48
28	TE	1	60994	0.64	0.44	0.03	0.35	0.62				-0.39	0.46			
29	TE	2	62753	0.66	0.56	0.00	0.15	0.37	0.47			-0.45	-0.17	0.49		
30	MC	1	62805	0.42	0.28	0.00		0.17	0.30	0.42	0.10		-0.08	-0.11	0.28	-0.19
31	TE	2	62808	0.52	0.33	0.00	0.18	0.60	0.22			-0.34	0.11	0.19		
32	ESR	2	62808	0.55	0.55	0.00	0.40	0.10	0.50			-0.48	-0.15	0.57		
33	MC	1	62768	0.57	0.51	0.00		0.57	0.15	0.15	0.13		0.51	-0.23	-0.24	-0.24
34	MC	1	62783	0.66	0.46	0.00		0.17	0.66	0.07	0.10		-0.18	0.46	-0.29	-0.24
35	MC	1	62780	0.51	0.48	0.00		0.10	0.25	0.14	0.50		-0.20	-0.17	-0.30	0.48
36	MC	1	62782	0.59	0.49	0.00		0.59	0.14	0.17	0.10		0.49	-0.30	-0.22	-0.17
37	TE	2	62700	0.71	0.42	0.01	0.08	0.41	0.50			-0.34	-0.15	0.35		
38	MC	1	62803	0.48	0.42	0.00		0.11	0.09	0.31	0.48		-0.15	-0.28	-0.17	0.43
39	MC	1	62799	0.60	0.43	0.00		0.60	0.11	0.14	0.15		0.43	-0.22	-0.25	-0.15

Table G-6. Item Statistics, ELA Grade 8

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Proportion Omit	Proportion 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	TDA	4	62977	0.47	0.61	0.00	0.10	0.24	0.36	0.24	0.05	-0.42	-0.28	0.10	0.35	0.24
2	MC	1	63076	0.64	0.34	0.00		0.15	0.07	0.64	0.14		-0.18	-0.19	0.34	-0.15
3	TE	2	62930	0.61	0.32	0.00	0.04	0.70	0.26			-0.12	-0.23	0.30		
4	MC	1	63032	0.68	0.47	0.00		0.10	0.12	0.11	0.68		-0.19	-0.27	-0.23	0.47
5	MC	1	62990	0.22	0.18	0.00		0.14	0.21	0.22	0.43		-0.19	-0.01	0.18	-0.01
6	TE	2	62703	0.52	0.48	0.01	0.18	0.59	0.22			-0.39	0.01	0.37		
7	MC	1	62976	0.73	0.47	0.00		0.07	0.09	0.11	0.73		-0.24	-0.24	-0.24	0.47
8	MC	1	62997	0.68	0.38	0.00		0.10	0.08	0.68	0.13		-0.18	-0.22	0.38	-0.18
9	MC	1	62997	0.81	0.35	0.00		0.08	0.81	0.03	0.09		-0.17	0.35	-0.23	-0.19
10	MC	1	62991	0.55	0.40	0.00		0.13	0.20	0.12	0.55		-0.21	-0.14	-0.22	0.40
11	TE	1	62769	0.49	0.35	0.01	0.50	0.49				-0.34	0.35			
12	MC	1	62942	0.54	0.29	0.00		0.14	0.54	0.18	0.13		-0.16	0.29	-0.16	-0.08
13	MC	1	62959	0.69	0.44	0.00		0.14	0.69	0.10	0.07		-0.17	0.44	-0.30	-0.22
14	MC	1	62976	0.57	0.43	0.00		0.13	0.11	0.19	0.57		-0.28	-0.22	-0.12	0.43
15	MC	1	62974	0.53	0.39	0.00		0.12	0.09	0.26	0.53		-0.15	-0.29	-0.13	0.39
16	MC	1	63042	0.60	0.25	0.00		0.20	0.14	0.60	0.05		-0.04	-0.22	0.25	-0.13
17	MC	1	62987	0.87	0.41	0.00		0.04	0.87	0.07	0.02		-0.20	0.41	-0.27	-0.21
18	ESR	2	63049	0.55	0.52	0.00	0.31	0.28	0.42			-0.42	-0.11	0.49		
19	ESR	2	63039	0.55	0.47	0.00	0.40	0.10	0.50			-0.41	-0.12	0.48		
20	MC	1	62985	0.70	0.49	0.00		0.09	0.12	0.70	0.09		-0.23	-0.27	0.49	-0.24

Note: TDA responses that received a condition code were converted to 0 and are reported as a score of 0.

Note: TDA item is weighted x 2 in computation of student scores.

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	Proportion Omit	Proportion 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	62997	0.47	0.36	0.00		0.47	0.08	0.25	0.20		0.36	-0.18	-0.20	-0.11
22	TE	1	62640	0.47	0.38	0.01	0.53	0.47				-0.36	0.38			
23	MC	1	62843	0.56	0.54	0.00		0.26	0.10	0.07	0.56		-0.23	-0.26	-0.32	0.54
24	ESR	2	62986	0.46	0.41	0.00	0.52	0.04	0.44			-0.36	-0.22	0.45		
25	MC	1	62857	0.63	0.45	0.00		0.16	0.10	0.63	0.11		-0.24	-0.28	0.46	-0.15
26	MC	1	62893	0.62	0.25	0.00		0.15	0.12	0.61	0.12		-0.07	-0.02	0.25	-0.27
27	MC	1	62900	0.60	0.49	0.00		0.03	0.05	0.32	0.60		-0.20	-0.30	-0.29	0.49
28	MC	1	62879	0.50	0.27	0.00		0.17	0.50	0.17	0.16		-0.12	0.27	-0.17	-0.05
29	TE	2	62782	0.69	0.57	0.00	0.09	0.44	0.47			-0.33	-0.35	0.55		
30	MC	1	62770	0.76	0.54	0.00		0.75	0.11	0.09	0.05		0.55	-0.30	-0.33	-0.21
31	MC	1	62795	0.47	0.24	0.00		0.06	0.29	0.17	0.47		-0.17	-0.16	-0.02	0.25
32	ESR	2	62952	0.61	0.53	0.00	0.29	0.20	0.50			-0.43	-0.17	0.53		
33	MC	1	62727	0.54	0.39	0.01		0.54	0.19	0.21	0.06		0.39	-0.24	-0.09	-0.24
34	ESR	2	62915	0.58	0.63	0.00	0.28	0.28	0.44			-0.58	0.01	0.53		
35	MC	1	62825	0.78	0.53	0.00		0.06	0.77	0.09	0.07		-0.22	0.53	-0.32	-0.27
36	MC	1	62843	0.67	0.44	0.00		0.08	0.10	0.67	0.14		-0.18	-0.21	0.45	-0.27
37	MC	1	62833	0.61	0.35	0.00		0.04	0.61	0.06	0.29		-0.27	0.35	-0.29	-0.10
38	TE	2	62838	0.57	0.42	0.00	0.08	0.69	0.22			-0.32	-0.09	0.32		
39	MC	1	62828	0.75	0.49	0.00		0.07	0.75	0.07	0.11		-0.26	0.50	-0.28	-0.23
40	MC	1	62844	0.59	0.43	0.00		0.59	0.19	0.13	0.09		0.43	-0.16	-0.21	-0.26

Table G-7. Item Statistics, Mathematics Grade 3

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Proportion Omit	Proportion 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	63182	0.64	0.46	0.00		0.21	0.64	0.10	0.05		-0.26	0.46	-0.26	-0.17
2	MC	1	63197	0.74	0.44	0.00		0.04	0.04	0.74	0.18		-0.16	-0.16	0.44	-0.34
3	MC	1	63174	0.75	0.48	0.00		0.15	0.03	0.75	0.07		-0.43	-0.15	0.48	-0.11
4	MC	1	63173	0.76	0.45	0.00		0.05	0.08	0.76	0.12		-0.13	-0.33	0.46	-0.24
5	MC	1	63174	0.53	0.46	0.00		0.25	0.05	0.17	0.53		-0.28	-0.08	-0.24	0.47
6	TE	1	63028	0.43	0.61	0.00	0.57	0.43				-0.60	0.61			
7	SA	1	62883	0.84	0.39	0.01	0.16	0.83				-0.38	0.39			
8	MC	1	63164	0.19	0.30	0.00		0.04	0.65	0.19	0.12		-0.19	-0.16	0.30	0.00
9	MC	1	63137	0.67	0.40	0.00		0.67	0.11	0.13	0.09		0.40	-0.26	-0.13	-0.21
10	SA	1	63172	0.61	0.56	0.00	0.39	0.61				-0.56	0.56			
11	MC	1	63135	0.44	0.36	0.00		0.19	0.18	0.43	0.18		-0.11	-0.13	0.36	-0.21
12	SA	1	63157	0.21	0.48	0.00	0.79	0.21				-0.47	0.48			
13	MC	1	62930	0.62	0.42	0.01		0.62	0.10	0.19	0.09		0.42	-0.12	-0.27	-0.21
14	MC	1	62869	0.47	0.32	0.01		0.32	0.08	0.13	0.47		-0.04	-0.27	-0.20	0.32
15	MC	1	63093	0.43	0.32	0.00		0.31	0.43	0.14	0.12		-0.21	0.32	-0.11	-0.05
16	SA	1	63101	0.33	0.51	0.00	0.67	0.32				-0.50	0.51			
17	MC	1	63148	0.40	0.44	0.00		0.32	0.06	0.21	0.40		-0.20	-0.25	-0.15	0.44
18	MC	1	63136	0.59	0.42	0.00		0.19	0.59	0.10	0.12		-0.21	0.42	-0.14	-0.26
19	MC	1	63144	0.72	0.49	0.00		0.05	0.10	0.13	0.72		-0.24	-0.27	-0.26	0.49
20	TE	1	62559	0.87	0.36	0.01	0.13	0.86				-0.35	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	Proportion Omit	Proportion 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	63164	0.82	0.45	0.00		0.82	0.10	0.05	0.03		0.45	-0.31	-0.22	-0.20
22	MC	1	63159	0.47	0.39	0.00		0.47	0.33	0.08	0.13		0.39	-0.06	-0.16	-0.37
23	MC	1	63143	0.44	0.39	0.00		0.21	0.44	0.09	0.25		-0.22	0.39	-0.20	-0.10
24	MC	1	63093	0.62	0.45	0.00		0.15	0.12	0.62	0.12		-0.23	-0.24	0.45	-0.19
25	MC	1	63134	0.61	0.39	0.00		0.14	0.15	0.61	0.10		-0.23	-0.19	0.39	-0.15
26	SA	1	63158	0.64	0.57	0.00	0.36	0.64				-0.57	0.57			
27	TE	1	63011	0.63	0.33	0.00	0.37	0.63				-0.33	0.33			
28	MC	1	62817	0.68	0.45	0.01		0.06	0.05	0.21	0.68		-0.25	-0.24	-0.24	0.45
29	MC	1	63065	0.42	0.38	0.00		0.13	0.14	0.31	0.41		-0.27	-0.20	-0.05	0.38
30	MC	1	63099	0.39	0.20	0.00		0.20	0.24	0.16	0.39		-0.10	0.00	-0.15	0.20
31	MC	1	63113	0.61	0.49	0.00		0.06	0.29	0.61	0.03		-0.27	-0.30	0.49	-0.20
32	MC	1	63081	0.41	0.39	0.00		0.30	0.14	0.15	0.41		-0.14	-0.19	-0.16	0.39
33	SA	1	63087	0.52	0.56	0.00	0.48	0.52				-0.56	0.56			
34	MC	1	62848	0.49	0.48	0.01		0.19	0.14	0.18	0.49		-0.26	-0.22	-0.15	0.48
35	MC	1	62851	0.66	0.51	0.01		0.21	0.66	0.07	0.05		-0.39	0.51	-0.19	-0.14
36	TE	1	62801	0.24	0.47	0.01	0.76	0.24				-0.45	0.47			
37	MC	1	63095	0.65	0.49	0.00		0.06	0.65	0.11	0.18		-0.27	0.49	-0.20	-0.27
38	SA	1	63084	0.59	0.60	0.00	0.41	0.59				-0.59	0.60			
39	MC	1	63106	0.80	0.43	0.00		0.11	0.79	0.05	0.04		-0.27	0.44	-0.23	-0.19
40	MC	1	63102	0.75	0.39	0.00		0.10	0.03	0.13	0.74		-0.13	-0.10	-0.34	0.39
41	MC	1	63074	0.52	0.57	0.00		0.07	0.52	0.36	0.04		-0.17	0.57	-0.42	-0.18
42	TE	1	62264	0.35	0.47	0.02	0.64	0.34				-0.43	0.47			

Table G-8. Item Statistics, Mathematics Grade 4

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Proportion Omit	Proportion 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	64183	0.68	0.45	0.00		0.68	0.16	0.07	0.09		0.45	-0.29	-0.21	-0.18
2	MC	1	64342	0.46	0.30	0.00		0.05	0.43	0.06	0.46		-0.20	-0.13	-0.16	0.30
3	MC	1	64347	0.57	0.49	0.00		0.57	0.12	0.16	0.16		0.49	-0.15	-0.26	-0.27
4	SA	1	64274	0.19	0.48	0.00	0.81	0.19				-0.47	0.48			
5	MC	1	64344	0.72	0.45	0.00		0.72	0.09	0.14	0.05		0.45	-0.26	-0.22	-0.23
6	TE	1	64230	0.38	0.49	0.00	0.62	0.38				-0.49	0.49			
7	MC	1	64308	0.64	0.41	0.00		0.08	0.64	0.12	0.15		-0.10	0.42	-0.17	-0.31
8	MC	1	64173	0.64	0.37	0.00		0.18	0.10	0.64	0.08		-0.20	-0.15	0.38	-0.21
9	MC	1	64332	0.73	0.39	0.00		0.19	0.05	0.03	0.73		-0.28	-0.18	-0.16	0.39
10	MC	1	64302	0.37	0.32	0.00		0.25	0.27	0.37	0.12		-0.17	-0.12	0.32	-0.08
11	MC	1	64302	0.59	0.37	0.00		0.15	0.58	0.16	0.10		-0.17	0.37	-0.14	-0.22
12	TE	1	64326	0.69	0.36	0.00	0.31	0.69				-0.35	0.36			
13	MC	1	64296	0.47	0.58	0.00		0.22	0.11	0.20	0.47		-0.27	-0.28	-0.22	0.58
14	MC	1	64280	0.41	0.28	0.00		0.14	0.33	0.11	0.41		-0.01	-0.24	-0.06	0.28
15	MC	1	64079	0.40	0.34	0.00		0.40	0.22	0.21	0.17		0.34	-0.14	-0.16	-0.12
16	SA	1	64093	0.36	0.52	0.00	0.63	0.36				-0.51	0.52			
17	MC	1	64299	0.88	0.32	0.00		0.01	0.08	0.88	0.03		-0.12	-0.25	0.32	-0.13
18	MC	1	64267	0.42	0.46	0.00		0.16	0.26	0.16	0.42		-0.28	-0.09	-0.24	0.46
19	MC	1	64284	0.58	0.33	0.00		0.58	0.17	0.17	0.08		0.33	-0.16	-0.23	-0.05
20	TE	1	64072	0.26	0.57	0.00	0.73	0.26				-0.56	0.57			
21	MC	1	64299	0.59	0.42	0.00		0.59	0.16	0.17	0.08		0.42	-0.19	-0.21	-0.22
22	MC	1	64283	0.83	0.31	0.00		0.05	0.83	0.05	0.08		-0.15	0.31	-0.19	-0.15
23	MC	1	64281	0.43	0.46	0.00		0.19	0.43	0.12	0.26		-0.14	0.46	-0.27	-0.19

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	Proportion Omit	Proportion 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	64148	0.82	0.40	0.00		0.82	0.07	0.06	0.05		0.41	-0.24	-0.22	-0.19
25	MC	1	64285	0.86	0.33	0.00		0.05	0.05	0.86	0.04		-0.21	-0.17	0.34	-0.16
26	MC	1	64301	0.51	0.41	0.00		0.11	0.20	0.18	0.51		-0.22	-0.13	-0.22	0.41
27	MC	1	64266	0.49	0.49	0.00		0.27	0.49	0.14	0.10		-0.37	0.49	-0.16	-0.08
28	MC	1	64270	0.33	0.57	0.00		0.33	0.07	0.46	0.13		0.57	-0.08	-0.29	-0.31
29	TE	1	64267	0.25	0.44	0.00	0.75	0.25				-0.44	0.44			
30	SA	1	64264	0.23	0.43	0.00	0.77	0.23				-0.43	0.43			
31	MC	1	64149	0.32	0.41	0.00		0.31	0.13	0.13	0.42		0.41	-0.13	-0.08	-0.24
32	MC	1	64262	0.40	0.24	0.00		0.11	0.21	0.28	0.40		-0.15	-0.11	-0.06	0.24
33	MC	1	64275	0.53	0.59	0.00		0.53	0.32	0.07	0.08		0.59	-0.50	-0.11	-0.12
34	MC	1	64274	0.43	0.43	0.00		0.06	0.45	0.06	0.43		-0.24	-0.21	-0.21	0.43
35	MC	1	64263	0.63	0.49	0.00		0.63	0.13	0.12	0.12		0.49	-0.22	-0.25	-0.25
36	TE	1	64071	0.19	0.49	0.00	0.81	0.19				-0.48	0.49			
37	MC	1	64245	0.52	0.31	0.00		0.19	0.52	0.14	0.14		-0.21	0.31	-0.22	0.02
38	MC	1	64133	0.58	0.41	0.00		0.10	0.17	0.58	0.15		-0.19	-0.19	0.41	-0.20
39	MC	1	64126	0.38	0.51	0.00		0.22	0.29	0.11	0.38		-0.25	-0.25	-0.08	0.51
40	MC	1	64259	0.49	0.49	0.00		0.41	0.49	0.07	0.04		-0.36	0.49	-0.14	-0.14
41	MC	1	64236	0.32	0.53	0.00		0.10	0.28	0.30	0.32		-0.21	-0.29	-0.11	0.53
42	MC	1	64240	0.49	0.38	0.00		0.14	0.19	0.49	0.18		-0.19	-0.19	0.38	-0.12
43	SA	1	64219	0.50	0.44	0.00	0.50	0.49				-0.44	0.45			
44	MC	1	64251	0.28	0.40	0.00		0.28	0.22	0.23	0.27		0.41	-0.12	-0.13	-0.17
45	MC	1	64260	0.56	0.40	0.00		0.22	0.55	0.15	0.08		-0.25	0.40	-0.18	-0.10
46	MC	1	64257	0.55	0.39	0.00		0.55	0.18	0.21	0.07		0.39	-0.21	-0.18	-0.16

Table G-9. Item Statistics, Mathematics Grade 5

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Proportion Omit	Proportion 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	SA	1	64756	0.71	0.46	0.00	0.29	0.71				-0.45	0.46			
2	MC	1	64912	0.47	0.52	0.00		0.47	0.47	0.04	0.02		0.52	-0.40	-0.23	-0.09
3	MC	1	64859	0.59	0.19	0.00		0.59	0.23	0.15	0.03		0.19	0.02	-0.23	-0.10
4	MC	1	64889	0.35	0.40	0.00		0.11	0.34	0.21	0.35		-0.23	-0.17	-0.10	0.40
5	MC	1	64858	0.42	0.35	0.00		0.15	0.25	0.42	0.17		-0.19	-0.19	0.35	-0.05
6	TE	1	64786	0.34	0.61	0.00	0.66	0.34				-0.60	0.60			
7	TE	1	64547	0.33	0.55	0.01	0.67	0.33				-0.53	0.55			
8	MC	1	64720	0.43	0.35	0.00		0.43	0.22	0.12	0.23		0.35	-0.14	-0.22	-0.10
9	MC	1	64885	0.67	0.43	0.00		0.06	0.67	0.11	0.16		-0.18	0.43	-0.22	-0.24
10	SA	1	64842	0.61	0.23	0.00	0.39	0.61				-0.23	0.24			
11	MC	1	64853	0.76	0.47	0.00		0.10	0.76	0.07	0.06		-0.28	0.48	-0.25	-0.21
12	SA	1	64777	0.34	0.51	0.00	0.66	0.34				-0.50	0.51			
13	MC	1	64856	0.55	0.56	0.00		0.26	0.11	0.08	0.55		-0.37	-0.23	-0.17	0.56
14	MC	1	64854	0.56	0.45	0.00		0.25	0.56	0.15	0.03		-0.32	0.45	-0.15	-0.18
15	TE	1	64341	0.23	0.56	0.01	0.77	0.22				-0.54	0.56			
16	SA	1	64725	0.32	0.35	0.00	0.68	0.32				-0.35	0.35			
17	MC	1	64846	0.47	0.34	0.00		0.47	0.24	0.13	0.15		0.34	-0.05	-0.23	-0.19
18	MC	1	64795	0.46	0.50	0.00		0.46	0.12	0.20	0.22		0.50	-0.18	-0.30	-0.17
19	MC	1	64820	0.44	0.46	0.00		0.32	0.15	0.09	0.44		-0.23	-0.18	-0.18	0.46
20	TE	1	64809	0.59	0.46	0.00	0.41	0.59				-0.46	0.46			
21	MC	1	64808	0.39	0.26	0.00		0.11	0.39	0.31	0.19		-0.01	0.26	-0.15	-0.13
22	SA	1	64768	0.32	0.60	0.00	0.68	0.32				-0.60	0.60			
23	MC	1	64796	0.69	0.44	0.00		0.68	0.17	0.11	0.04		0.45	-0.24	-0.27	-0.18

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	Proportion Omit	Proportion 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	64737	0.56	0.33	0.00		0.10	0.19	0.56	0.14		-0.15	-0.09	0.33	-0.23
25	SA	1	64826	0.48	0.47	0.00	0.52	0.48				-0.47	0.47			
26	MC	1	64853	0.40	0.37	0.00		0.24	0.23	0.40	0.13		-0.08	-0.11	0.37	-0.30
27	MC	1	64829	0.57	0.44	0.00		0.06	0.19	0.57	0.18		-0.20	-0.22	0.44	-0.21
28	MC	1	64788	0.37	0.17	0.00		0.21	0.37	0.24	0.18		-0.08	0.17	-0.06	-0.07
29	TE	1	64707	0.44	0.50	0.00	0.56	0.44				-0.50	0.51			
30	SA	1	64754	0.56	0.53	0.00	0.43	0.56				-0.52	0.53			
31	SA	1	64635	0.11	0.42	0.00	0.89	0.10				-0.41	0.42			
32	MC	1	64821	0.54	0.49	0.00		0.09	0.17	0.54	0.20		-0.21	-0.19	0.49	-0.28
33	SA	1	64777	0.25	0.45	0.00	0.74	0.25				-0.45	0.45			
34	SA	1	64760	0.36	0.52	0.00	0.63	0.36				-0.51	0.52			
35	TE	1	64109	0.23	0.49	0.01	0.76	0.23				-0.45	0.49			
36	MC	1	64804	0.52	0.49	0.00		0.52	0.12	0.26	0.10		0.49	-0.29	-0.18	-0.23
37	TE	1	64831	0.34	0.55	0.00	0.66	0.34				-0.55	0.55			
38	MC	1	64589	0.50	0.43	0.01		0.15	0.20	0.50	0.15		-0.26	-0.11	0.43	-0.22
39	MC	1	64708	0.54	0.29	0.00		0.21	0.08	0.54	0.17		-0.14	-0.18	0.30	-0.10
40	MC	1	64797	0.58	0.45	0.00		0.07	0.16	0.19	0.58		-0.20	-0.19	-0.25	0.45
41	MC	1	64791	0.29	0.41	0.00		0.42	0.18	0.29	0.10		-0.34	0.00	0.41	-0.04
42	MC	1	64810	0.44	0.42	0.00		0.34	0.12	0.10	0.43		-0.12	-0.26	-0.21	0.42
43	MC	1	64787	0.54	0.47	0.00		0.25	0.54	0.09	0.12		-0.20	0.47	-0.22	-0.26
44	SA	1	64747	0.12	0.44	0.00	0.88	0.12				-0.43	0.44			
45	MC	1	64739	0.28	0.37	0.00		0.21	0.30	0.20	0.28		-0.17	-0.05	-0.18	0.37
46	MC	1	64816	0.79	0.26	0.00		0.08	0.09	0.79	0.03		-0.13	-0.14	0.26	-0.16

Table G-10. Item Statistics, Mathematics Grade 6

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Proportion Omit	Proportion 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	63505	0.51	0.46	0.00		0.18	0.22	0.51	0.08		-0.24	-0.23	0.46	-0.15
2	SA	1	63367	0.52	0.49	0.00	0.48	0.52				-0.49	0.49			
3	MC	1	63455	0.63	0.40	0.00		0.17	0.14	0.62	0.06		-0.27	-0.16	0.40	-0.15
4	TE	1	63294	0.59	0.52	0.00	0.41	0.58				-0.51	0.52			
5	MC	1	63480	0.54	0.34	0.00		0.25	0.11	0.54	0.10		-0.09	-0.17	0.34	-0.24
6	SA	1	63266	0.25	0.52	0.00	0.75	0.25				-0.51	0.52			
7	MC	1	63458	0.74	0.45	0.00		0.04	0.10	0.12	0.74		-0.15	-0.28	-0.26	0.45
8	MC	1	63419	0.52	0.33	0.00		0.14	0.52	0.21	0.12		-0.10	0.33	-0.21	-0.13
9	SA	1	63402	0.51	0.55	0.00	0.49	0.51				-0.55	0.55			
10	MC	1	63462	0.80	0.41	0.00		0.02	0.05	0.13	0.80		-0.16	-0.17	-0.31	0.41
11	SA	1	63257	0.19	0.51	0.00	0.81	0.19				-0.50	0.51			
12	MC	1	63401	0.40	0.29	0.00		0.08	0.40	0.42	0.11		-0.07	0.29	-0.13	-0.19
13	SA	1	63410	0.14	0.46	0.00	0.85	0.14				-0.45	0.46			
14	MC	1	63410	0.50	0.35	0.00		0.16	0.50	0.19	0.15		-0.26	0.35	-0.12	-0.09
15	TE	1	63189	0.38	0.59	0.01	0.61	0.38				-0.57	0.59			
16	SA	1	63281	0.20	0.51	0.00	0.79	0.20				-0.49	0.51			
17	MC	1	63401	0.77	0.39	0.00		0.02	0.17	0.04	0.77		-0.14	-0.29	-0.18	0.39
18	MC	1	63362	0.68	0.38	0.00		0.06	0.68	0.13	0.12		-0.15	0.39	-0.19	-0.23
19	SA	1	63125	0.24	0.59	0.01	0.76	0.24				-0.57	0.59			
20	MC	1	63425	0.36	0.51	0.00		0.16	0.36	0.31	0.17		-0.29	0.51	-0.09	-0.26
21	MC	1	63398	0.71	0.28	0.00		0.09	0.71	0.13	0.07		-0.01	0.29	-0.20	-0.23
22	MC	1	63375	0.60	0.53	0.00		0.60	0.21	0.10	0.09		0.53	-0.36	-0.21	-0.17
23	MC	1	63359	0.51	0.43	0.00		0.28	0.51	0.16	0.05		-0.26	0.43	-0.19	-0.12

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	Proportion Omit	Proportion 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	TE	1	62873	0.73	0.34	0.01	0.27	0.72				-0.33	0.35			
25	MC	1	63371	0.42	0.32	0.00		0.28	0.16	0.13	0.42		-0.12	-0.20	-0.09	0.33
26	MC	1	63226	0.52	0.26	0.00		0.14	0.23	0.51	0.12		-0.12	-0.18	0.26	-0.03
27	SA	1	63105	0.31	0.44	0.01	0.69	0.30				-0.43	0.44			
28	MC	1	63247	0.33	0.34	0.00		0.46	0.33	0.12	0.08		-0.19	0.34	-0.23	0.03
29	MC	1	63198	0.58	0.32	0.01		0.11	0.58	0.25	0.05		-0.11	0.33	-0.21	-0.14
30	TE	1	63383	0.15	0.36	0.00	0.85	0.15				-0.35	0.36			
31	SA	1	63172	0.15	0.49	0.01	0.84	0.15				-0.47	0.49			
32	MC	1	63334	0.64	0.46	0.00		0.13	0.16	0.64	0.07		-0.25	-0.25	0.46	-0.17
33	MC	1	63310	0.31	0.13	0.00		0.27	0.25	0.31	0.17		-0.03	-0.01	0.13	-0.11
34	MC	1	63358	0.44	0.46	0.00		0.38	0.07	0.11	0.44		-0.17	-0.25	-0.26	0.47
35	MC	1	63367	0.61	0.48	0.00		0.13	0.09	0.17	0.61		-0.18	-0.24	-0.28	0.49
36	MC	1	63359	0.44	0.46	0.00		0.25	0.22	0.44	0.09		-0.22	-0.15	0.46	-0.24
37	TE	1	63022	0.39	0.51	0.01	0.60	0.39				-0.50	0.52			
38	MC	1	63256	0.40	0.25	0.00		0.17	0.35	0.40	0.07		-0.15	-0.03	0.25	-0.19
39	MC	1	63220	0.40	0.26	0.00		0.12	0.16	0.32	0.40		-0.15	-0.16	-0.03	0.26
40	MC	1	63158	0.35	0.55	0.01		0.26	0.34	0.08	0.30		-0.26	0.55	-0.10	-0.24
41	SA	1	63173	0.22	0.50	0.01	0.78	0.22				-0.48	0.50			
42	MC	1	63342	0.60	0.46	0.00		0.12	0.14	0.14	0.60		-0.22	-0.24	-0.20	0.46
43	MC	1	63332	0.33	0.23	0.00		0.24	0.28	0.33	0.15		-0.22	-0.03	0.23	0.01
44	TE	1	63194	0.45	0.59	0.01	0.55	0.45				-0.58	0.60			
45	MC	1	63283	0.64	0.41	0.00		0.63	0.13	0.15	0.09		0.41	-0.22	-0.16	-0.23
46	MC	1	63339	0.32	0.49	0.00		0.32	0.11	0.11	0.46		0.49	-0.05	-0.11	-0.35

Table G-11. Item Statistics, Mathematics Grade 7

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Proportion Omit	Proportion 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	63063	0.23	0.40	0.00		0.18	0.23	0.41	0.18		-0.05	0.40	-0.31	0.01
2	MC	1	63079	0.47	0.54	0.00		0.30	0.47	0.21	0.02		-0.25	0.54	-0.34	-0.12
3	MC	1	63077	0.52	0.33	0.00		0.52	0.13	0.21	0.14		0.33	-0.11	-0.19	-0.14
4	SA	1	62974	0.24	0.54	0.00	0.76	0.24				-0.53	0.54			
5	MC	1	63051	0.50	0.43	0.00		0.07	0.50	0.14	0.29		-0.12	0.43	-0.27	-0.20
6	MC	1	63058	0.40	0.44	0.00		0.40	0.08	0.29	0.22		0.44	-0.23	-0.20	-0.14
7	MC	1	63075	0.35	0.27	0.00		0.31	0.11	0.23	0.35		0.04	-0.23	-0.18	0.27
8	MC	1	63078	0.49	0.37	0.00		0.27	0.11	0.13	0.49		-0.02	-0.28	-0.26	0.37
9	TE	1	63049	0.20	0.45	0.00	0.80	0.20				-0.45	0.45			
10	MC	1	63056	0.47	0.38	0.00		0.47	0.20	0.27	0.06		0.38	-0.33	-0.03	-0.19
11	MC	1	63072	0.50	0.43	0.00		0.37	0.05	0.08	0.50		-0.27	-0.15	-0.19	0.43
12	MC	1	63018	0.49	0.41	0.00		0.49	0.16	0.19	0.15		0.41	-0.15	-0.10	-0.30
13	MC	1	62950	0.62	0.48	0.00		0.62	0.11	0.15	0.12		0.48	-0.18	-0.26	-0.25
14	SA	1	62562	0.25	0.56	0.01	0.74	0.25				-0.55	0.56			
15	MC	1	62929	0.72	0.26	0.00		0.17	0.06	0.06	0.71		-0.12	-0.14	-0.17	0.26
16	MC	1	62870	0.25	0.49	0.00		0.25	0.14	0.25	0.35		0.49	-0.10	-0.05	-0.31
17	MC	1	62951	0.36	0.48	0.00		0.28	0.24	0.12	0.36		-0.23	-0.15	-0.19	0.48
18	MC	1	62956	0.42	0.22	0.00		0.04	0.19	0.42	0.35		-0.10	-0.03	0.22	-0.16
19	TE	1	62637	0.38	0.61	0.01	0.61	0.38				-0.60	0.61			
20	MC	1	62943	0.37	0.28	0.00		0.28	0.25	0.37	0.10		-0.16	-0.12	0.28	-0.03
21	SA	1	62786	0.37	0.47	0.01	0.63	0.36				-0.46	0.47			
22	MC	1	62899	0.30	0.26	0.00		0.37	0.28	0.29	0.05		-0.08	-0.15	0.26	-0.04
23	MC	1	62829	0.61	0.27	0.00		0.22	0.11	0.61	0.06		-0.05	-0.26	0.28	-0.12

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	Proportion Omit	Proportion 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	TE	1	62614	0.15	0.40	0.01	0.85	0.14				-0.37	0.40			
25	MC	1	62795	0.32	0.27	0.01		0.30	0.21	0.32	0.17		-0.06	-0.09	0.28	-0.15
26	SA	1	62388	0.20	0.54	0.01	0.79	0.20				-0.51	0.54			
27	MC	1	62765	0.29	0.30	0.01		0.41	0.18	0.29	0.11		-0.19	-0.12	0.30	0.04
28	MC	1	62853	0.50	0.47	0.00		0.50	0.14	0.16	0.20		0.47	-0.29	-0.24	-0.12
29	MC	1	62858	0.55	0.47	0.00		0.05	0.55	0.23	0.17		-0.12	0.47	-0.24	-0.27
30	MC	1	62810	0.38	0.44	0.00		0.14	0.13	0.35	0.37		-0.17	-0.18	-0.18	0.44
31	MC	1	62847	0.60	0.55	0.00		0.09	0.17	0.14	0.60		-0.13	-0.37	-0.26	0.55
32	MC	1	62860	0.55	0.18	0.00		0.09	0.13	0.23	0.55		-0.03	-0.08	-0.12	0.18
33	SA	1	62818	0.38	0.55	0.00	0.62	0.37				-0.54	0.55			
34	MC	1	62750	0.51	0.51	0.01		0.51	0.22	0.20	0.08		0.51	-0.29	-0.17	-0.24
35	MC	1	62764	0.65	0.39	0.01		0.05	0.13	0.65	0.15		-0.16	-0.27	0.39	-0.15
36	MC	1	62772	0.57	0.33	0.01		0.11	0.57	0.18	0.14		-0.21	0.33	-0.22	-0.03
37	MC	1	62683	0.40	0.45	0.01		0.16	0.27	0.17	0.39		-0.14	-0.16	-0.25	0.46
38	MC	1	62724	0.55	0.25	0.01		0.03	0.54	0.37	0.05		-0.14	0.25	-0.12	-0.18
39	MC	1	62703	0.31	0.24	0.01		0.17	0.31	0.29	0.23		0.00	0.24	0.02	-0.27
40	TE	1	62668	0.30	0.64	0.01	0.70	0.30				-0.62	0.64			
41	SA	1	62446	0.19	0.47	0.01	0.80	0.19				-0.44	0.47			
42	MC	1	62742	0.37	0.45	0.01		0.37	0.22	0.20	0.20		0.45	-0.18	-0.23	-0.10
43	MC	1	62754	0.51	0.52	0.01		0.12	0.15	0.21	0.51		-0.20	-0.25	-0.24	0.52
44	TE	1	62368	0.19	0.57	0.01	0.80	0.19				-0.54	0.57			
45	MC	1	62752	0.65	0.36	0.01		0.11	0.65	0.16	0.08		-0.17	0.36	-0.22	-0.13
46	MC	1	62774	0.43	0.31	0.01		0.13	0.18	0.26	0.42		-0.11	-0.09	-0.17	0.31

Table G-12. Item Statistics, Mathematics Grade 8

Item Number	Item Type	Maximum Points	Number of Students	Item p-value	Item-Total Test Correlation	Proportion Omit	Proportion 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	63142	0.24	0.20	0.00		0.12	0.19	0.24	0.45		0.00	-0.10	0.20	-0.09
2	SA	1	62947	0.42	0.62	0.00	0.58	0.42				-0.62	0.62			
3	MC	1	63131	0.54	0.39	0.00		0.23	0.09	0.54	0.14		-0.15	-0.16	0.39	-0.25
4	MC	1	63145	0.47	0.23	0.00		0.47	0.16	0.25	0.12		0.23	-0.03	-0.16	-0.11
5	SA	1	62978	0.21	0.53	0.00	0.79	0.21				-0.52	0.53			
6	MC	1	63108	0.35	0.44	0.00		0.36	0.14	0.14	0.35		-0.17	-0.16	-0.20	0.44
7	MC	1	63153	0.41	0.31	0.00		0.40	0.27	0.13	0.19		0.31	-0.16	-0.14	-0.08
8	MC	1	63170	0.50	0.49	0.00		0.11	0.28	0.50	0.12		-0.27	-0.16	0.49	-0.28
9	MC	1	63130	0.55	0.34	0.00		0.09	0.17	0.19	0.54		-0.03	-0.26	-0.15	0.34
10	MC	1	63096	0.46	0.33	0.00		0.46	0.26	0.18	0.10		0.34	-0.05	-0.25	-0.16
11	TE	1	62931	0.25	0.49	0.00	0.75	0.25				-0.48	0.49			
12	MC	1	63129	0.41	0.36	0.00		0.41	0.34	0.16	0.08		0.36	-0.12	-0.19	-0.18
13	SA	1	62887	0.20	0.51	0.01	0.79	0.20				-0.49	0.51			
14	MC	1	63055	0.35	0.41	0.00		0.15	0.35	0.05	0.44		-0.20	0.41	-0.14	-0.18
15	MC	1	63000	0.64	0.27	0.00		0.15	0.13	0.64	0.08		-0.09	-0.17	0.27	-0.13
16	TE	1	62913	0.49	0.51	0.00	0.51	0.49				-0.50	0.51			
17	MC	1	62967	0.61	0.45	0.00		0.10	0.61	0.15	0.14		-0.20	0.45	-0.23	-0.22
18	MC	1	63029	0.58	0.43	0.00		0.11	0.16	0.58	0.15		-0.17	-0.27	0.43	-0.16
19	MC	1	63020	0.45	0.40	0.00		0.09	0.45	0.38	0.07		-0.12	0.40	-0.27	-0.12
20	MC	1	63038	0.56	0.50	0.00		0.03	0.11	0.31	0.55		-0.14	-0.24	-0.32	0.50
21	MC	1	62984	0.50	0.43	0.00		0.49	0.18	0.20	0.12		0.43	-0.19	-0.19	-0.20
22	MC	1	62963	0.44	0.44	0.00		0.21	0.23	0.13	0.44		-0.15	-0.18	-0.25	0.44
23	MC	1	63013	0.35	0.27	0.00		0.09	0.35	0.16	0.40		-0.10	0.28	-0.25	-0.02

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	Proportion Omit	Proportion 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	TE	1	62620	0.20	0.46	0.01	0.79	0.20				-0.43	0.46			
25	MC	1	62871	0.55	0.39	0.01		0.54	0.23	0.15	0.07		0.39	-0.15	-0.25	-0.15
26	MC	1	62858	0.67	0.41	0.01		0.09	0.67	0.10	0.13		-0.12	0.41	-0.26	-0.22
27	SA	1	62401	0.30	0.54	0.01	0.69	0.30				-0.52	0.54			
28	MC	1	62925	0.55	0.57	0.00		0.06	0.15	0.23	0.55		-0.22	-0.25	-0.33	0.57
29	TE	1	62779	0.11	0.42	0.01	0.88	0.11				-0.39	0.42			
30	MC	1	62870	0.30	0.39	0.01		0.30	0.35	0.27	0.08		0.39	-0.10	-0.23	-0.07
31	MC	1	62922	0.49	0.27	0.00		0.20	0.49	0.22	0.08		-0.06	0.27	-0.10	-0.22
32	TE	1	62662	0.34	0.51	0.01	0.65	0.34				-0.49	0.51			
33	MC	1	62920	0.57	0.40	0.00		0.14	0.57	0.17	0.12		-0.12	0.40	-0.24	-0.19
34	MC	1	62793	0.36	0.18	0.01		0.23	0.10	0.36	0.30		-0.08	-0.23	0.19	0.04
35	MC	1	62857	0.51	0.31	0.01		0.33	0.51	0.10	0.06		-0.09	0.32	-0.24	-0.17
36	MC	1	62823	0.56	0.47	0.01		0.14	0.55	0.21	0.09		-0.13	0.47	-0.33	-0.17
37	TE	1	62642	0.36	0.37	0.01	0.64	0.36				-0.35	0.37			
38	MC	1	62808	0.70	0.48	0.01		0.08	0.70	0.13	0.09		-0.24	0.48	-0.24	-0.24
39	MC	1	62768	0.42	0.39	0.01		0.42	0.18	0.24	0.16		0.39	-0.26	-0.14	-0.08
40	MC	1	62884	0.75	0.43	0.00		0.12	0.06	0.75	0.07		-0.24	-0.21	0.44	-0.24
41	MC	1	62819	0.36	0.30	0.01		0.36	0.16	0.24	0.24		0.30	-0.13	-0.20	-0.01
42	MC	1	62875	0.51	0.54	0.01		0.14	0.10	0.24	0.51		-0.16	-0.23	-0.33	0.54
43	SA	1	62456	0.15	0.37	0.01	0.84	0.14				-0.33	0.37			
44	MC	1	62861	0.56	0.33	0.01		0.08	0.21	0.56	0.16		-0.15	-0.13	0.33	-0.19
45	MC	1	62881	0.49	0.37	0.01		0.09	0.49	0.32	0.11		-0.12	0.37	-0.23	-0.13
46	MC	1	62872	0.56	0.47	0.01		0.55	0.21	0.16	0.07		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	Proportion Omit	Proportion 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	64327	0.76	0.38	0.00		0.76	0.11	0.05	0.08		0.39	-0.24	-0.21	-0.15
2	MC	1	64340	0.87	0.44	0.00		0.07	0.04	0.02	0.87		-0.25	-0.26	-0.22	0.44
3	MC	1	64351	0.92	0.18	0.00		0.04	0.92	0.02	0.02		-0.13	0.18	-0.06	-0.11
4	MC	1	64332	0.94	0.31	0.00		0.03	0.01	0.01	0.94		-0.19	-0.15	-0.18	0.31
5	MC	1	64326	0.57	0.42	0.00		0.02	0.12	0.28	0.57		-0.20	-0.23	-0.23	0.42
6	MC	1	64314	0.79	0.43	0.00		0.10	0.07	0.79	0.04		-0.26	-0.22	0.43	-0.20
7	MC	1	64332	0.79	0.46	0.00		0.07	0.10	0.79	0.04		-0.26	-0.27	0.46	-0.20
8	MC	1	64222	0.73	0.46	0.00		0.11	0.05	0.11	0.73		-0.22	-0.20	-0.28	0.46
9	MC	1	64041	0.35	0.22	0.01		0.26	0.12	0.34	0.27		0.03	-0.18	0.22	-0.12
10	MC	1	64234	0.66	0.50	0.00		0.12	0.09	0.66	0.13		-0.23	-0.24	0.50	-0.26
11	TE	1	63990	0.41	0.24	0.01	0.58	0.41				-0.23	0.24			
12	MC	1	64307	0.60	0.49	0.00		0.22	0.10	0.08	0.60		-0.21	-0.29	-0.25	0.49
13	MC	1	64325	0.89	0.45	0.00		0.04	0.04	0.04	0.89		-0.25	-0.26	-0.24	0.45
14	MC	1	64304	0.62	0.36	0.00		0.19	0.08	0.62	0.11		-0.12	-0.27	0.37	-0.17
15	MC	1	64344	0.87	0.28	0.00		0.87	0.03	0.02	0.07		0.28	-0.14	-0.12	-0.19
16	MC	1	64342	0.81	0.38	0.00		0.04	0.11	0.04	0.81		-0.21	-0.19	-0.25	0.38
17	MC	1	64230	0.80	0.44	0.00		0.10	0.79	0.03	0.08		-0.28	0.44	-0.15	-0.25
18	MC	1	64197	0.57	0.36	0.00		0.19	0.05	0.57	0.19		-0.12	-0.22	0.37	-0.21
19	MC	1	64321	0.76	0.28	0.00		0.76	0.05	0.08	0.11		0.28	-0.22	-0.18	-0.08
20	MC	1	64317	0.53	0.42	0.00		0.13	0.09	0.25	0.53		-0.22	-0.24	-0.15	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	Proportion Omit	Proportion 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	64319	0.85	0.35	0.00		0.08	0.85	0.03	0.04		-0.13	0.35	-0.22	-0.26
22	MC	1	64318	0.92	0.24	0.00		0.92	0.01	0.04	0.03		0.24	-0.13	-0.15	-0.13
23	MC	1	64286	0.84	0.42	0.00		0.83	0.06	0.06	0.04		0.42	-0.25	-0.23	-0.19
24	MC	1	64298	0.63	0.36	0.00		0.13	0.14	0.10	0.63		-0.33	-0.09	-0.11	0.36
25	MC	1	64295	0.46	0.13	0.00		0.04	0.12	0.46	0.38		-0.21	-0.21	0.13	0.09
26	MC	1	64281	0.58	0.34	0.00		0.10	0.11	0.58	0.21		-0.21	-0.18	0.34	-0.11
27	MC	1	64231	0.64	0.39	0.00		0.22	0.06	0.08	0.64		-0.17	-0.21	-0.24	0.39
28	MC	1	64303	0.55	0.29	0.00		0.03	0.25	0.18	0.54		-0.10	-0.12	-0.20	0.29
29	MC	1	64185	0.75	0.32	0.00		0.11	0.74	0.02	0.13		-0.10	0.32	-0.19	-0.24
30	MC	1	64073	0.67	0.41	0.00		0.66	0.10	0.14	0.09		0.41	-0.16	-0.22	-0.23
31	MC	1	64196	0.72	0.37	0.00		0.71	0.10	0.07	0.11		0.38	-0.19	-0.15	-0.23
32	MC	1	64244	0.74	0.48	0.00		0.13	0.05	0.08	0.74		-0.25	-0.24	-0.27	0.48
33	MC	1	64256	0.71	0.49	0.00		0.11	0.08	0.10	0.71		-0.22	-0.24	-0.28	0.49
34	MC	1	64272	0.79	0.44	0.00		0.79	0.10	0.05	0.06		0.44	-0.23	-0.27	-0.21
35	MC	1	64258	0.82	0.44	0.00		0.04	0.05	0.09	0.82		-0.25	-0.24	-0.24	0.44
36	MC	1	64276	0.40	0.20	0.00		0.22	0.11	0.40	0.26		-0.07	-0.13	0.20	-0.06
37	MC	1	64279	0.67	0.34	0.00		0.11	0.18	0.67	0.03		-0.18	-0.19	0.35	-0.18
38	MC	1	64209	0.46	0.33	0.00		0.11	0.25	0.18	0.46		-0.19	-0.21	-0.03	0.34
39	MC	1	64107	0.56	0.45	0.00		0.15	0.15	0.56	0.13		-0.21	-0.28	0.45	-0.13
40	MC	1	64232	0.77	0.47	0.00		0.11	0.06	0.06	0.77		-0.22	-0.27	-0.27	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	Proportion Omit	Proportion 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	63142	0.85	0.41	0.00		0.04	0.03	0.85	0.08		-0.20	-0.25	0.41	-0.24
2	MC	1	63115	0.87	0.40	0.00		0.03	0.87	0.05	0.05		-0.19	0.40	-0.21	-0.25
3	MC	1	63127	0.91	0.38	0.00		0.02	0.91	0.04	0.03		-0.16	0.38	-0.21	-0.26
4	MC	1	63116	0.79	0.38	0.00		0.04	0.06	0.12	0.79		-0.22	-0.25	-0.17	0.38
5	MC	1	63116	0.86	0.29	0.00		0.03	0.07	0.03	0.86		-0.16	-0.18	-0.13	0.29
6	MC	1	63107	0.85	0.38	0.00		0.02	0.85	0.08	0.05		-0.14	0.38	-0.23	-0.25
7	MC	1	63115	0.78	0.44	0.00		0.78	0.07	0.12	0.03		0.44	-0.29	-0.24	-0.18
8	MC	1	63072	0.84	0.41	0.00		0.04	0.84	0.06	0.06		-0.27	0.41	-0.29	-0.12
9	MC	1	62891	0.62	0.46	0.00		0.62	0.22	0.12	0.04		0.46	-0.24	-0.29	-0.14
10	MC	1	62995	0.46	0.20	0.00		0.19	0.07	0.28	0.45		-0.08	-0.22	-0.02	0.20
11	MC	1	63022	0.66	0.45	0.00		0.66	0.13	0.12	0.09		0.45	-0.23	-0.22	-0.22
12	MC	1	63038	0.45	0.26	0.00		0.14	0.21	0.45	0.19		-0.09	-0.13	0.26	-0.11
13	MC	1	63108	0.77	0.29	0.00		0.12	0.09	0.77	0.02		-0.22	-0.09	0.29	-0.18
14	MC	1	63097	0.90	0.32	0.00		0.02	0.03	0.90	0.05		-0.15	-0.25	0.32	-0.14
15	MC	1	63107	0.75	0.35	0.00		0.07	0.09	0.09	0.75		-0.13	-0.15	-0.27	0.35
16	MC	1	63113	0.75	0.37	0.00		0.08	0.75	0.07	0.09		-0.19	0.37	-0.27	-0.12
17	MC	1	63076	0.61	0.29	0.00		0.61	0.20	0.08	0.10		0.29	-0.20	-0.14	-0.07
18	MC	1	63059	0.69	0.45	0.00		0.07	0.14	0.09	0.69		-0.24	-0.21	-0.23	0.45
19	MC	1	63067	0.87	0.50	0.00		0.86	0.05	0.05	0.03		0.50	-0.29	-0.30	-0.23
20	MC	1	63071	0.68	0.44	0.00		0.04	0.04	0.24	0.67		-0.22	-0.29	-0.24	0.44

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	Proportion Omit	Proportion 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	63089	0.77	0.34	0.00		0.16	0.77	0.03	0.04		-0.16	0.34	-0.23	-0.22
22	MC	1	63035	0.72	0.36	0.00		0.71	0.04	0.05	0.19		0.36	-0.21	-0.16	-0.22
23	MC	1	63011	0.74	0.49	0.00		0.74	0.13	0.08	0.05		0.49	-0.26	-0.31	-0.18
24	MC	1	63023	0.56	0.26	0.00		0.06	0.14	0.56	0.24		-0.06	-0.09	0.26	-0.20
25	MC	1	62997	0.54	0.27	0.00		0.18	0.10	0.54	0.18		-0.05	-0.25	0.27	-0.10
26	MC	1	63003	0.78	0.50	0.00		0.06	0.07	0.08	0.78		-0.25	-0.27	-0.26	0.50
27	MC	1	63004	0.62	0.33	0.00		0.06	0.09	0.23	0.62		-0.11	-0.23	-0.16	0.33
28	MC	1	63022	0.64	0.48	0.00		0.64	0.11	0.10	0.14		0.48	-0.27	-0.24	-0.20
29	MC	1	62986	0.73	0.45	0.00		0.05	0.12	0.10	0.72		-0.27	-0.24	-0.21	0.45
30	MC	1	62842	0.67	0.52	0.00		0.67	0.15	0.14	0.04		0.52	-0.22	-0.33	-0.24
31	MC	1	62843	0.50	0.33	0.00		0.11	0.20	0.50	0.18		-0.17	-0.19	0.33	-0.08
32	MC	1	62950	0.32	0.14	0.00		0.21	0.24	0.32	0.23		-0.09	-0.09	0.14	0.02
33	MC	1	62954	0.36	0.22	0.00		0.13	0.13	0.38	0.36		-0.15	-0.20	0.03	0.22
34	MC	1	62998	0.86	0.46	0.00		0.04	0.86	0.07	0.03		-0.26	0.46	-0.29	-0.20
35	MC	1	62999	0.72	0.42	0.00		0.12	0.72	0.10	0.06		-0.13	0.42	-0.30	-0.23
36	MC	1	62997	0.66	0.51	0.00		0.66	0.11	0.10	0.12		0.51	-0.33	-0.29	-0.14
37	MC	1	63005	0.84	0.51	0.00		0.04	0.06	0.05	0.84		-0.26	-0.29	-0.29	0.51
38	MC	1	62984	0.69	0.38	0.00		0.03	0.22	0.06	0.69		-0.21	-0.17	-0.29	0.38
39	MC	1	62902	0.71	0.46	0.00		0.11	0.14	0.71	0.04		-0.27	-0.22	0.46	-0.22
40	MC	1	62929	0.74	0.45	0.00		0.09	0.07	0.74	0.09		-0.18	-0.25	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	Proportion Omit	Proportion 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	64359	0.74	0.38	0.00		0.15	0.74	0.08	0.03		-0.24	0.38	-0.18	-0.20
2	MC	1	64317	0.84	0.42	0.00		0.08	0.01	0.84	0.07		-0.23	-0.16	0.42	-0.30
3	MC	1	64323	0.85	0.32	0.00		0.03	0.07	0.04	0.85		-0.21	-0.11	-0.24	0.33
4	MC	1	64301	0.59	0.40	0.00		0.04	0.28	0.59	0.09		-0.20	-0.29	0.40	-0.08
5	MC	1	64316	0.58	0.44	0.00		0.58	0.15	0.10	0.17		0.44	-0.19	-0.22	-0.22
6	TE	1	64307	0.69	0.41	0.00	0.31	0.68				-0.41	0.41			
7	MC	1	64320	0.81	0.46	0.00		0.05	0.81	0.10	0.04		-0.27	0.46	-0.24	-0.26
8	MC	1	64300	0.71	0.43	0.00		0.10	0.09	0.10	0.71		-0.20	-0.23	-0.22	0.43
9	MC	1	64321	0.78	0.46	0.00		0.06	0.10	0.78	0.06		-0.27	-0.26	0.46	-0.20
10	TE	1	64320	0.56	0.38	0.00	0.44	0.56				-0.38	0.38			
11	MC	1	64317	0.54	0.23	0.00		0.16	0.54	0.13	0.17		-0.07	0.23	-0.11	-0.13
12	MC	1	64218	0.61	0.32	0.00		0.26	0.08	0.61	0.05		-0.15	-0.15	0.32	-0.22
13	MC	1	64299	0.44	0.41	0.00		0.20	0.23	0.14	0.43		-0.18	-0.12	-0.23	0.42
14	MC	1	64297	0.34	0.27	0.00		0.18	0.38	0.34	0.10		-0.17	-0.06	0.27	-0.11
15	MC	1	64325	0.73	0.40	0.00		0.16	0.05	0.06	0.73		-0.26	-0.20	-0.16	0.40
16	MC	1	64313	0.83	0.36	0.00		0.83	0.04	0.09	0.03		0.36	-0.15	-0.23	-0.21
17	MC	1	64309	0.66	0.47	0.00		0.06	0.08	0.20	0.66		-0.21	-0.23	-0.27	0.47
18	MC	1	64310	0.60	0.35	0.00		0.60	0.22	0.10	0.07		0.35	-0.13	-0.24	-0.18
19	MC	1	64329	0.85	0.48	0.00		0.08	0.85	0.03	0.04		-0.29	0.48	-0.24	-0.26
20	MC	1	64322	0.83	0.54	0.00		0.05	0.05	0.06	0.83		-0.29	-0.28	-0.31	0.54

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	Proportion Omit	Proportion 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	64301	0.80	0.37	0.00		0.79	0.03	0.14	0.03		0.37	-0.21	-0.20	-0.26
22	MC	1	64298	0.83	0.47	0.00		0.05	0.83	0.03	0.08		-0.27	0.48	-0.27	-0.25
23	MC	1	64282	0.58	0.36	0.00		0.21	0.15	0.58	0.06		-0.10	-0.22	0.36	-0.24
24	MC	1	64297	0.61	0.29	0.00		0.06	0.28	0.61	0.05		-0.30	-0.08	0.29	-0.14
25	MC	1	64299	0.61	0.32	0.00		0.22	0.09	0.61	0.08		-0.09	-0.21	0.32	-0.20
26	TE	1	64281	0.54	0.40	0.00	0.46	0.54				-0.40	0.40			
27	MC	1	64057	0.58	0.38	0.00		0.11	0.58	0.12	0.19		-0.15	0.38	-0.25	-0.14
28	TE	1	64239	0.45	0.47	0.00	0.55	0.44				-0.47	0.47			
29	MC	1	64258	0.75	0.46	0.00		0.09	0.08	0.75	0.07		-0.24	-0.29	0.47	-0.20
30	MC	1	64273	0.64	0.20	0.00		0.64	0.19	0.07	0.09		0.20	-0.04	-0.16	-0.13
31	MC	1	64265	0.66	0.47	0.00		0.65	0.12	0.13	0.10		0.47	-0.28	-0.19	-0.23
32	MC	1	64278	0.73	0.48	0.00		0.73	0.16	0.07	0.04		0.48	-0.27	-0.30	-0.21
33	MC	1	64161	0.69	0.44	0.00		0.14	0.05	0.69	0.12		-0.27	-0.27	0.44	-0.15
34	MC	1	64085	0.83	0.46	0.00		0.83	0.07	0.06	0.04		0.46	-0.23	-0.27	-0.25
35	MC	1	64233	0.70	0.45	0.00		0.12	0.06	0.70	0.11		-0.25	-0.21	0.45	-0.23
36	MC	1	64270	0.82	0.46	0.00		0.82	0.03	0.08	0.07		0.46	-0.24	-0.25	-0.27
37	MC	1	64272	0.85	0.47	0.00		0.08	0.03	0.04	0.85		-0.26	-0.26	-0.29	0.48
38	MC	1	64292	0.78	0.51	0.00		0.06	0.78	0.12	0.04		-0.28	0.51	-0.32	-0.21

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	Proportion Omit	Proportion 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	63076	0.85	0.48	0.00		0.05	0.03	0.06	0.85		-0.25	-0.23	-0.30	0.48
2	MC	1	63055	0.79	0.40	0.00		0.14	0.05	0.78	0.03		-0.22	-0.23	0.40	-0.23
3	MC	1	63074	0.82	0.44	0.00		0.12	0.04	0.82	0.02		-0.26	-0.27	0.44	-0.21
4	MC	1	63079	0.82	0.42	0.00		0.06	0.08	0.05	0.82		-0.19	-0.25	-0.23	0.42
5	MC	1	63081	0.80	0.40	0.00		0.07	0.06	0.80	0.07		-0.20	-0.24	0.40	-0.20
6	MC	1	63051	0.83	0.51	0.00		0.03	0.83	0.06	0.07		-0.24	0.51	-0.31	-0.28
7	MC	1	63029	0.60	0.44	0.00		0.10	0.60	0.16	0.14		-0.10	0.44	-0.31	-0.21
8	MC	1	62954	0.56	0.33	0.00		0.56	0.17	0.18	0.08		0.33	-0.15	-0.12	-0.22
9	MC	1	63022	0.76	0.28	0.00		0.75	0.09	0.12	0.04		0.28	-0.22	-0.07	-0.19
10	MC	1	63048	0.74	0.48	0.00		0.05	0.74	0.16	0.04		-0.19	0.48	-0.32	-0.25
11	MC	1	63053	0.52	0.40	0.00		0.15	0.08	0.25	0.52		-0.23	-0.20	-0.14	0.40
12	TE	1	63020	0.41	0.44	0.00	0.59	0.40				-0.43	0.44			
13	MC	1	63047	0.64	0.30	0.00		0.20	0.64	0.13	0.03		-0.03	0.30	-0.30	-0.17
14	MC	1	63022	0.67	0.53	0.00		0.67	0.12	0.12	0.10		0.53	-0.29	-0.32	-0.17
15	MC	1	62949	0.81	0.50	0.00		0.81	0.03	0.12	0.04		0.50	-0.22	-0.33	-0.25
16	MC	1	63006	0.71	0.42	0.00		0.12	0.12	0.71	0.06		-0.16	-0.26	0.43	-0.23
17	TE	1	63024	0.51	0.58	0.00	0.49	0.51				-0.57	0.58			
18	MC	1	63029	0.66	0.40	0.00		0.15	0.10	0.09	0.66		-0.18	-0.22	-0.20	0.40
19	MC	1	63039	0.59	0.49	0.00		0.27	0.08	0.06	0.59		-0.22	-0.26	-0.29	0.49
20	MC	1	63055	0.74	0.49	0.00		0.74	0.09	0.10	0.06		0.49	-0.24	-0.29	-0.22

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	Proportion Omit	Proportion 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	62950	0.80	0.44	0.00		0.11	0.80	0.06	0.03		-0.26	0.45	-0.28	-0.17
22	TE	1	62890	0.73	0.39	0.00	0.27	0.72				-0.38	0.39			
23	MC	1	62983	0.71	0.42	0.00		0.04	0.19	0.71	0.06		-0.24	-0.23	0.42	-0.21
24	MC	1	62907	0.59	0.42	0.00		0.59	0.09	0.16	0.16		0.42	-0.29	-0.23	-0.10
25	MC	1	62929	0.59	0.32	0.00		0.18	0.07	0.59	0.16		-0.11	-0.26	0.33	-0.13
26	MC	1	62948	0.77	0.39	0.00		0.09	0.06	0.77	0.07		-0.14	-0.20	0.40	-0.28
27	TE	1	62809	0.57	0.41	0.00	0.43	0.57				-0.40	0.42			
28	MC	1	62934	0.60	0.27	0.00		0.08	0.60	0.08	0.24		-0.19	0.28	-0.25	-0.03
29	MC	1	62949	0.46	0.39	0.00		0.46	0.14	0.05	0.36		0.39	-0.19	-0.20	-0.18
30	MC	1	62958	0.90	0.46	0.00		0.04	0.03	0.02	0.90		-0.24	-0.28	-0.25	0.46
31	MC	1	62965	0.83	0.46	0.00		0.03	0.83	0.04	0.10		-0.21	0.46	-0.30	-0.25
32	MC	1	62931	0.65	0.44	0.00		0.07	0.65	0.17	0.12		-0.22	0.44	-0.23	-0.21
33	MC	1	62900	0.74	0.39	0.00		0.05	0.15	0.74	0.06		-0.20	-0.17	0.40	-0.27
34	MC	1	62934	0.52	0.33	0.00		0.52	0.08	0.20	0.20		0.33	-0.30	-0.15	-0.05
35	MC	1	62942	0.62	0.45	0.00		0.62	0.12	0.15	0.12		0.45	-0.24	-0.26	-0.14
36	MC	1	62934	0.69	0.55	0.00		0.15	0.09	0.07	0.69		-0.22	-0.30	-0.34	0.55
37	MC	1	62947	0.58	0.40	0.00		0.10	0.58	0.17	0.15		-0.10	0.40	-0.29	-0.16
38	MC	1	62928	0.51	0.31	0.00		0.26	0.51	0.13	0.10		-0.02	0.31	-0.28	-0.18
39	MC	1	62946	0.82	0.51	0.00		0.82	0.06	0.06	0.06		0.51	-0.29	-0.29	-0.24
40	MC	1	62953	0.83	0.48	0.00		0.03	0.83	0.08	0.06		-0.23	0.48	-0.27	-0.28

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	Proportion Omit	Proportion 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	62335	0.61	0.25	0.00		0.13	0.17	0.10	0.61		-0.10	-0.20	-0.06	0.25
2	MC	1	62285	0.84	0.38	0.00		0.01	0.08	0.84	0.07		-0.15	-0.24	0.38	-0.22
3	MC	1	62228	0.65	0.40	0.00		0.07	0.15	0.13	0.65		-0.14	-0.17	-0.28	0.40
4	MC	1	62294	0.52	0.37	0.00		0.10	0.27	0.52	0.11		-0.17	-0.16	0.38	-0.21
5	MC	1	62317	0.80	0.45	0.00		0.05	0.06	0.80	0.08		-0.21	-0.29	0.45	-0.22
6	MC	1	62312	0.51	0.30	0.00		0.20	0.51	0.20	0.10		-0.17	0.30	-0.15	-0.08
7	TE	1	62016	0.88	0.33	0.01	0.12	0.88				-0.33	0.35			
8	MC	1	62293	0.80	0.37	0.00		0.01	0.14	0.80	0.04		-0.15	-0.25	0.37	-0.20
9	TE	1	62047	0.40	0.35	0.01	0.59	0.40				-0.33	0.35			
10	MC	1	62196	0.64	0.40	0.00		0.64	0.18	0.12	0.06		0.40	-0.21	-0.23	-0.13
11	MC	1	62169	0.56	0.32	0.00		0.10	0.07	0.27	0.56		-0.20	-0.28	-0.05	0.32
12	MC	1	62261	0.62	0.40	0.00		0.12	0.62	0.22	0.04		-0.19	0.40	-0.19	-0.25
13	MC	1	62247	0.64	0.37	0.00		0.05	0.14	0.64	0.17		-0.27	-0.13	0.37	-0.19
14	MC	1	62213	0.55	0.29	0.00		0.20	0.55	0.10	0.15		-0.05	0.30	-0.25	-0.14
15	MC	1	62220	0.56	0.41	0.00		0.56	0.14	0.20	0.10		0.41	-0.17	-0.22	-0.17
16	MC	1	62191	0.66	0.41	0.00		0.16	0.65	0.09	0.10		-0.14	0.41	-0.29	-0.20
17	MC	1	62143	0.70	0.46	0.00		0.16	0.70	0.06	0.07		-0.28	0.47	-0.27	-0.17
18	MC	1	62120	0.64	0.45	0.00		0.64	0.14	0.15	0.07		0.45	-0.15	-0.31	-0.20
19	MC	1	62190	0.64	0.50	0.00		0.63	0.16	0.13	0.08		0.50	-0.16	-0.34	-0.24
20	MC	1	62168	0.51	0.45	0.00		0.09	0.29	0.11	0.51		-0.22	-0.15	-0.29	0.46

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	Proportion Omit	Proportion 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	62180	0.59	0.53	0.00		0.59	0.16	0.14	0.10		0.53	-0.21	-0.31	-0.23
22	MC	1	62178	0.62	0.46	0.00		0.62	0.16	0.11	0.11		0.46	-0.26	-0.28	-0.11
23	MC	1	62185	0.58	0.39	0.00		0.57	0.11	0.16	0.15		0.39	-0.25	-0.20	-0.09
24	MC	1	62205	0.69	0.33	0.00		0.05	0.08	0.69	0.17		-0.21	-0.25	0.34	-0.09
25	MC	1	62201	0.69	0.41	0.00		0.16	0.69	0.09	0.06		-0.16	0.41	-0.25	-0.23
26	MC	1	61814	0.73	0.43	0.00		0.10	0.10	0.08	0.73		-0.19	-0.26	-0.21	0.43
27	MC	1	61734	0.75	0.48	0.00		0.11	0.08	0.75	0.06		-0.23	-0.32	0.49	-0.20
28	MC	1	61695	0.58	0.55	0.00		0.58	0.12	0.18	0.11		0.55	-0.28	-0.25	-0.25
29	MC	1	61773	0.46	0.47	0.00		0.20	0.16	0.18	0.46		-0.13	-0.26	-0.22	0.47
30	TE	1	61579	0.77	0.47	0.01	0.23	0.76				-0.46	0.47			
31	MC	1	61740	0.68	0.48	0.00		0.15	0.67	0.15	0.03		-0.24	0.48	-0.29	-0.20
32	MC	1	61750	0.61	0.50	0.00		0.10	0.13	0.61	0.16		-0.28	-0.26	0.51	-0.20
33	MC	1	61712	0.64	0.46	0.00		0.11	0.64	0.12	0.13		-0.20	0.47	-0.34	-0.14
34	MC	1	61697	0.72	0.50	0.00		0.09	0.12	0.72	0.07		-0.23	-0.27	0.50	-0.27
35	MC	1	61628	0.47	0.40	0.00		0.47	0.24	0.10	0.19		0.40	-0.15	-0.31	-0.10
36	MC	1	61578	0.62	0.48	0.01		0.14	0.12	0.13	0.62		-0.17	-0.25	-0.27	0.48
37	TE	1	61700	0.38	0.43	0.00	0.62	0.38				-0.42	0.43			
38	MC	1	61675	0.57	0.50	0.00		0.19	0.11	0.13	0.56		-0.13	-0.29	-0.29	0.50
39	MC	1	61654	0.71	0.34	0.00		0.07	0.70	0.10	0.12		-0.20	0.35	-0.15	-0.18
40	MC	1	61687	0.80	0.55	0.00		0.06	0.09	0.80	0.05		-0.29	-0.33	0.55	-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	Proportion Omit	Proportion 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	61681	0.66	0.43	0.00		0.03	0.10	0.21	0.66		-0.19	-0.16	-0.30	0.44
42	MC	1	61705	0.68	0.51	0.00		0.15	0.09	0.08	0.68		-0.24	-0.25	-0.29	0.51
43	MC	1	61668	0.58	0.36	0.00		0.12	0.18	0.58	0.11		-0.11	-0.22	0.36	-0.16
44	MC	1	61518	0.57	0.41	0.01		0.17	0.56	0.19	0.07		-0.15	0.41	-0.23	-0.20
45	MC	1	61602	0.50	0.25	0.01		0.06	0.50	0.36	0.08		-0.17	0.26	-0.05	-0.23
46	MC	1	61618	0.43	0.23	0.00		0.32	0.43	0.17	0.08		0.02	0.23	-0.20	-0.16
47	MC	1	61612	0.68	0.49	0.00		0.09	0.67	0.12	0.11		-0.22	0.49	-0.30	-0.21
48	MC	1	61629	0.74	0.45	0.00		0.11	0.09	0.06	0.74		-0.31	-0.18	-0.20	0.46
49	MC	1	61626	0.68	0.44	0.00		0.67	0.10	0.15	0.07		0.45	-0.26	-0.19	-0.22
50	MC	1	61623	0.72	0.36	0.00		0.12	0.09	0.06	0.72		-0.17	-0.13	-0.26	0.36

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 θ .

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

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

with

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

and

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

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

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

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

Estimation of the Prior Distribution of T_j

The k items in each test are scaled together using a generalized IRT model (3PL/2PPC) that fits a three-parameter logistic model (3PL) to the 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 (α_i) is the item discrimination, and gamma (γ_{ih}) is related to the difficulty of the item levels; the trace lines for adjacent score levels intersect at γ_{ih} / α_i .

Item parameters estimated from the national standardization sample are used to obtain SPI values.

$T_{ij}(\theta)$ is the expected score for item i in strand j , and θ 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 θ (Lord, 1980, p.71),

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

where

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

Given these results, Lord (1980, p. 79 and 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

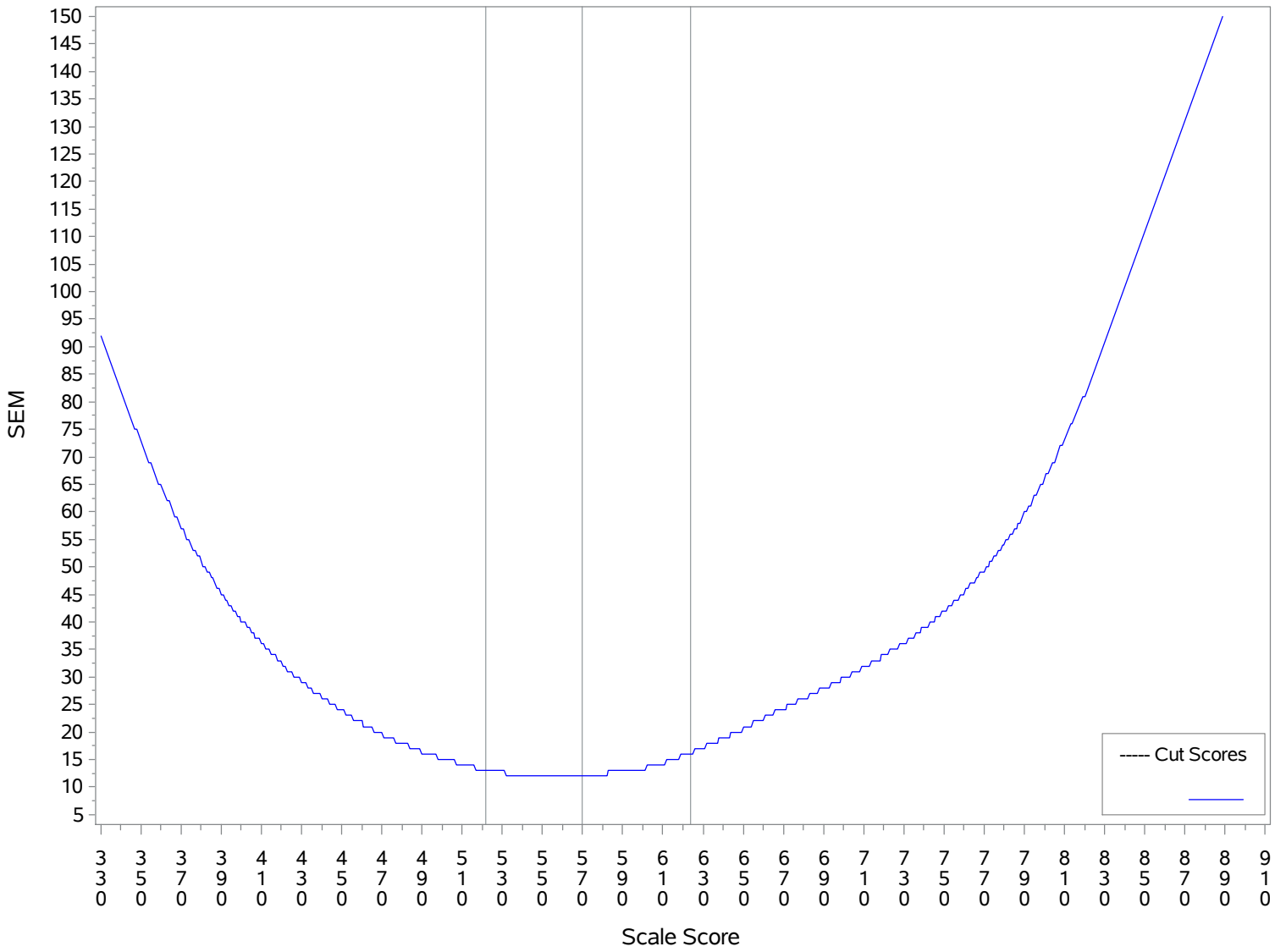


Figure I-2 CSEM with cut scores, ELA Grade 4

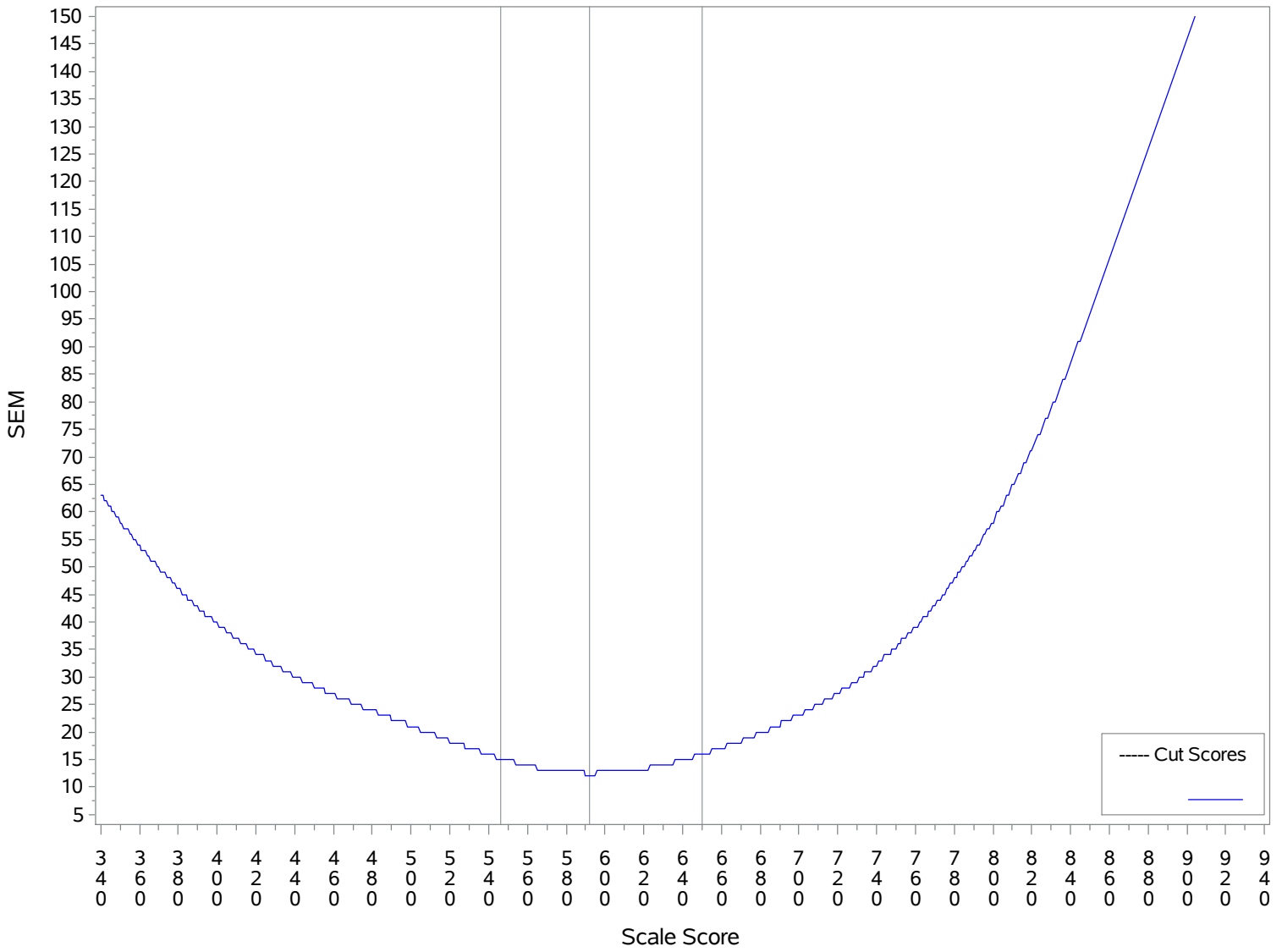


Figure I-3 CSEM with cut scores, ELA Grade 5

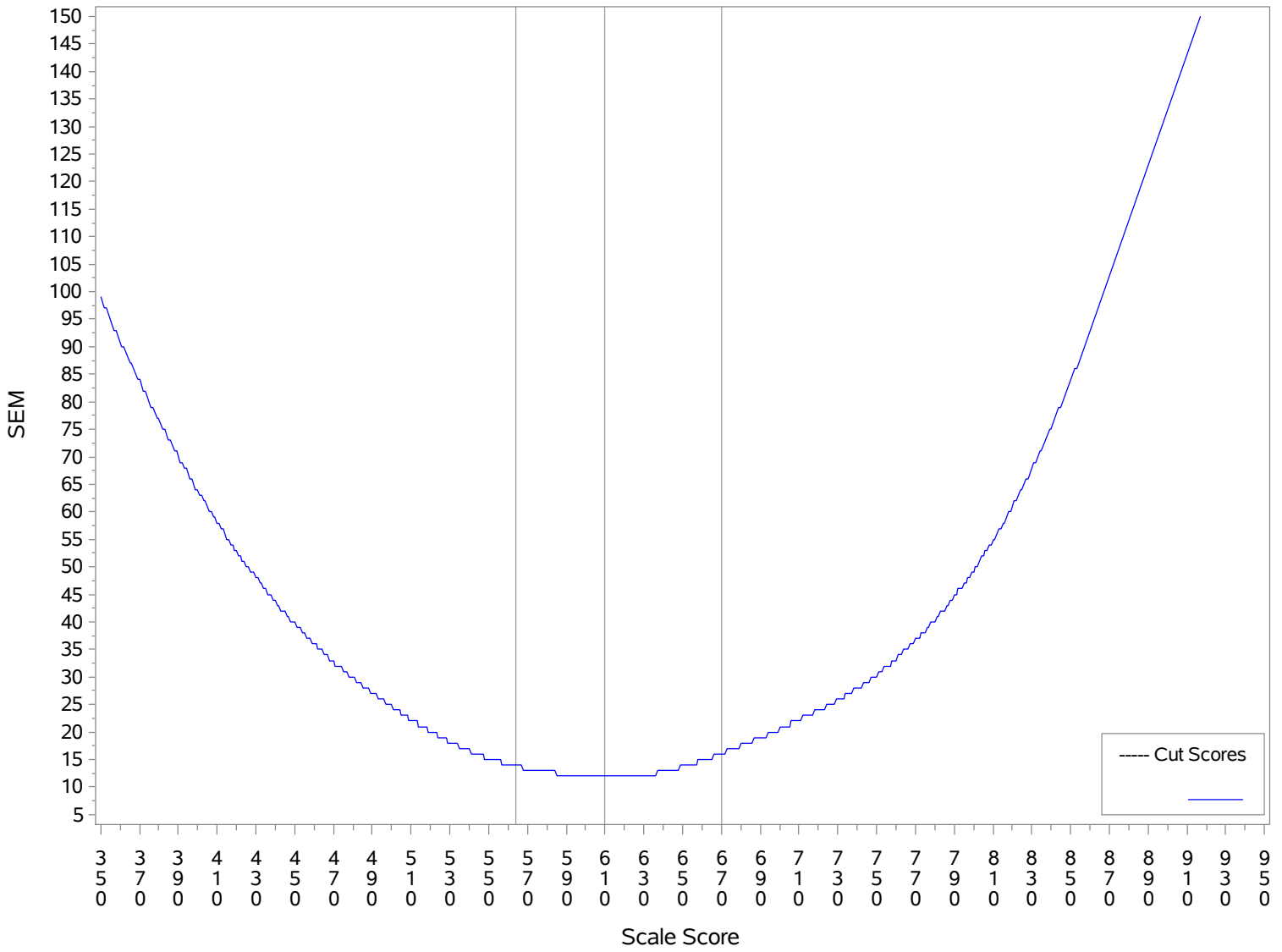


Figure I-4 CSEM with cut scores, ELA Grade 6

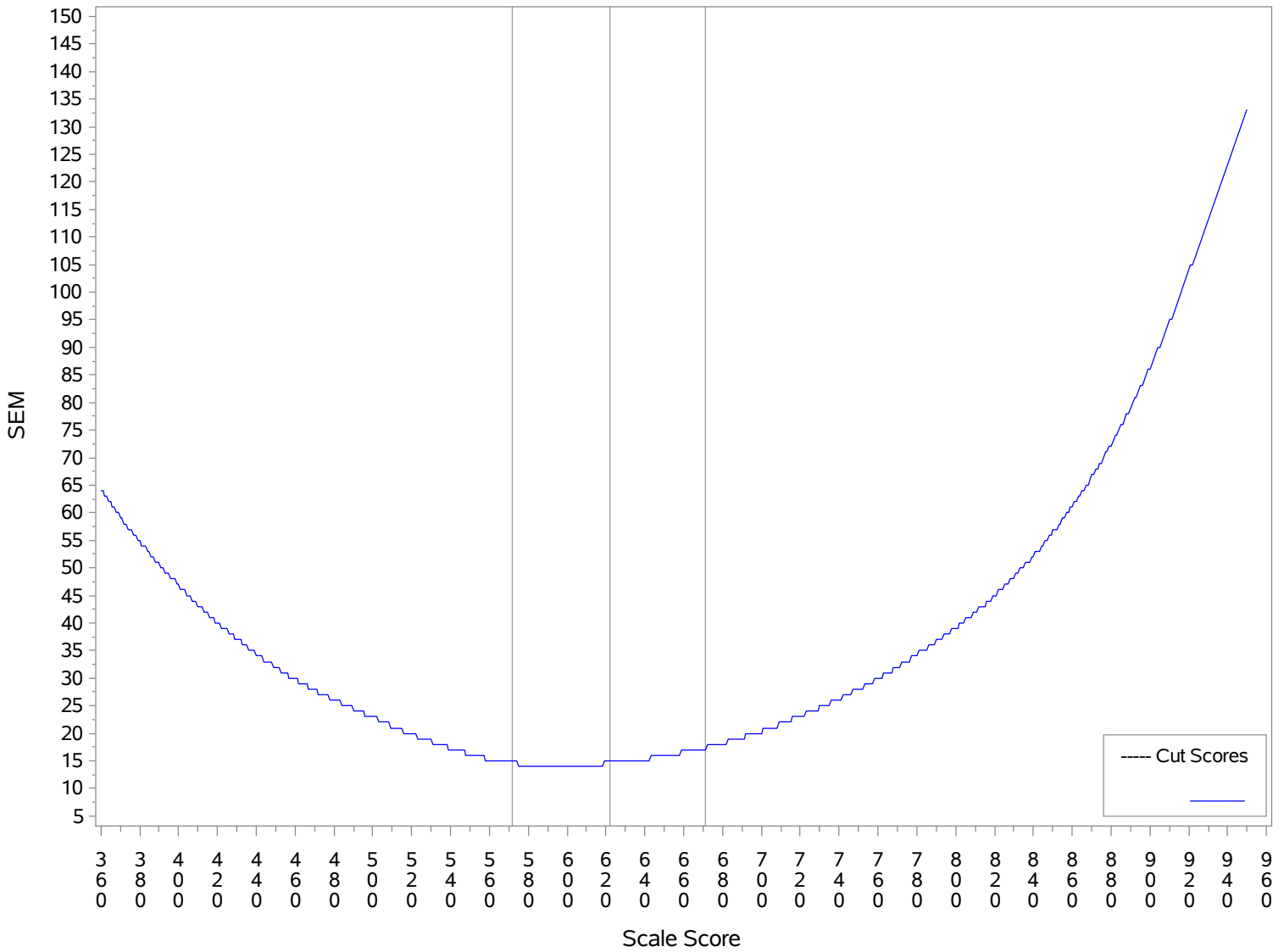


Figure I-5 CSEM with cut scores, ELA Grade 7

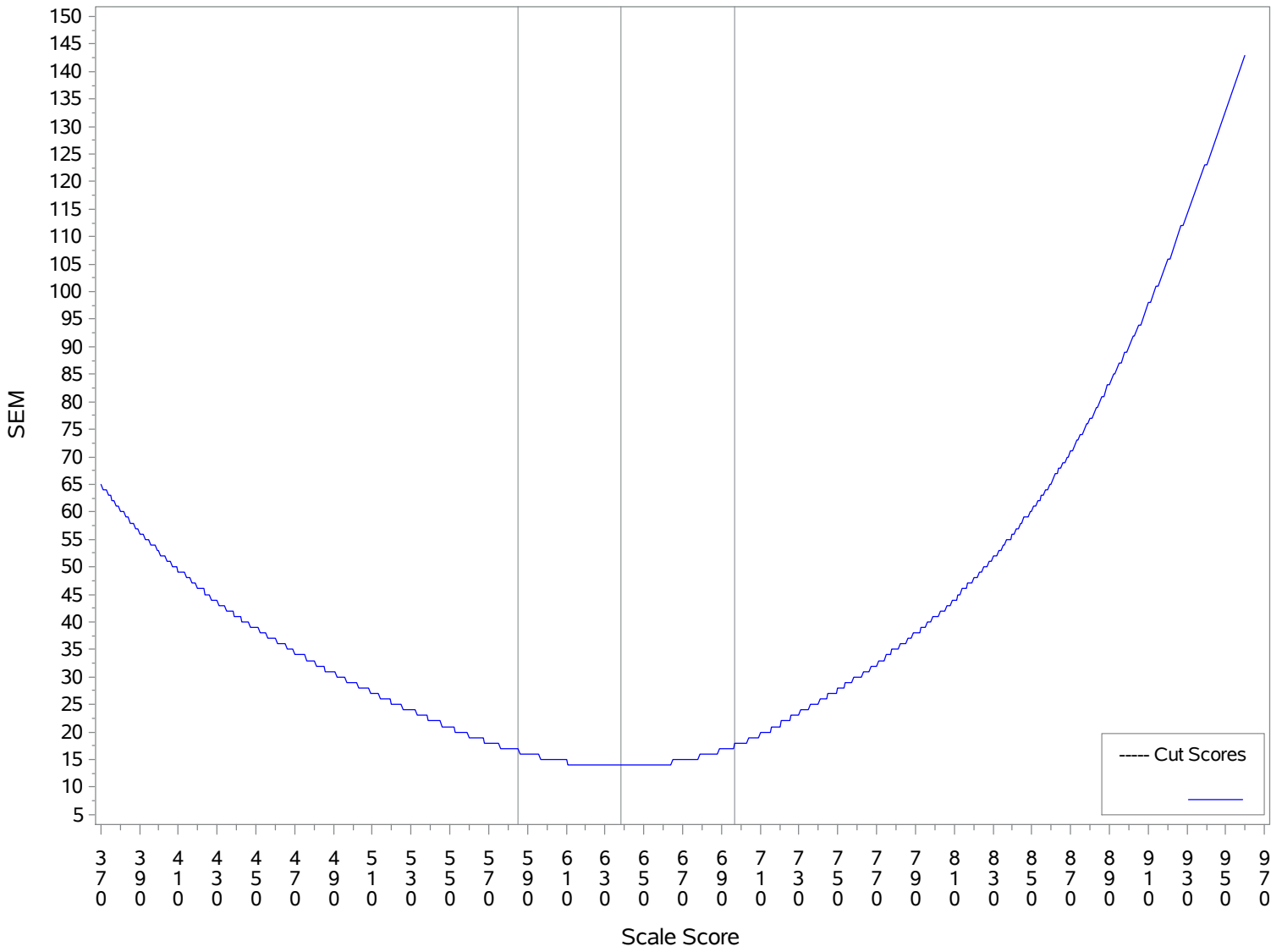


Figure I-6 CSEM with cut scores, ELA Grade 8

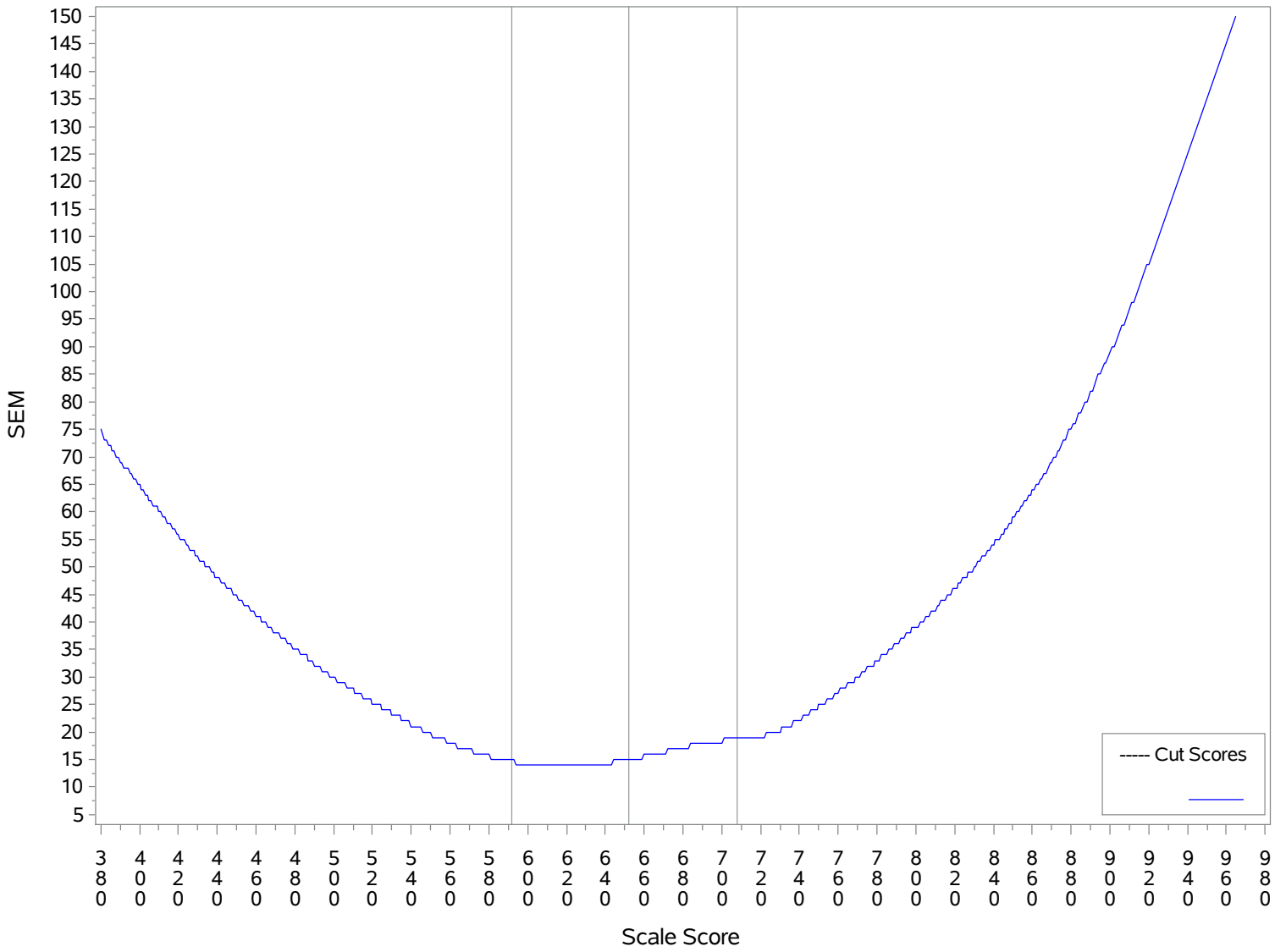


Figure I-7 CSEM with cut scores, Mathematics Grade 3

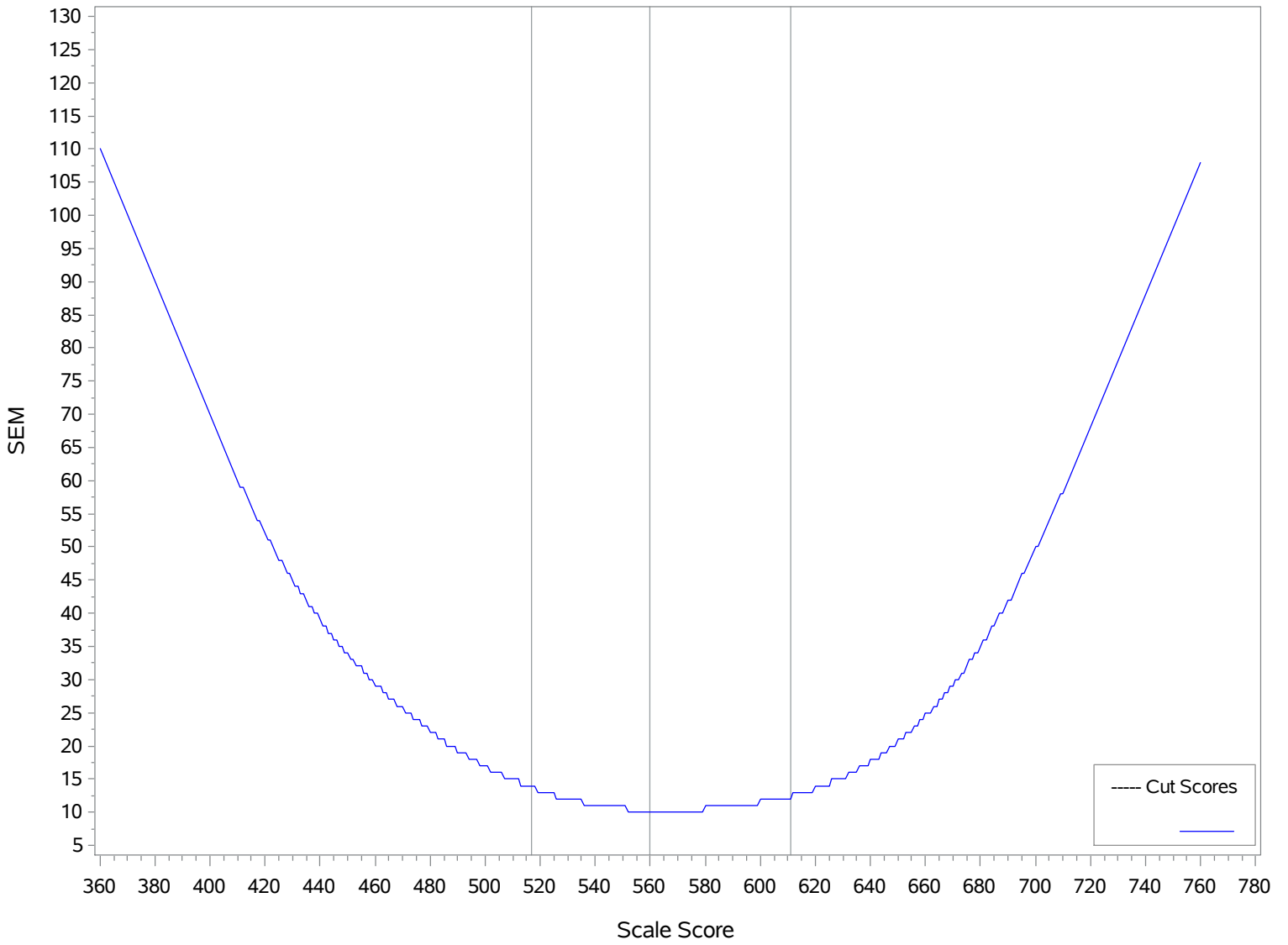


Figure I-8 CSEM with cut scores, Mathematics Grade 4

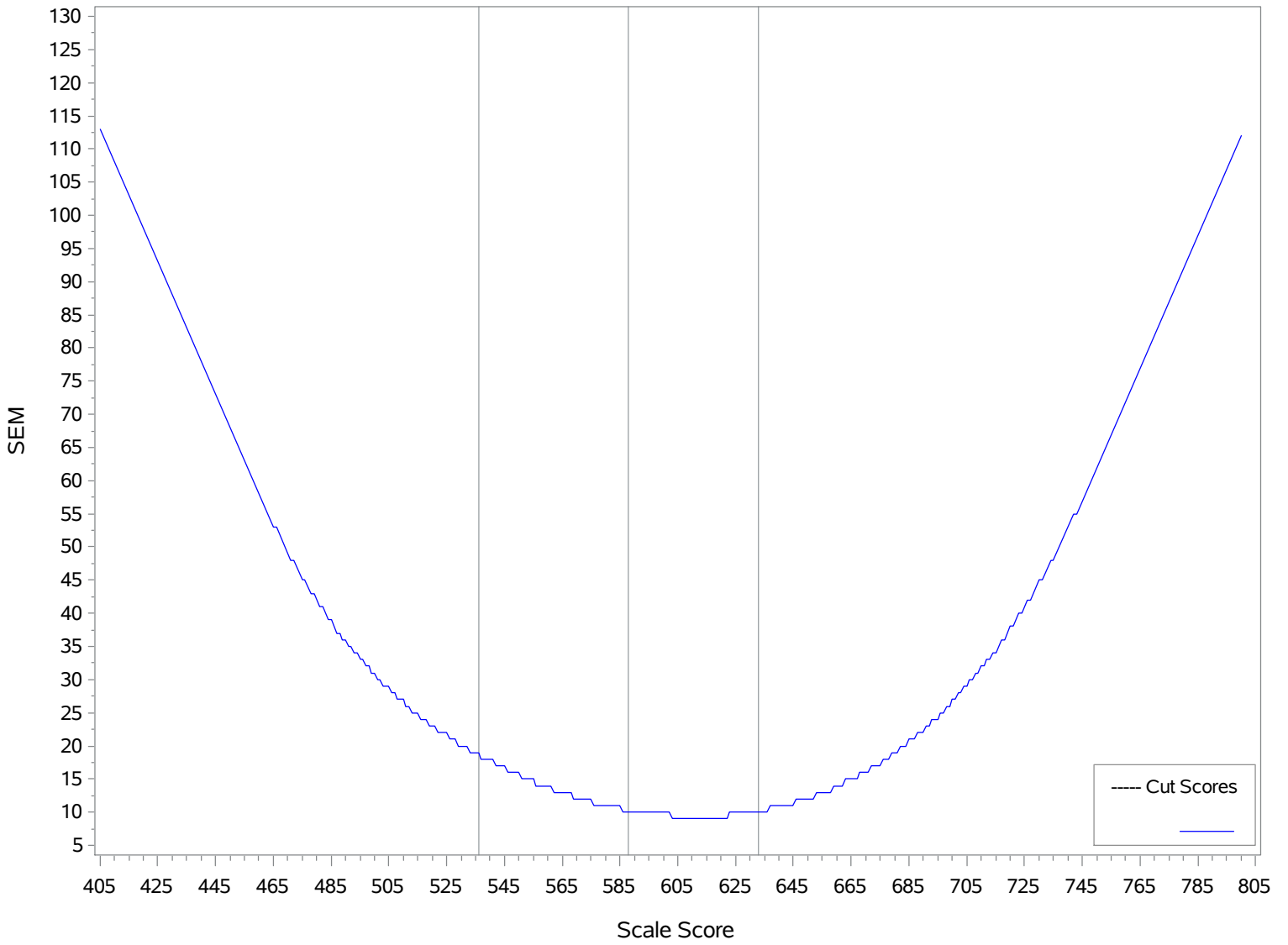


Figure I-9 CSEM with cut scores, Mathematics Grade 5

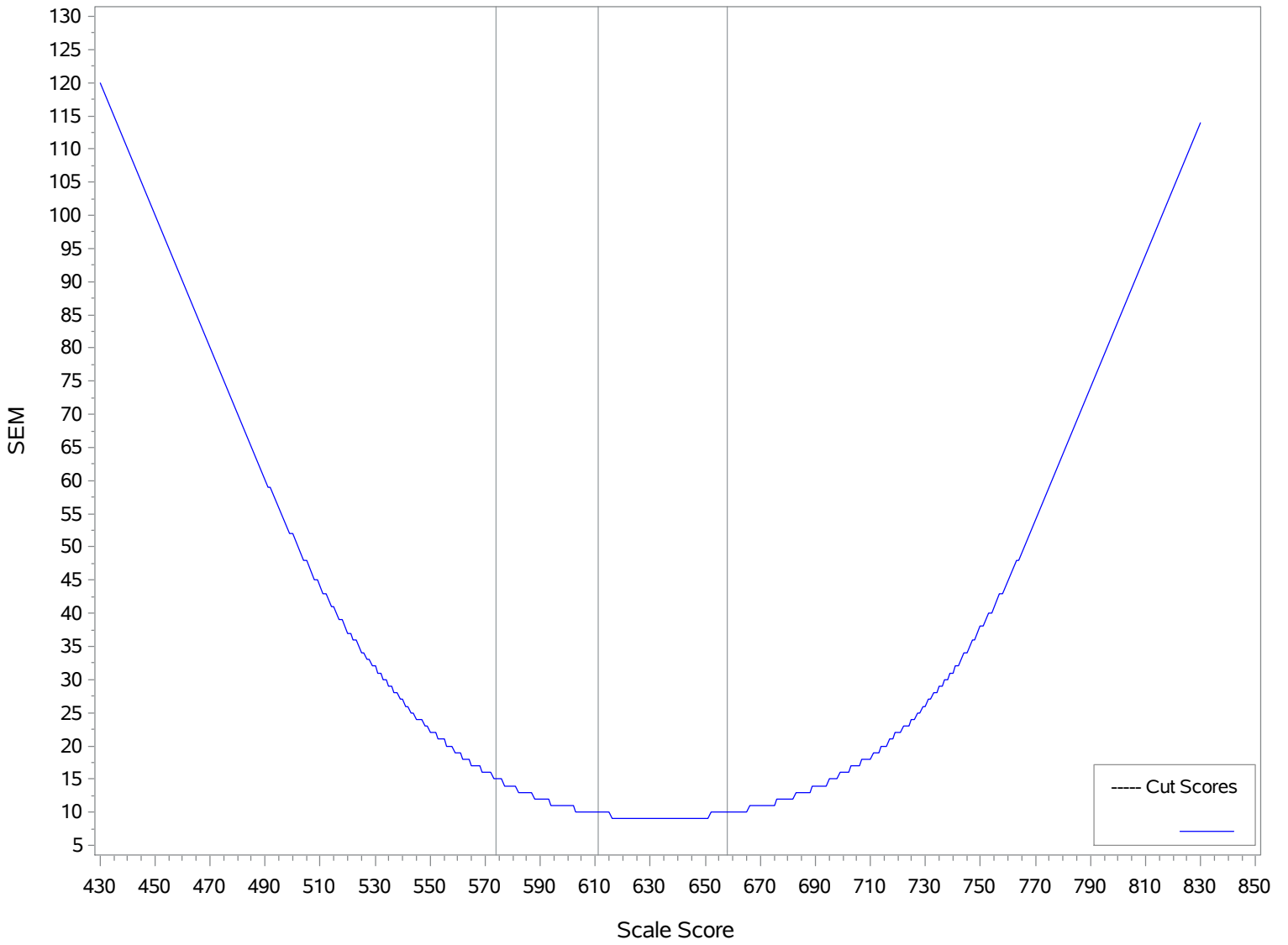


Figure I-10 CSEM with cut scores, Mathematics Grade 6

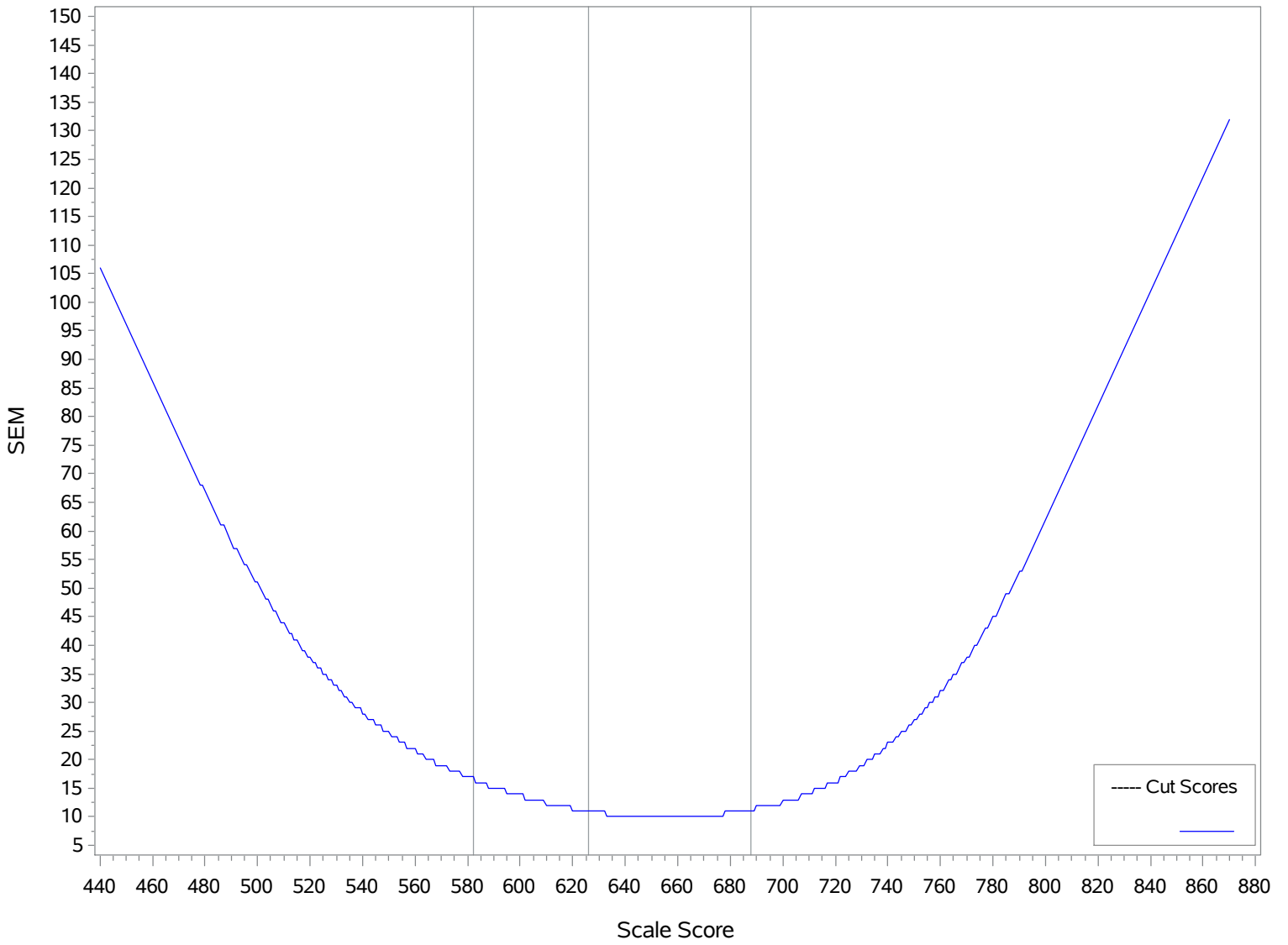


Figure I-11 CSEM with cut scores, Mathematics Grade 7

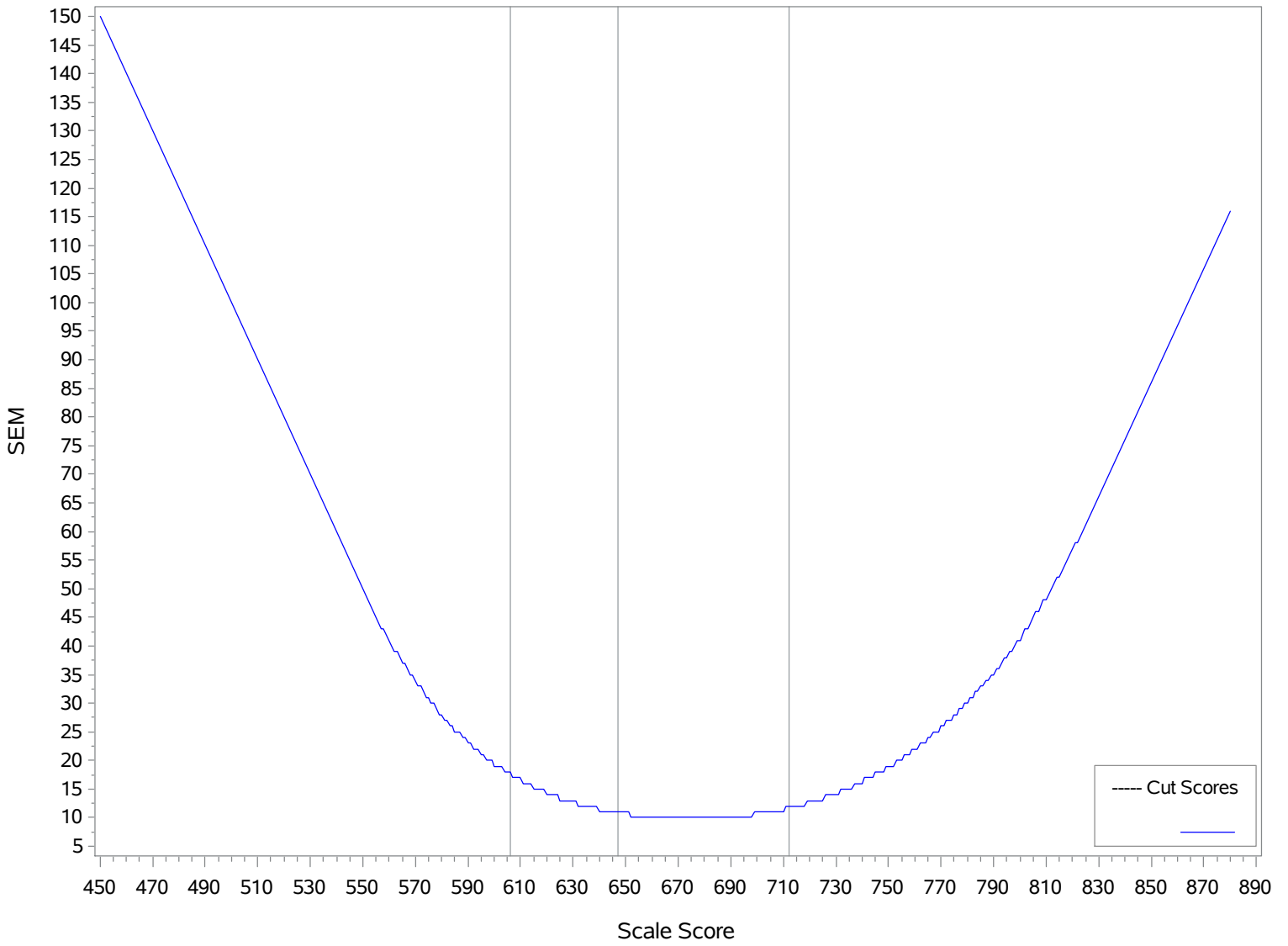


Figure I-12 CSEM with cut scores, Mathematics Grade 8

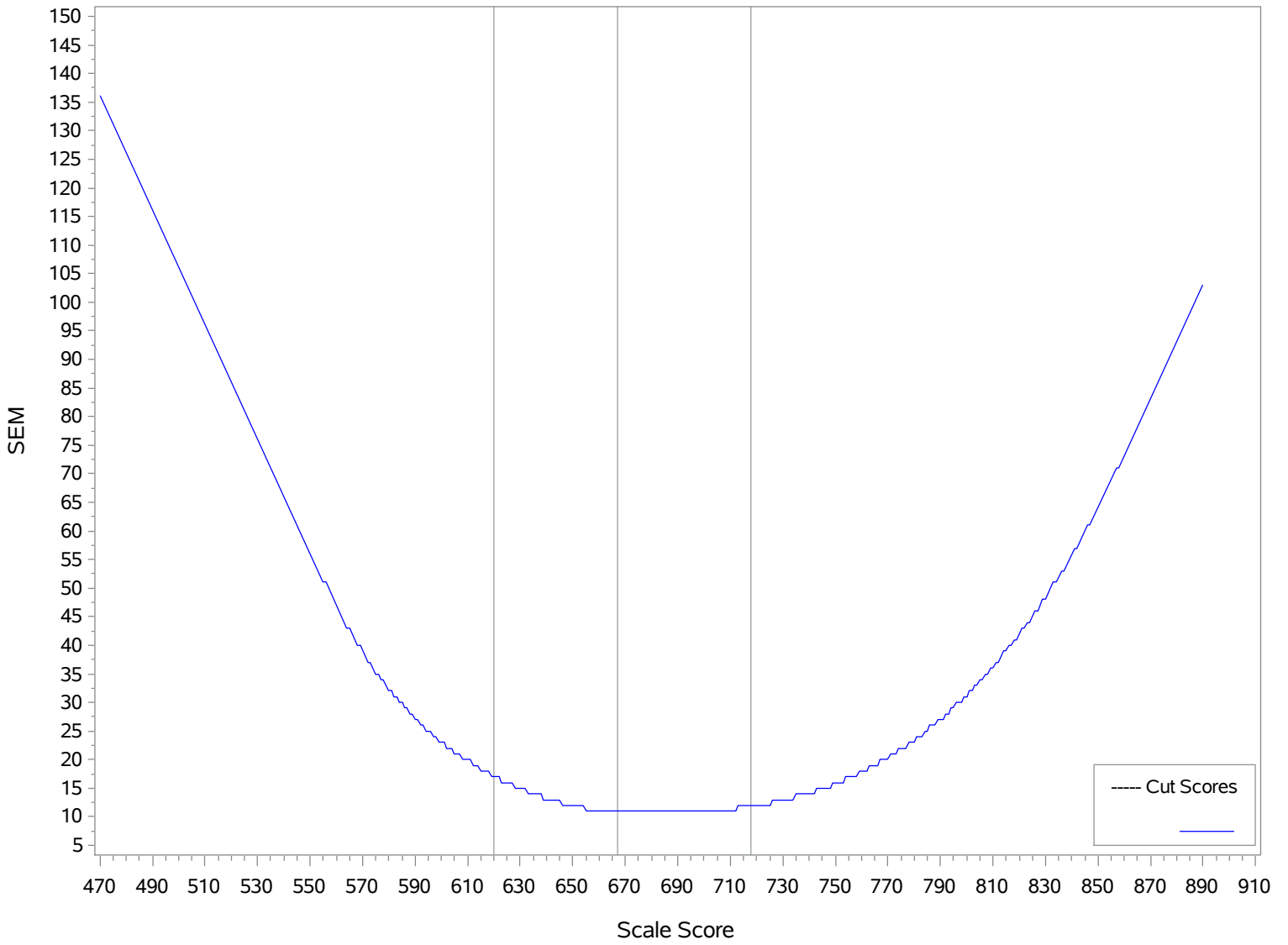


Figure I-13 CSEM with cut scores, Science Grade 4

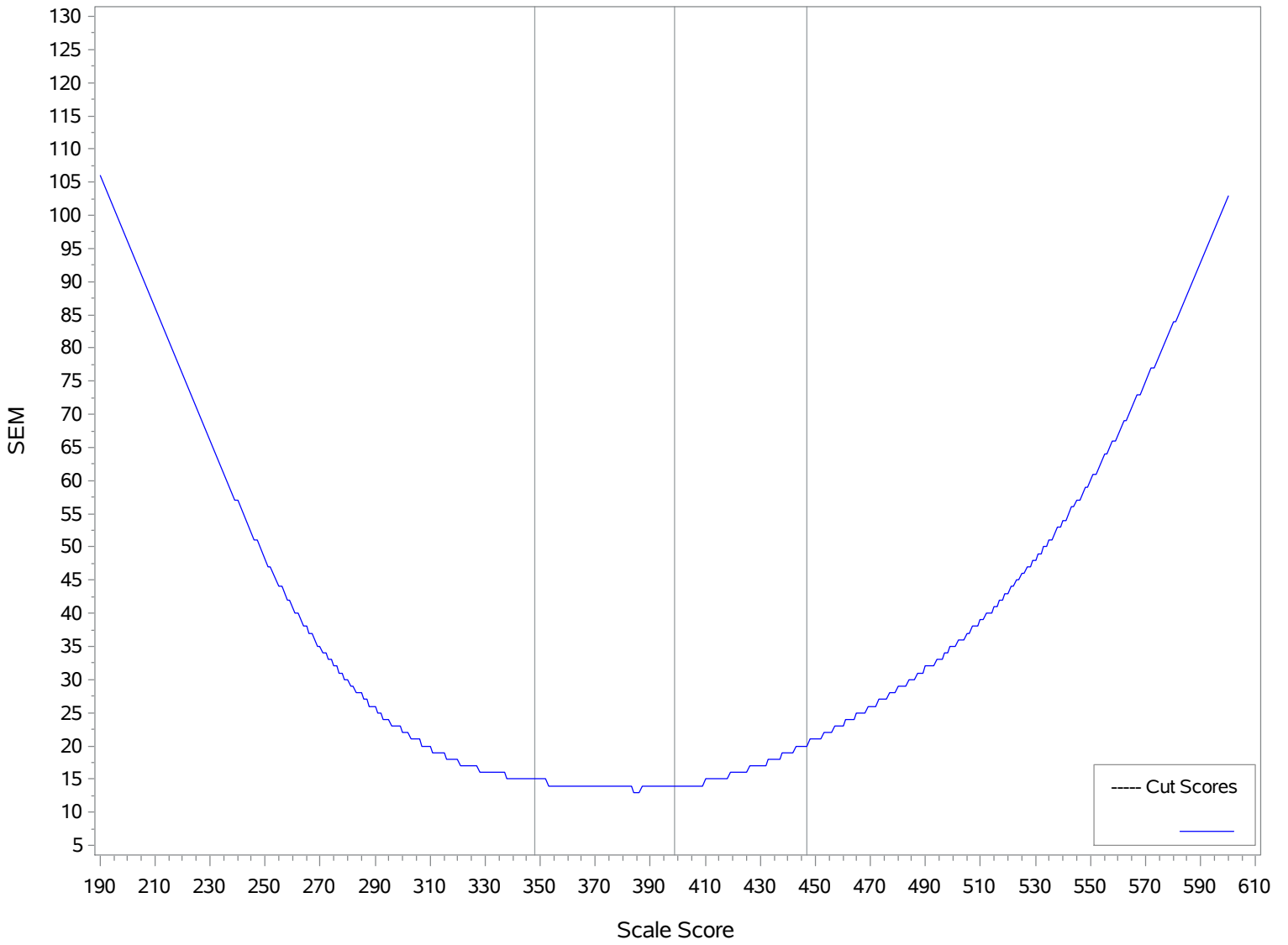


Figure I-14 CSEM with cut scores, Science Grade 8

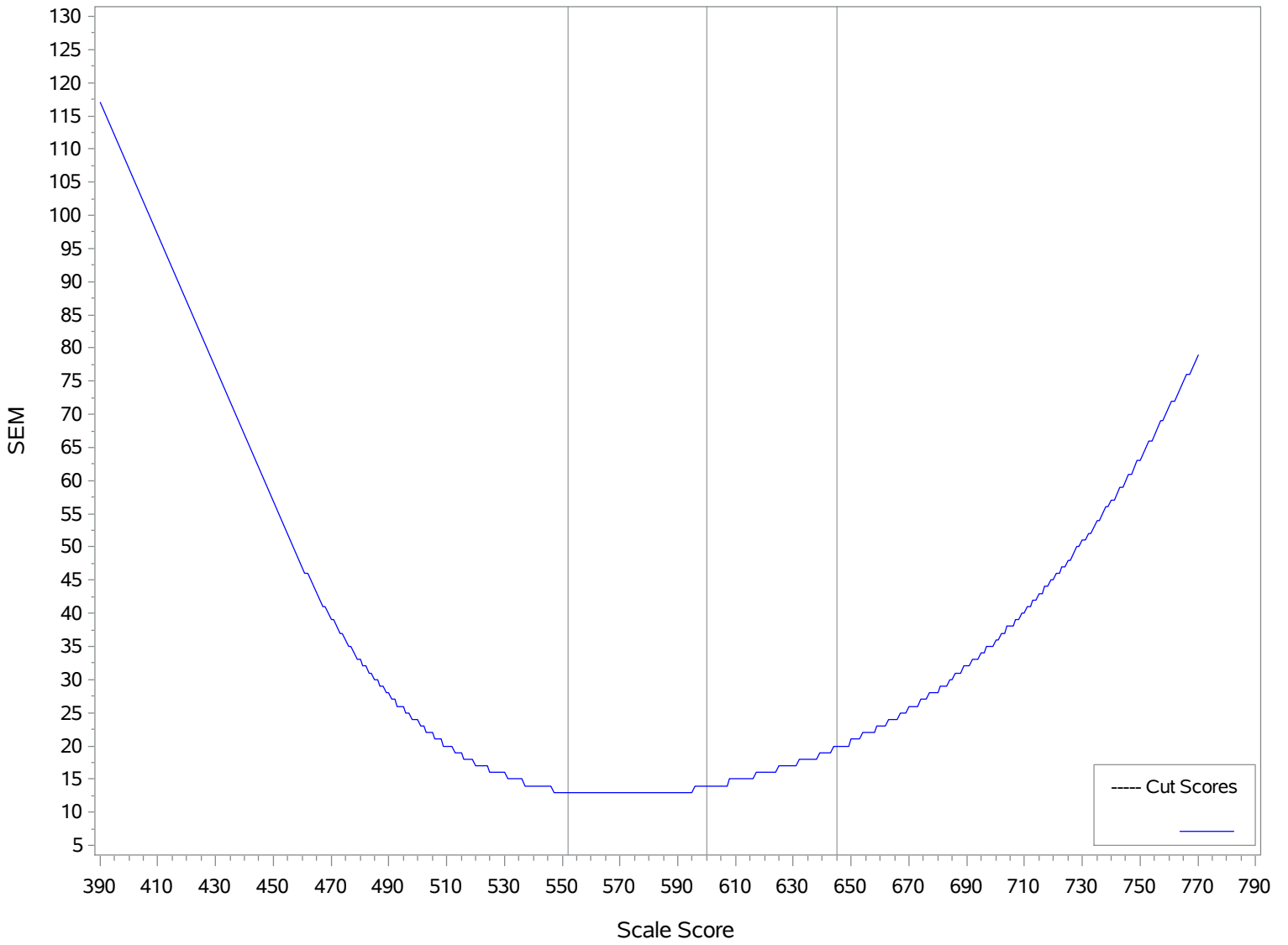


Figure I-15 CSEM with cut scores, Social Studies Grade 4

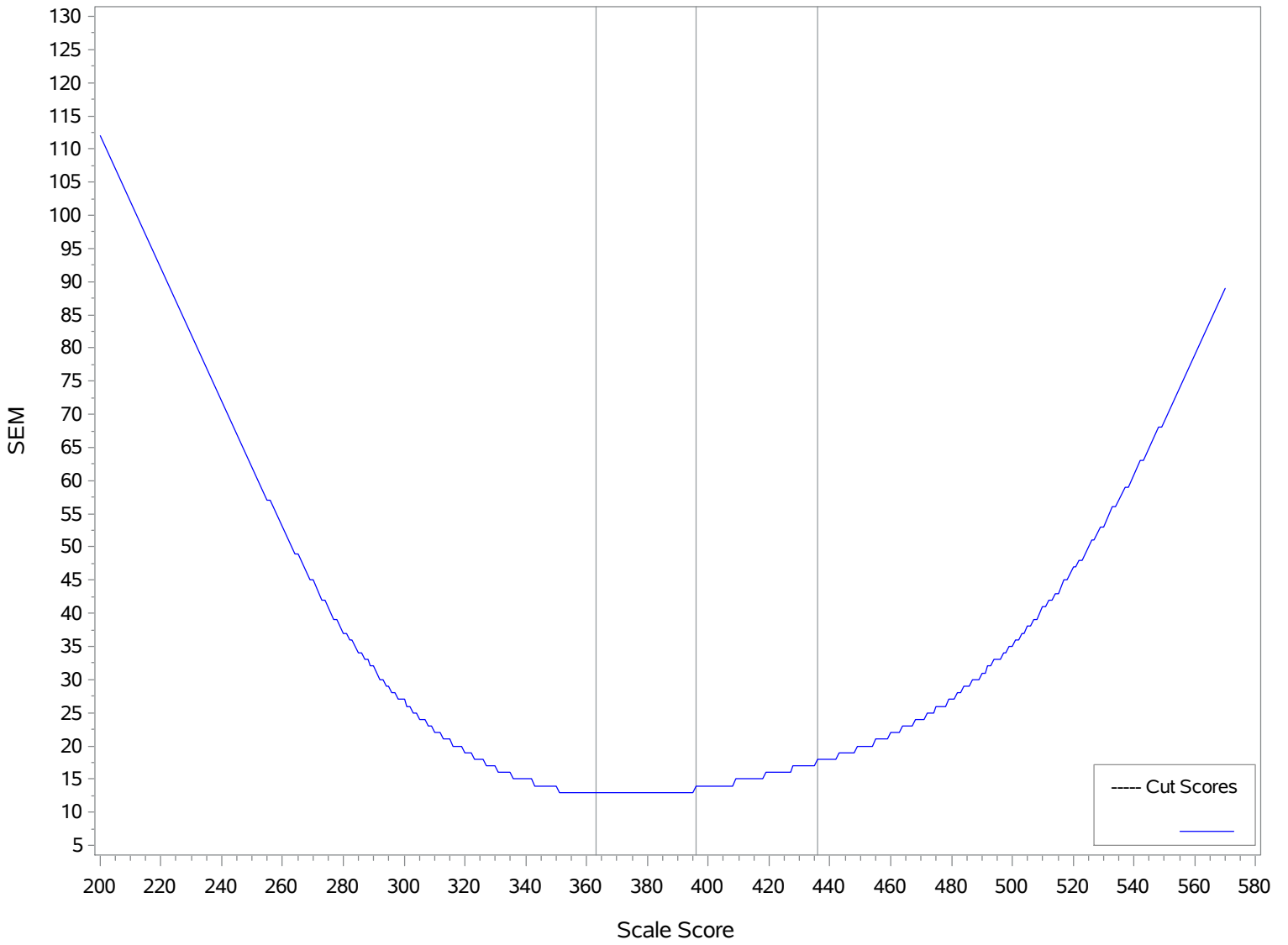


Figure I-16 CSEM with cut scores, Social Studies Grade 8

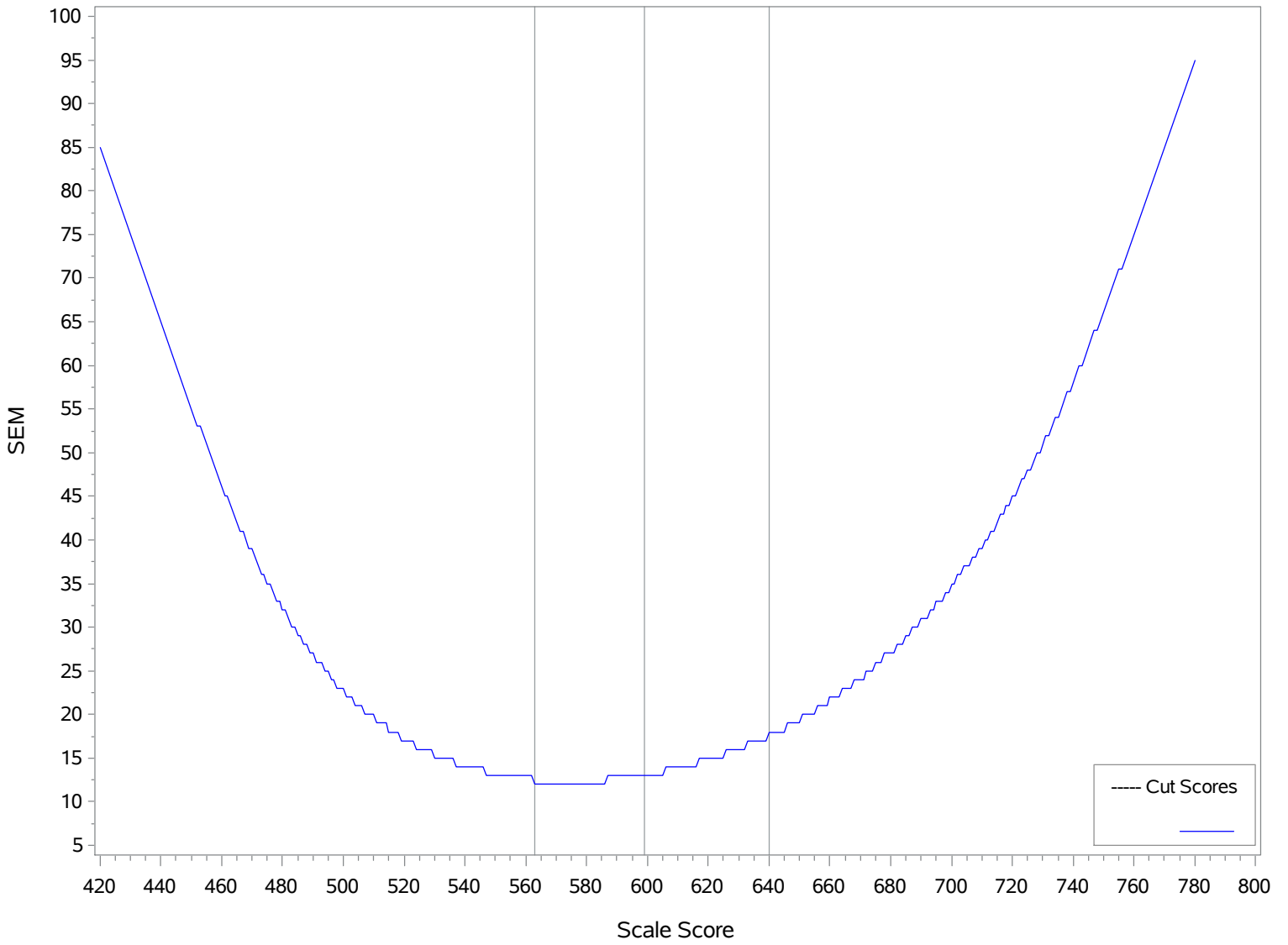
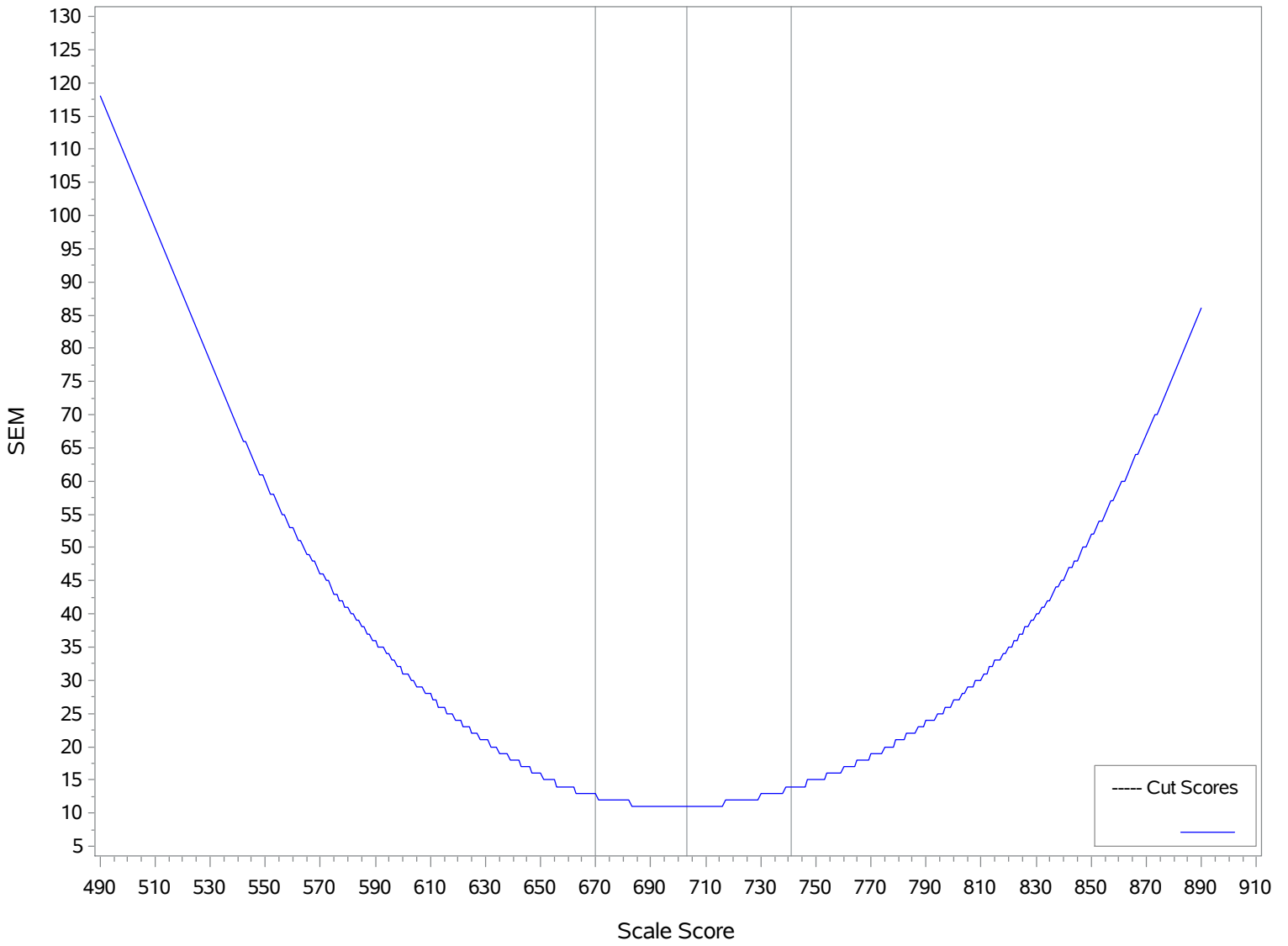


Figure I-17 CSEM with cut scores, Social Studies Grade 10



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)	0.91	0.89	0.93	0.73	
		Probability of Chance	0.66	0.51	0.82	0.29	
		Kappa (k)	0.73	0.77	0.62	0.62	
		Classification Accuracy	0.93	0.92	0.95	0.81	
	Male	Classification Consistency (P)	0.90	0.89	0.95	0.73	
		Probability of Chance	0.62	0.53	0.87	0.30	
		Kappa (k)	0.73	0.76	0.58	0.62	
		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	0.74	0.50	0.81	0.31	
		Kappa (k)	0.69	0.75	0.60	0.59	
		Classification Accuracy	0.94	0.91	0.95	0.80	
	African-American	Classification Consistency (P)	0.86	0.94	0.99	0.79	
		Probability of Chance	0.50	0.77	0.97	0.41	
		Kappa (k)	0.72	0.74	0.54	0.64	
		Classification Accuracy	0.90	0.96	0.99	0.85	
	Hispanic	Classification Consistency (P)	0.86	0.91	0.97	0.74	
		Probability of Chance	0.54	0.64	0.94	0.34	
		Kappa (k)	0.70	0.75	0.54	0.61	
		Classification Accuracy	0.90	0.94	0.98	0.82	
	Asian	Classification Consistency (P)	0.90	0.89	0.94	0.73	
		Probability of Chance	0.64	0.52	0.83	0.29	
		Kappa (k)	0.71	0.77	0.65	0.62	
		Classification Accuracy	0.93	0.92	0.96	0.81	
	American Indian	Classification Consistency (P)	0.87	0.90	0.98	0.75	
		Probability of Chance	0.55	0.64	0.95	0.34	
		Kappa (k)	0.71	0.73	0.48	0.62	
		Classification Accuracy	0.91	0.93	0.98	0.82	
	Two or More	Classification Consistency (P)	0.89	0.89	0.95	0.73	
		Probability of Chance	0.62	0.54	0.86	0.30	
		Kappa (k)	0.71	0.77	0.62	0.62	
		Classification Accuracy	0.92	0.93	0.96	0.81	
	Limited English Proficiency	Yes	Classification Consistency (P)	0.85	0.91	0.98	0.74
			Probability of Chance	0.52	0.70	0.97	0.36
			Kappa (k)	0.68	0.71	0.51	0.60
			Classification Accuracy	0.89	0.94	0.99	0.82
Disability Status	Yes	Classification Consistency (P)	0.86	0.93	0.98	0.77	
		Probability of Chance	0.50	0.72	0.95	0.38	
		Kappa (k)	0.72	0.76	0.55	0.63	
		Classification Accuracy	0.90	0.95	0.99	0.84	

Table J-1 Indexes for Classification Consistency and Accuracy, ELA Grade 3 (cont.)

Category	Group	Indexes	Cut 1	Cut 2	Cut 3	All Cuts
SES Disadvantaged	Yes	Classification Consistency (P)	0.87	0.90	0.97	0.75
		Probability of Chance	0.54	0.63	0.93	0.33
		Kappa (k)	0.72	0.75	0.56	0.62
		Classification Accuracy	0.91	0.93	0.98	0.82
Accommodation Use	Yes	Classification Consistency (P)	0.90	0.93	0.97	0.80
		Probability of Chance	0.50	0.61	0.95	0.34
		Kappa (k)	0.80	0.83	0.44	0.70
		Classification Accuracy	0.94	0.95	0.98	0.87

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	0.93	0.74
		Probability of Chance	0.66	0.50	0.80	0.29
		Kappa (k)	0.76	0.78	0.63	0.63
		Classification Accuracy	0.94	0.92	0.95	0.81
	Male	Classification Consistency (P)	0.91	0.89	0.94	0.74
		Probability of Chance	0.60	0.52	0.86	0.29
		Kappa (k)	0.77	0.78	0.60	0.64
		Classification Accuracy	0.93	0.93	0.96	0.82
Race/Ethnicity	White	Classification Consistency (P)	0.93	0.88	0.92	0.73
		Probability of Chance	0.73	0.50	0.79	0.30
		Kappa (k)	0.73	0.77	0.61	0.61
		Classification Accuracy	0.94	0.92	0.94	0.81
	African-American	Classification Consistency (P)	0.88	0.94	0.99	0.80
		Probability of Chance	0.51	0.76	0.97	0.43
		Kappa (k)	0.75	0.75	0.55	0.66
		Classification Accuracy	0.91	0.96	0.99	0.86
	Hispanic	Classification Consistency (P)	0.88	0.90	0.97	0.75
		Probability of Chance	0.54	0.60	0.93	0.32
		Kappa (k)	0.74	0.76	0.58	0.63
		Classification Accuracy	0.91	0.93	0.98	0.82
	Asian	Classification Consistency (P)	0.91	0.89	0.93	0.73
		Probability of Chance	0.64	0.51	0.81	0.28
		Kappa (k)	0.75	0.78	0.62	0.63
		Classification Accuracy	0.94	0.92	0.95	0.81
	American Indian	Classification Consistency (P)	0.87	0.91	0.97	0.75
		Probability of Chance	0.52	0.63	0.94	0.33
		Kappa (k)	0.74	0.75	0.48	0.63
		Classification Accuracy	0.91	0.93	0.98	0.82
Two or More	Classification Consistency (P)	0.91	0.90	0.94	0.74	
	Probability of Chance	0.61	0.51	0.83	0.29	
	Kappa (k)	0.76	0.79	0.63	0.64	
	Classification Accuracy	0.93	0.93	0.96	0.82	
Limited English Proficiency	Yes	Classification Consistency (P)	0.87	0.90	0.99	0.76
		Probability of Chance	0.51	0.67	0.97	0.35
		Kappa (k)	0.73	0.70	0.44	0.62
		Classification Accuracy	0.90	0.93	0.99	0.82
Disability Status	Yes	Classification Consistency (P)	0.88	0.93	0.98	0.79
		Probability of Chance	0.50	0.70	0.95	0.38
		Kappa (k)	0.77	0.77	0.63	0.67
		Classification Accuracy	0.91	0.95	0.99	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
SES Disadvantaged	Yes	Classification Consistency (P)	0.88	0.90	0.97	0.76
		Probability of Chance	0.53	0.60	0.93	0.32
		Kappa (k)	0.75	0.76	0.55	0.64
		Classification Accuracy	0.91	0.93	0.98	0.82
Accommodation Use	Yes	Classification Consistency (P)	0.92	0.93	0.97	0.82
		Probability of Chance	0.50	0.58	0.92	0.34
		Kappa (k)	0.85	0.83	0.63	0.73
		Classification Accuracy	0.94	0.96	0.98	0.87

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)	0.92	0.90	0.94	0.76
		Probability of Chance	0.70	0.50	0.82	0.30
		Kappa (k)	0.74	0.80	0.65	0.65
		Classification Accuracy	0.94	0.92	0.96	0.82
	Male	Classification Consistency (P)	0.91	0.90	0.95	0.76
		Probability of Chance	0.63	0.52	0.88	0.30
		Kappa (k)	0.76	0.79	0.61	0.66
		Classification Accuracy	0.93	0.92	0.97	0.82
Race/Ethnicity	White	Classification Consistency (P)	0.93	0.89	0.93	0.75
		Probability of Chance	0.75	0.50	0.82	0.32
		Kappa (k)	0.72	0.78	0.63	0.64
		Classification Accuracy	0.95	0.91	0.95	0.81
	African-American	Classification Consistency (P)	0.87	0.94	0.99	0.80
		Probability of Chance	0.50	0.75	0.97	0.40
		Kappa (k)	0.73	0.76	0.59	0.66
		Classification Accuracy	0.90	0.95	0.99	0.85
	Hispanic	Classification Consistency (P)	0.88	0.91	0.97	0.76
		Probability of Chance	0.56	0.59	0.93	0.32
		Kappa (k)	0.73	0.78	0.59	0.65
		Classification Accuracy	0.91	0.93	0.98	0.82
	Asian	Classification Consistency (P)	0.92	0.89	0.95	0.76
		Probability of Chance	0.66	0.51	0.84	0.30
		Kappa (k)	0.75	0.77	0.68	0.65
		Classification Accuracy	0.94	0.92	0.96	0.82
	American Indian	Classification Consistency (P)	0.85	0.91	0.98	0.74
		Probability of Chance	0.54	0.64	0.95	0.34
		Kappa (k)	0.67	0.74	0.65	0.61
		Classification Accuracy	0.89	0.94	0.99	0.82
Two or More	Classification Consistency (P)	0.90	0.89	0.96	0.74	
	Probability of Chance	0.64	0.52	0.87	0.30	
	Kappa (k)	0.72	0.76	0.68	0.63	
	Classification Accuracy	0.93	0.92	0.97	0.82	
Limited English Proficiency	Yes	Classification Consistency (P)	0.85	0.92	1.00	0.77
		Probability of Chance	0.50	0.78	0.99	0.40
		Kappa (k)	0.70	0.66	0.44	0.62
		Classification Accuracy	0.89	0.94	1.00	0.83
Disability Status	Yes	Classification Consistency (P)	0.87	0.94	0.99	0.80
		Probability of Chance	0.50	0.75	0.97	0.41
		Kappa (k)	0.74	0.78	0.61	0.67
		Classification Accuracy	0.91	0.96	0.99	0.85

Table J-3 Indexes for Classification Consistency and Accuracy, ELA Grade 5 (cont.)

Category	Group	Indexes	Cut 1	Cut 2	Cut 3	All Cuts
SES Disadvantaged	Yes	Classification Consistency (P)	0.88	0.91	0.98	0.77
		Probability of Chance	0.54	0.61	0.94	0.33
		Kappa (k)	0.74	0.77	0.58	0.66
		Classification Accuracy	0.91	0.93	0.98	0.83
Accommodation Use	Yes	Classification Consistency (P)	0.86	0.94	0.97	0.78
		Probability of Chance	0.50	0.67	0.93	0.36
		Kappa (k)	0.72	0.82	0.63	0.65
		Classification Accuracy	0.90	0.96	0.98	0.84

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)	0.92	0.88	0.90	0.70
		Probability of Chance	0.70	0.50	0.76	0.28
		Kappa (k)	0.73	0.75	0.60	0.58
		Classification Accuracy	0.94	0.91	0.93	0.78
	Male	Classification Consistency (P)	0.90	0.89	0.93	0.72
		Probability of Chance	0.61	0.53	0.82	0.28
		Kappa (k)	0.75	0.76	0.59	0.60
		Classification Accuracy	0.93	0.92	0.95	0.80
Race/Ethnicity	White	Classification Consistency (P)	0.93	0.87	0.90	0.69
		Probability of Chance	0.74	0.50	0.75	0.29
		Kappa (k)	0.71	0.74	0.59	0.57
		Classification Accuracy	0.95	0.90	0.93	0.78
	African-American	Classification Consistency (P)	0.87	0.93	0.98	0.78
		Probability of Chance	0.50	0.76	0.96	0.41
		Kappa (k)	0.73	0.72	0.48	0.63
		Classification Accuracy	0.91	0.95	0.99	0.84
	Hispanic	Classification Consistency (P)	0.87	0.89	0.96	0.72
		Probability of Chance	0.55	0.61	0.91	0.32
		Kappa (k)	0.72	0.73	0.53	0.59
		Classification Accuracy	0.91	0.92	0.97	0.80
	Asian	Classification Consistency (P)	0.91	0.88	0.92	0.70
		Probability of Chance	0.66	0.51	0.77	0.27
		Kappa (k)	0.73	0.75	0.65	0.59
		Classification Accuracy	0.94	0.92	0.94	0.79
	American Indian	Classification Consistency (P)	0.87	0.91	0.97	0.75
		Probability of Chance	0.52	0.66	0.93	0.34
		Kappa (k)	0.72	0.73	0.58	0.62
		Classification Accuracy	0.91	0.93	0.98	0.83
Two or More	Classification Consistency (P)	0.90	0.87	0.93	0.71	
	Probability of Chance	0.62	0.53	0.82	0.29	
	Kappa (k)	0.74	0.73	0.62	0.59	
	Classification Accuracy	0.93	0.91	0.95	0.79	
Limited English Proficiency	Yes	Classification Consistency (P)	0.82	0.93	0.99	0.75
		Probability of Chance	0.50	0.85	0.99	0.44
		Kappa (k)	0.64	0.57	0.44	0.55
		Classification Accuracy	0.88	0.95	1.00	0.83
Disability Status	Yes	Classification Consistency (P)	0.87	0.95	0.98	0.80
		Probability of Chance	0.52	0.80	0.96	0.45
		Kappa (k)	0.72	0.72	0.58	0.63
		Classification Accuracy	0.91	0.96	0.99	0.86

Table J-4 Indexes for Classification Consistency and Accuracy, ELA Grade 6 (cont.)

Category	Group	Indexes	Cut 1	Cut 2	Cut 3	All Cuts
SES Disadvantaged	Yes	Classification Consistency (P)	0.88	0.90	0.96	0.73
		Probability of Chance	0.54	0.62	0.91	0.32
		Kappa (k)	0.73	0.73	0.53	0.60
		Classification Accuracy	0.91	0.92	0.97	0.81
Accommodation Use	Yes	Classification Consistency (P)	0.89	0.94	0.97	0.80
		Probability of Chance	0.52	0.71	0.94	0.42
		Kappa (k)	0.78	0.80	0.51	0.66
		Classification Accuracy	0.92	0.96	0.98	0.86

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	0.89	0.92	0.73
		Probability of Chance	0.72	0.50	0.77	0.29
		Kappa (k)	0.72	0.77	0.68	0.62
		Classification Accuracy	0.95	0.92	0.94	0.80
	Male	Classification Consistency (P)	0.90	0.89	0.94	0.74
		Probability of Chance	0.61	0.52	0.84	0.29
		Kappa (k)	0.74	0.78	0.66	0.63
		Classification Accuracy	0.93	0.92	0.96	0.81
Race/Ethnicity	White	Classification Consistency (P)	0.92	0.88	0.92	0.73
		Probability of Chance	0.74	0.50	0.77	0.30
		Kappa (k)	0.69	0.76	0.66	0.61
		Classification Accuracy	0.95	0.91	0.94	0.80
	African-American	Classification Consistency (P)	0.87	0.93	0.98	0.78
		Probability of Chance	0.50	0.72	0.96	0.39
		Kappa (k)	0.73	0.76	0.63	0.64
		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	0.56	0.57	0.90	0.31
		Kappa (k)	0.72	0.76	0.63	0.63
		Classification Accuracy	0.92	0.93	0.97	0.82
	Asian	Classification Consistency (P)	0.92	0.89	0.92	0.73
		Probability of Chance	0.71	0.50	0.76	0.28
		Kappa (k)	0.72	0.78	0.66	0.62
		Classification Accuracy	0.94	0.92	0.94	0.81
	American Indian	Classification Consistency (P)	0.86	0.88	0.98	0.73
		Probability of Chance	0.56	0.63	0.97	0.35
		Kappa (k)	0.69	0.66	0.51	0.58
		Classification Accuracy	0.89	0.91	0.98	0.78
Two or More	Classification Consistency (P)	0.90	0.89	0.94	0.73	
	Probability of Chance	0.62	0.52	0.84	0.29	
	Kappa (k)	0.73	0.77	0.63	0.62	
	Classification Accuracy	0.93	0.92	0.96	0.81	
Limited English Proficiency	Yes	Classification Consistency (P)	0.83	0.95	1.00	0.77
		Probability of Chance	0.51	0.84	0.99	0.45
		Kappa (k)	0.66	0.66	0.49	0.59
		Classification Accuracy	0.89	0.96	1.00	0.84
Disability Status	Yes	Classification Consistency (P)	0.86	0.95	0.99	0.79
		Probability of Chance	0.52	0.80	0.97	0.45
		Kappa (k)	0.70	0.75	0.61	0.63
		Classification Accuracy	0.90	0.96	0.99	0.86

Table J-5 Indexes for Classification Consistency and Accuracy, ELA Grade 7 (cont.)

Category	Group	Indexes	Cut 1	Cut 2	Cut 3	All Cuts
SES Disadvantaged	Yes	Classification Consistency (P)	0.87	0.90	0.97	0.75
		Probability of Chance	0.54	0.59	0.92	0.32
		Kappa (k)	0.73	0.76	0.62	0.63
		Classification Accuracy	0.91	0.93	0.98	0.82
Accommodation Use	Yes	Classification Consistency (P)	0.87	0.94	0.96	0.77
		Probability of Chance	0.50	0.67	0.91	0.37
		Kappa (k)	0.74	0.82	0.55	0.64
		Classification Accuracy	0.91	0.95	0.97	0.83

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	0.92	0.73
		Probability of Chance	0.68	0.51	0.77	0.28
		Kappa (k)	0.77	0.76	0.66	0.63
		Classification Accuracy	0.95	0.91	0.94	0.80
	Male	Classification Consistency (P)	0.94	0.93	0.92	0.79
		Probability of Chance	0.55	0.51	0.75	0.27
		Kappa (k)	0.87	0.85	0.69	0.72
		Classification Accuracy	0.95	0.90	0.92	0.76
Race/Ethnicity	White	Classification Consistency (P)	0.93	0.88	0.92	0.73
		Probability of Chance	0.71	0.51	0.78	0.29
		Kappa (k)	0.76	0.75	0.64	0.61
		Classification Accuracy	0.95	0.91	0.94	0.80
	African-American	Classification Consistency (P)	0.89	0.95	0.98	0.82
		Probability of Chance	0.51	0.80	0.96	0.44
		Kappa (k)	0.77	0.73	0.58	0.68
		Classification Accuracy	0.92	0.96	0.99	0.87
	Hispanic	Classification Consistency (P)	0.89	0.91	0.96	0.77
		Probability of Chance	0.53	0.65	0.91	0.34
		Kappa (k)	0.77	0.75	0.60	0.65
		Classification Accuracy	0.92	0.93	0.98	0.83
	Asian	Classification Consistency (P)	0.92	0.88	0.92	0.72
		Probability of Chance	0.67	0.51	0.76	0.28
		Kappa (k)	0.75	0.76	0.68	0.62
		Classification Accuracy	0.95	0.92	0.95	0.81
	American Indian	Classification Consistency (P)	0.89	0.91	0.97	0.77
		Probability of Chance	0.51	0.68	0.93	0.35
		Kappa (k)	0.77	0.71	0.57	0.64
		Classification Accuracy	0.92	0.93	0.98	0.84
Two or More	Classification Consistency (P)	0.91	0.89	0.94	0.75	
	Probability of Chance	0.60	0.55	0.83	0.29	
	Kappa (k)	0.78	0.75	0.66	0.64	
	Classification Accuracy	0.94	0.92	0.96	0.82	
Limited English Proficiency	Yes	Classification Consistency (P)	0.86	0.97	1.00	0.83
		Probability of Chance	0.54	0.92	0.99	0.52
		Kappa (k)	0.69	0.60	0.49	0.64
		Classification Accuracy	0.89	0.98	1.00	0.87
Disability Status	Yes	Classification Consistency (P)	0.89	0.96	0.99	0.84
		Probability of Chance	0.55	0.86	0.98	0.50
		Kappa (k)	0.76	0.73	0.57	0.68
		Classification Accuracy	0.92	0.97	0.99	0.88

Table J-6 Indexes for Classification Consistency and Accuracy, ELA Grade 8 (cont.)

Category	Group	Indexes	Cut 1	Cut 2	Cut 3	All Cuts
SES Disadvantaged	Yes	Classification Consistency (P)	0.90	0.91	0.97	0.78
		Probability of Chance	0.52	0.66	0.92	0.34
		Kappa (k)	0.78	0.74	0.60	0.66
		Classification Accuracy	0.92	0.93	0.98	0.83
Accommodation Use	Yes	Classification Consistency (P)	0.90	0.93	0.96	0.79
		Probability of Chance	0.50	0.70	0.92	0.37
		Kappa (k)	0.80	0.76	0.50	0.67
		Classification Accuracy	0.93	0.95	0.97	0.85

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)	0.92	0.89	0.94	0.75
		Probability of Chance	0.69	0.50	0.80	0.29
		Kappa (k)	0.75	0.78	0.68	0.64
		Classification Accuracy	0.94	0.92	0.95	0.82
	Male	Classification Consistency (P)	0.93	0.90	0.93	0.75
		Probability of Chance	0.69	0.50	0.76	0.28
		Kappa (k)	0.77	0.79	0.70	0.65
		Classification Accuracy	0.95	0.93	0.95	0.82
Race/Ethnicity	White	Classification Consistency (P)	0.94	0.89	0.91	0.74
		Probability of Chance	0.80	0.52	0.72	0.31
		Kappa (k)	0.72	0.76	0.68	0.63
		Classification Accuracy	0.96	0.92	0.94	0.82
	African-American	Classification Consistency (P)	0.87	0.93	0.99	0.79
		Probability of Chance	0.50	0.71	0.97	0.38
		Kappa (k)	0.74	0.75	0.58	0.66
		Classification Accuracy	0.91	0.95	0.99	0.84
	Hispanic	Classification Consistency (P)	0.89	0.89	0.97	0.75
		Probability of Chance	0.57	0.57	0.91	0.31
		Kappa (k)	0.73	0.76	0.65	0.64
		Classification Accuracy	0.92	0.92	0.98	0.82
	Asian	Classification Consistency (P)	0.92	0.89	0.93	0.75
		Probability of Chance	0.69	0.50	0.73	0.27
		Kappa (k)	0.73	0.79	0.76	0.65
		Classification Accuracy	0.94	0.92	0.95	0.81
	American Indian	Classification Consistency (P)	0.89	0.91	0.97	0.77
		Probability of Chance	0.56	0.58	0.92	0.31
		Kappa (k)	0.74	0.79	0.65	0.66
		Classification Accuracy	0.91	0.94	0.98	0.83
Two or More	Classification Consistency (P)	0.91	0.90	0.95	0.75	
	Probability of Chance	0.64	0.51	0.82	0.29	
	Kappa (k)	0.74	0.80	0.71	0.65	
	Classification Accuracy	0.93	0.93	0.96	0.82	
Limited English Proficiency	Yes	Classification Consistency (P)	0.88	0.90	0.97	0.75
		Probability of Chance	0.56	0.59	0.93	0.32
		Kappa (k)	0.73	0.75	0.63	0.63
		Classification Accuracy	0.91	0.93	0.98	0.82
Disability Status	Yes	Classification Consistency (P)	0.89	0.93	0.97	0.79
		Probability of Chance	0.51	0.62	0.91	0.33
		Kappa (k)	0.78	0.80	0.69	0.68
		Classification Accuracy	0.92	0.95	0.98	0.84

Table J-7 Indexes for Classification Consistency and Accuracy, Mathematics Grade 3 (cont.)

Category	Group	Indexes	Cut 1	Cut 2	Cut 3	All Cuts
SES Disadvantaged	Yes	Classification Consistency (P)	0.89	0.90	0.97	0.76
		Probability of Chance	0.57	0.56	0.91	0.31
		Kappa (k)	0.75	0.77	0.64	0.65
		Classification Accuracy	0.92	0.93	0.98	0.82
Accommodation Use	Yes	Classification Consistency (P)	0.87	0.95	1.00	0.82
		Probability of Chance	0.52	0.83	0.99	0.47
		Kappa (k)	0.74	0.70	0.55	0.66
		Classification Accuracy	0.91	0.96	1.00	0.87

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	0.89	0.95	0.74
		Probability of Chance	0.68	0.51	0.81	0.29
		Kappa (k)	0.67	0.78	0.75	0.63
		Classification Accuracy	0.92	0.93	0.97	0.82
	Male	Classification Consistency (P)	0.93	0.95	0.96	0.84
		Probability of Chance	0.63	0.51	0.66	0.26
		Kappa (k)	0.80	0.90	0.88	0.78
		Classification Accuracy	0.92	0.96	0.91	0.79
Race/Ethnicity	White	Classification Consistency (P)	0.93	0.89	0.93	0.75
		Probability of Chance	0.79	0.51	0.74	0.30
		Kappa (k)	0.65	0.77	0.75	0.64
		Classification Accuracy	0.95	0.92	0.96	0.82
	African-American	Classification Consistency (P)	0.82	0.94	0.99	0.75
		Probability of Chance	0.50	0.78	0.97	0.40
		Kappa (k)	0.63	0.73	0.69	0.58
		Classification Accuracy	0.87	0.96	0.99	0.82
	Hispanic	Classification Consistency (P)	0.85	0.91	0.98	0.73
		Probability of Chance	0.57	0.61	0.92	0.33
		Kappa (k)	0.66	0.76	0.72	0.61
		Classification Accuracy	0.89	0.94	0.98	0.81
	Asian	Classification Consistency (P)	0.89	0.90	0.95	0.74
		Probability of Chance	0.69	0.50	0.72	0.27
		Kappa (k)	0.64	0.80	0.81	0.64
		Classification Accuracy	0.92	0.93	0.96	0.82
	American Indian	Classification Consistency (P)	0.84	0.89	0.98	0.71
		Probability of Chance	0.57	0.64	0.93	0.34
		Kappa (k)	0.62	0.70	0.68	0.56
		Classification Accuracy	0.89	0.93	0.98	0.80
Two or More	Classification Consistency (P)	0.89	0.90	0.95	0.74	
	Probability of Chance	0.66	0.52	0.81	0.29	
	Kappa (k)	0.67	0.79	0.75	0.63	
	Classification Accuracy	0.92	0.93	0.97	0.82	
Limited English Proficiency	Yes	Classification Consistency (P)	0.83	0.91	0.98	0.73
		Probability of Chance	0.54	0.67	0.94	0.35
		Kappa (k)	0.64	0.74	0.72	0.59
		Classification Accuracy	0.88	0.94	0.99	0.81
Disability Status	Yes	Classification Consistency (P)	0.84	0.93	0.98	0.75
		Probability of Chance	0.51	0.67	0.91	0.35
		Kappa (k)	0.68	0.80	0.75	0.62
		Classification Accuracy	0.89	0.96	0.99	0.83

Table J-8 Indexes for Classification Consistency and Accuracy, Mathematics Grade 4 (cont.)

Category	Group	Indexes	Cut 1	Cut 2	Cut 3	All Cuts
SES Disadvantaged	Yes	Classification Consistency (P)	0.85	0.90	0.98	0.73
		Probability of Chance	0.57	0.61	0.92	0.32
		Kappa (k)	0.66	0.76	0.71	0.61
		Classification Accuracy	0.90	0.93	0.98	0.81
Accommodation Use	Yes	Classification Consistency (P)	0.81	0.96	1.00	0.76
		Probability of Chance	0.52	0.88	0.99	0.48
		Kappa (k)	0.60	0.66	0.45	0.55
		Classification Accuracy	0.87	0.97	1.00	0.84

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)	0.90	0.90	0.95	0.75
		Probability of Chance	0.63	0.50	0.81	0.28
		Kappa (k)	0.73	0.80	0.74	0.65
		Classification Accuracy	0.93	0.93	0.97	0.82
	Male	Classification Consistency (P)	0.91	0.91	0.95	0.76
		Probability of Chance	0.61	0.50	0.77	0.27
		Kappa (k)	0.76	0.82	0.76	0.67
		Classification Accuracy	0.93	0.93	0.96	0.83
Race/Ethnicity	White	Classification Consistency (P)	0.92	0.90	0.94	0.75
		Probability of Chance	0.72	0.51	0.75	0.29
		Kappa (k)	0.71	0.79	0.74	0.65
		Classification Accuracy	0.94	0.92	0.96	0.82
	African-American	Classification Consistency (P)	0.86	0.94	0.99	0.80
		Probability of Chance	0.52	0.77	0.97	0.44
		Kappa (k)	0.71	0.75	0.69	0.63
		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	0.60	0.92	0.32
		Kappa (k)	0.71	0.78	0.72	0.64
		Classification Accuracy	0.90	0.93	0.98	0.82
	Asian	Classification Consistency (P)	0.90	0.90	0.95	0.75
		Probability of Chance	0.65	0.50	0.75	0.27
		Kappa (k)	0.71	0.80	0.81	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	0.62	0.93	0.33
		Kappa (k)	0.70	0.78	0.72	0.63
		Classification Accuracy	0.90	0.94	0.99	0.83
Two or More	Classification Consistency (P)	0.89	0.91	0.96	0.75	
	Probability of Chance	0.58	0.52	0.82	0.28	
	Kappa (k)	0.73	0.80	0.76	0.65	
	Classification Accuracy	0.92	0.94	0.97	0.83	
Limited English Proficiency	Yes	Classification Consistency (P)	0.84	0.92	0.99	0.75
		Probability of Chance	0.50	0.72	0.98	0.38
		Kappa (k)	0.68	0.70	0.63	0.60
		Classification Accuracy	0.89	0.94	0.99	0.82
Disability Status	Yes	Classification Consistency (P)	0.87	0.94	0.99	0.80
		Probability of Chance	0.51	0.71	0.94	0.41
		Kappa (k)	0.74	0.80	0.78	0.66
		Classification Accuracy	0.91	0.96	0.99	0.86

Table J-9 Indexes for Classification Consistency and Accuracy, Mathematics Grade 5 (cont.)

Category	Group	Indexes	Cut 1	Cut 2	Cut 3	All Cuts
SES Disadvantaged	Yes	Classification Consistency (P)	0.87	0.91	0.98	0.76
		Probability of Chance	0.52	0.60	0.92	0.32
		Kappa (k)	0.73	0.78	0.71	0.65
		Classification Accuracy	0.91	0.94	0.98	0.83
Accommodation Use	Yes	Classification Consistency (P)	0.87	0.96	1.00	0.83
		Probability of Chance	0.59	0.87	0.99	0.56
		Kappa (k)	0.68	0.71	0.73	0.61
		Classification Accuracy	0.91	0.97	1.00	0.88

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)	0.89	0.89	0.96	0.75
		Probability of Chance	0.63	0.51	0.88	0.31
		Kappa (k)	0.71	0.78	0.69	0.64
		Classification Accuracy	0.93	0.93	0.97	0.83
	Male	Classification Consistency (P)	0.90	0.90	0.96	0.76
		Probability of Chance	0.60	0.51	0.86	0.30
		Kappa (k)	0.74	0.81	0.71	0.66
		Classification Accuracy	0.93	0.94	0.97	0.84
Race/Ethnicity	White	Classification Consistency (P)	0.91	0.89	0.95	0.75
		Probability of Chance	0.72	0.50	0.84	0.32
		Kappa (k)	0.69	0.78	0.69	0.63
		Classification Accuracy	0.94	0.93	0.97	0.83
	African-American	Classification Consistency (P)	0.85	0.94	1.00	0.79
		Probability of Chance	0.52	0.79	0.99	0.45
		Kappa (k)	0.70	0.73	0.62	0.62
		Classification Accuracy	0.90	0.96	1.00	0.86
	Hispanic	Classification Consistency (P)	0.85	0.92	0.99	0.76
		Probability of Chance	0.52	0.63	0.96	0.34
		Kappa (k)	0.69	0.79	0.63	0.64
		Classification Accuracy	0.90	0.94	0.99	0.83
	Asian	Classification Consistency (P)	0.90	0.90	0.96	0.76
		Probability of Chance	0.65	0.50	0.80	0.28
		Kappa (k)	0.71	0.80	0.80	0.66
		Classification Accuracy	0.93	0.93	0.97	0.83
	American Indian	Classification Consistency (P)	0.85	0.93	0.99	0.77
		Probability of Chance	0.50	0.69	0.97	0.37
		Kappa (k)	0.69	0.78	0.62	0.63
		Classification Accuracy	0.89	0.95	0.99	0.83
Two or More	Classification Consistency (P)	0.88	0.92	0.97	0.77	
	Probability of Chance	0.58	0.54	0.90	0.30	
	Kappa (k)	0.71	0.82	0.75	0.67	
	Classification Accuracy	0.91	0.94	0.98	0.83	
Limited English Proficiency	Yes	Classification Consistency (P)	0.83	0.95	1.00	0.77
		Probability of Chance	0.51	0.83	0.99	0.44
		Kappa (k)	0.64	0.70	0.72	0.59
		Classification Accuracy	0.88	0.96	1.00	0.84
Disability Status	Yes	Classification Consistency (P)	0.86	0.95	0.99	0.80
		Probability of Chance	0.53	0.78	0.97	0.47
		Kappa (k)	0.69	0.79	0.69	0.63
		Classification Accuracy	0.90	0.97	0.99	0.86

Table J-10 Indexes for Classification Consistency and Accuracy, Mathematics Grade 6 (cont.)

Category	Group	Indexes	Cut 1	Cut 2	Cut 3	All Cuts
SES Disadvantaged	Yes	Classification Consistency (P)	0.85	0.91	0.99	0.76
		Probability of Chance	0.52	0.62	0.96	0.34
		Kappa (k)	0.70	0.78	0.64	0.63
		Classification Accuracy	0.90	0.94	0.99	0.83
Accommodation Use	Yes	Classification Consistency (P)	0.86	0.98	1.00	0.83
		Probability of Chance	0.64	0.93	1.00	0.62
		Kappa (k)	0.60	0.68	0.72	0.56
		Classification Accuracy	0.91	0.98	1.00	0.89

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	0.70	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	0.56	0.52	0.89	0.30
		Kappa (k)	0.74	0.80	0.71	0.65
		Classification Accuracy	0.92	0.93	0.98	0.83
Race/Ethnicity	White	Classification Consistency (P)	0.89	0.89	0.96	0.75
		Probability of Chance	0.65	0.50	0.88	0.31
		Kappa (k)	0.70	0.78	0.69	0.63
		Classification Accuracy	0.93	0.92	0.97	0.82
	African-American	Classification Consistency (P)	0.85	0.96	1.00	0.81
		Probability of Chance	0.56	0.83	0.99	0.52
		Kappa (k)	0.66	0.75	0.61	0.60
		Classification Accuracy	0.90	0.97	1.00	0.87
	Hispanic	Classification Consistency (P)	0.85	0.92	0.99	0.76
		Probability of Chance	0.50	0.67	0.97	0.37
		Kappa (k)	0.70	0.77	0.66	0.62
		Classification Accuracy	0.90	0.94	0.99	0.84
	Asian	Classification Consistency (P)	0.89	0.92	0.96	0.78
		Probability of Chance	0.58	0.51	0.82	0.28
		Kappa (k)	0.75	0.85	0.78	0.69
		Classification Accuracy	0.92	0.94	0.97	0.84
	American Indian	Classification Consistency (P)	0.90	0.95	0.99	0.84
		Probability of Chance	0.52	0.71	0.97	0.44
		Kappa (k)	0.79	0.83	0.67	0.72
		Classification Accuracy	0.88	0.97	0.99	0.83
Two or More	Classification Consistency (P)	0.87	0.91	0.98	0.76	
	Probability of Chance	0.54	0.56	0.92	0.31	
	Kappa (k)	0.71	0.80	0.74	0.65	
	Classification Accuracy	0.91	0.94	0.99	0.83	
Limited English Proficiency	Yes	Classification Consistency (P)	0.84	0.96	1.00	0.80
		Probability of Chance	0.58	0.88	0.99	0.55
		Kappa (k)	0.62	0.68	0.75	0.56
		Classification Accuracy	0.90	0.97	1.00	0.87
Disability Status	Yes	Classification Consistency (P)	0.85	0.96	1.00	0.81
		Probability of Chance	0.58	0.83	0.98	0.53
		Kappa (k)	0.65	0.78	0.69	0.59
		Classification Accuracy	0.90	0.97	1.00	0.87

Table J-11 Indexes for Classification Consistency and Accuracy, Mathematics Grade 7

Category	Group	Indexes	Cut 1	Cut 2	Cut 3	All Cuts
SES Disadvantaged	Yes	Classification Consistency (P)	0.85	0.92	0.99	0.76
		Probability of Chance	0.50	0.67	0.97	0.37
		Kappa (k)	0.70	0.76	0.65	0.62
		Classification Accuracy	0.90	0.94	0.99	0.84
Accommodation Use	Yes	Classification Consistency (P)	0.85	0.98	1.00	0.83
		Probability of Chance	0.69	0.95	1.00	0.68
		Kappa (k)	0.52	0.67	0.26	0.48
		Classification Accuracy	0.90	0.99	1.00	0.89

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)	0.88	0.89	0.96	0.73
		Probability of Chance	0.62	0.53	0.85	0.29
		Kappa (k)	0.69	0.77	0.72	0.62
		Classification Accuracy	0.92	0.93	0.97	0.81
	Male	Classification Consistency (P)	0.88	0.91	0.96	0.74
		Probability of Chance	0.58	0.54	0.84	0.29
		Kappa (k)	0.71	0.80	0.74	0.64
		Classification Accuracy	0.91	0.94	0.97	0.82
Race/Ethnicity	White	Classification Consistency (P)	0.89	0.89	0.95	0.73
		Probability of Chance	0.68	0.51	0.81	0.29
		Kappa (k)	0.67	0.77	0.72	0.62
		Classification Accuracy	0.93	0.92	0.96	0.81
	African-American	Classification Consistency (P)	0.83	0.96	0.99	0.79
		Probability of Chance	0.54	0.85	0.98	0.49
		Kappa (k)	0.63	0.75	0.70	0.58
		Classification Accuracy	0.88	0.97	1.00	0.85
	Hispanic	Classification Consistency (P)	0.84	0.93	0.98	0.75
		Probability of Chance	0.51	0.70	0.94	0.36
		Kappa (k)	0.67	0.76	0.68	0.61
		Classification Accuracy	0.89	0.95	0.99	0.83
	Asian	Classification Consistency (P)	0.89	0.91	0.95	0.75
		Probability of Chance	0.63	0.51	0.76	0.27
		Kappa (k)	0.70	0.81	0.80	0.66
		Classification Accuracy	0.92	0.93	0.97	0.82
	American Indian	Classification Consistency (P)	0.84	0.94	0.99	0.78
		Probability of Chance	0.50	0.72	0.96	0.38
		Kappa (k)	0.69	0.80	0.74	0.64
		Classification Accuracy	0.89	0.96	0.99	0.84
Two or More	Classification Consistency (P)	0.86	0.91	0.97	0.74	
	Probability of Chance	0.55	0.57	0.87	0.30	
	Kappa (k)	0.69	0.80	0.74	0.63	
	Classification Accuracy	0.90	0.94	0.98	0.82	
Limited English Proficiency	Yes	Classification Consistency (P)	0.82	0.97	1.00	0.79
		Probability of Chance	0.55	0.90	0.99	0.51
		Kappa (k)	0.61	0.71	0.59	0.57
		Classification Accuracy	0.88	0.98	1.00	0.86
Disability Status	Yes	Classification Consistency (P)	0.83	0.97	0.99	0.79
		Probability of Chance	0.55	0.86	0.98	0.51
		Kappa (k)	0.62	0.77	0.71	0.57
		Classification Accuracy	0.88	0.98	1.00	0.86

Table J-12 Indexes for Classification Consistency and Accuracy, Mathematics Grade 8 (cont.)

Category	Group	Indexes	Cut 1	Cut 2	Cut 3	All Cuts
SES Disadvantaged	Yes	Classification Consistency (P)	0.84	0.93	0.98	0.75
		Probability of Chance	0.51	0.69	0.95	0.36
		Kappa (k)	0.67	0.76	0.69	0.61
		Classification Accuracy	0.89	0.95	0.99	0.82
Accommodation Use	Yes	Classification Consistency (P)	0.90	0.99	1.00	0.89
		Probability of Chance	0.74	0.97	1.00	0.74
		Kappa (k)	0.60	0.72	0.69	0.57
		Classification Accuracy	0.92	0.99	1.00	0.91

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	0.89	0.69
		Probability of Chance	0.74	0.50	0.70	0.28
		Kappa (k)	0.72	0.75	0.61	0.58
		Classification Accuracy	0.95	0.91	0.92	0.78
	Male	Classification Consistency (P)	0.96	0.95	0.94	0.85
		Probability of Chance	0.71	0.52	0.52	0.28
		Kappa (k)	0.87	0.88	0.88	0.79
		Classification Accuracy	0.97	0.89	0.77	0.63
Race/Ethnicity	White	Classification Consistency (P)	0.95	0.87	0.86	0.69
		Probability of Chance	0.84	0.52	0.64	0.30
		Kappa (k)	0.69	0.73	0.62	0.55
		Classification Accuracy	0.96	0.90	0.90	0.77
	African-American	Classification Consistency (P)	0.86	0.92	0.97	0.75
		Probability of Chance	0.51	0.71	0.93	0.36
		Kappa (k)	0.71	0.73	0.57	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	0.62	0.56	0.85	0.31
		Kappa (k)	0.71	0.73	0.57	0.58
		Classification Accuracy	0.92	0.91	0.96	0.79
	Asian	Classification Consistency (P)	0.91	0.88	0.90	0.70
		Probability of Chance	0.70	0.50	0.72	0.28
		Kappa (k)	0.71	0.76	0.66	0.59
		Classification Accuracy	0.94	0.92	0.93	0.79
	American Indian	Classification Consistency (P)	0.88	0.88	0.94	0.70
		Probability of Chance	0.60	0.57	0.85	0.31
		Kappa (k)	0.69	0.72	0.59	0.57
		Classification Accuracy	0.91	0.92	0.96	0.79
Two or More	Classification Consistency (P)	0.92	0.88	0.89	0.70	
	Probability of Chance	0.71	0.50	0.71	0.27	
	Kappa (k)	0.74	0.75	0.63	0.58	
	Classification Accuracy	0.94	0.91	0.92	0.78	
Limited English Proficiency	Yes	Classification Consistency (P)	0.87	0.89	0.96	0.72
		Probability of Chance	0.57	0.63	0.91	0.34
		Kappa (k)	0.70	0.70	0.50	0.58
		Classification Accuracy	0.91	0.92	0.97	0.80
Disability Status	Yes	Classification Consistency (P)	0.88	0.91	0.94	0.74
		Probability of Chance	0.54	0.60	0.85	0.31
		Kappa (k)	0.74	0.77	0.63	0.62
		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	0.62	0.55	0.83	0.30
		Kappa (k)	0.72	0.74	0.59	0.59
		Classification Accuracy	0.92	0.91	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	0.87	0.88	0.68
		Probability of Chance	0.73	0.50	0.72	0.28
		Kappa (k)	0.74	0.73	0.58	0.56
		Classification Accuracy	0.95	0.90	0.92	0.77
	Male	Classification Consistency (P)	0.93	0.88	0.89	0.70
		Probability of Chance	0.68	0.50	0.71	0.27
		Kappa (k)	0.78	0.76	0.61	0.59
		Classification Accuracy	0.95	0.91	0.92	0.78
Race/Ethnicity	White	Classification Consistency (P)	0.95	0.86	0.86	0.68
		Probability of Chance	0.81	0.51	0.67	0.29
		Kappa (k)	0.72	0.72	0.58	0.54
		Classification Accuracy	0.96	0.90	0.90	0.77
	African-American	Classification Consistency (P)	0.87	0.93	0.97	0.77
		Probability of Chance	0.50	0.75	0.94	0.39
		Kappa (k)	0.74	0.71	0.54	0.63
		Classification Accuracy	0.91	0.95	0.98	0.84
	Hispanic	Classification Consistency (P)	0.89	0.88	0.94	0.72
		Probability of Chance	0.58	0.59	0.87	0.31
		Kappa (k)	0.74	0.71	0.56	0.59
		Classification Accuracy	0.93	0.91	0.96	0.80
	Asian	Classification Consistency (P)	0.93	0.87	0.89	0.70
		Probability of Chance	0.73	0.50	0.72	0.28
		Kappa (k)	0.75	0.73	0.61	0.58
		Classification Accuracy	0.95	0.90	0.93	0.78
	American Indian	Classification Consistency (P)	0.88	0.88	0.94	0.70
		Probability of Chance	0.57	0.59	0.87	0.31
		Kappa (k)	0.73	0.70	0.55	0.57
		Classification Accuracy	0.91	0.91	0.96	0.78
Two or More	Classification Consistency (P)	0.92	0.87	0.90	0.70	
	Probability of Chance	0.68	0.51	0.75	0.28	
	Kappa (k)	0.76	0.74	0.61	0.58	
	Classification Accuracy	0.95	0.90	0.93	0.79	
Limited English Proficiency	Yes	Classification Consistency (P)	0.85	0.94	0.99	0.78
		Probability of Chance	0.50	0.84	0.98	0.44
		Kappa (k)	0.70	0.62	0.37	0.61
		Classification Accuracy	0.90	0.96	0.99	0.85
Disability Status	Yes	Classification Consistency (P)	0.87	0.93	0.97	0.78
		Probability of Chance	0.50	0.73	0.92	0.39
		Kappa (k)	0.75	0.75	0.59	0.63
		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	0.57	0.85	0.30
		Kappa (k)	0.76	0.73	0.56	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)	0.92	0.88	0.88	0.68	
		Probability of Chance	0.67	0.50	0.64	0.25	
		Kappa (k)	0.75	0.75	0.67	0.57	
		Classification Accuracy	0.94	0.91	0.91	0.77	
	Male	Classification Consistency (P)	0.91	0.89	0.88	0.69	
		Probability of Chance	0.63	0.50	0.63	0.25	
		Kappa (k)	0.76	0.78	0.68	0.59	
		Classification Accuracy	0.94	0.92	0.92	0.78	
Race/Ethnicity	White	Classification Consistency (P)	0.93	0.88	0.86	0.67	
		Probability of Chance	0.76	0.53	0.58	0.27	
		Kappa (k)	0.71	0.74	0.66	0.55	
		Classification Accuracy	0.95	0.91	0.90	0.76	
	African-American	Classification Consistency (P)	0.88	0.92	0.96	0.76	
		Probability of Chance	0.51	0.68	0.89	0.39	
		Kappa (k)	0.75	0.74	0.64	0.60	
		Classification Accuracy	0.91	0.94	0.97	0.83	
	Hispanic	Classification Consistency (P)	0.88	0.88	0.92	0.69	
		Probability of Chance	0.55	0.54	0.79	0.28	
		Kappa (k)	0.74	0.74	0.64	0.57	
		Classification Accuracy	0.92	0.91	0.94	0.78	
	Asian	Classification Consistency (P)	0.90	0.88	0.91	0.69	
		Probability of Chance	0.62	0.50	0.66	0.25	
		Kappa (k)	0.73	0.76	0.72	0.59	
		Classification Accuracy	0.93	0.91	0.93	0.78	
	American Indian	Classification Consistency (P)	0.88	0.86	0.93	0.68	
		Probability of Chance	0.54	0.55	0.82	0.29	
		Kappa (k)	0.75	0.68	0.62	0.55	
		Classification Accuracy	0.92	0.91	0.95	0.78	
	Two or More	Classification Consistency (P)	0.92	0.88	0.88	0.68	
		Probability of Chance	0.64	0.50	0.65	0.25	
		Kappa (k)	0.77	0.75	0.66	0.57	
		Classification Accuracy	0.94	0.91	0.91	0.76	
	Limited English Proficiency	Yes	Classification Consistency (P)	0.87	0.88	0.94	0.70
			Probability of Chance	0.51	0.59	0.87	0.31
			Kappa (k)	0.73	0.71	0.58	0.56
			Classification Accuracy	0.91	0.92	0.96	0.78
Disability Status	Yes	Classification Consistency (P)	0.89	0.91	0.94	0.74	
		Probability of Chance	0.50	0.59	0.82	0.33	
		Kappa (k)	0.78	0.77	0.67	0.62	
		Classification Accuracy	0.92	0.93	0.96	0.81	
SES Disadvantaged	Yes	Classification Consistency (P)	0.89	0.88	0.92	0.70	
		Probability of Chance	0.54	0.54	0.78	0.28	
		Kappa (k)	0.75	0.75	0.65	0.58	
		Classification Accuracy	0.92	0.91	0.94	0.78	

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)	0.92	0.87	0.88	0.68	
		Probability of Chance	0.67	0.50	0.65	0.26	
		Kappa (k)	0.77	0.73	0.67	0.57	
		Classification Accuracy	0.94	0.91	0.92	0.77	
	Male	Classification Consistency (P)	0.92	0.88	0.89	0.70	
		Probability of Chance	0.62	0.50	0.66	0.25	
		Kappa (k)	0.80	0.76	0.68	0.60	
		Classification Accuracy	0.94	0.92	0.92	0.78	
Race/Ethnicity	White	Classification Consistency (P)	0.94	0.87	0.87	0.68	
		Probability of Chance	0.74	0.52	0.60	0.27	
		Kappa (k)	0.76	0.72	0.66	0.56	
		Classification Accuracy	0.95	0.91	0.90	0.76	
	African-American	Classification Consistency (P)	0.88	0.91	0.97	0.77	
		Probability of Chance	0.51	0.72	0.92	0.41	
		Kappa (k)	0.76	0.70	0.62	0.61	
		Classification Accuracy	0.91	0.94	0.98	0.83	
	Hispanic	Classification Consistency (P)	0.89	0.87	0.93	0.70	
		Probability of Chance	0.54	0.55	0.81	0.28	
		Kappa (k)	0.76	0.71	0.64	0.58	
		Classification Accuracy	0.92	0.91	0.95	0.78	
	Asian	Classification Consistency (P)	0.92	0.88	0.89	0.69	
		Probability of Chance	0.67	0.50	0.66	0.26	
		Kappa (k)	0.76	0.76	0.67	0.59	
		Classification Accuracy	0.94	0.92	0.92	0.78	
	American Indian	Classification Consistency (P)	0.89	0.89	0.93	0.72	
		Probability of Chance	0.52	0.57	0.84	0.30	
		Kappa (k)	0.77	0.75	0.59	0.59	
		Classification Accuracy	0.92	0.92	0.95	0.79	
	Two or More	Classification Consistency (P)	0.92	0.89	0.89	0.70	
		Probability of Chance	0.61	0.50	0.69	0.25	
		Kappa (k)	0.80	0.77	0.65	0.60	
		Classification Accuracy	0.94	0.92	0.92	0.78	
	Limited English Proficiency	Yes	Classification Consistency (P)	0.85	0.92	0.99	0.77
			Probability of Chance	0.52	0.81	0.97	0.45
			Kappa (k)	0.70	0.59	0.54	0.57
			Classification Accuracy	0.89	0.95	0.99	0.83
Disability Status	Yes	Classification Consistency (P)	0.89	0.93	0.97	0.79	
		Probability of Chance	0.52	0.72	0.91	0.44	
		Kappa (k)	0.78	0.73	0.69	0.64	
		Classification Accuracy	0.92	0.95	0.98	0.85	
SES Disadvantaged	Yes	Classification Consistency (P)	0.90	0.88	0.93	0.71	
		Probability of Chance	0.53	0.55	0.81	0.29	
		Kappa (k)	0.78	0.73	0.64	0.59	
		Classification Accuracy	0.92	0.91	0.95	0.79	

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	0.89	0.91	0.72	
		Probability of Chance	0.61	0.50	0.67	0.25	
		Kappa (k)	0.77	0.78	0.73	0.62	
		Classification Accuracy	0.94	0.92	0.93	0.79	
	Male	Classification Consistency (P)	0.91	0.91	0.91	0.74	
		Probability of Chance	0.58	0.50	0.65	0.25	
		Kappa (k)	0.80	0.81	0.76	0.65	
		Classification Accuracy	0.94	0.93	0.94	0.81	
Race/Ethnicity	White	Classification Consistency (P)	0.92	0.89	0.90	0.72	
		Probability of Chance	0.66	0.51	0.62	0.25	
		Kappa (k)	0.76	0.79	0.74	0.62	
		Classification Accuracy	0.94	0.93	0.93	0.79	
	African-American	Classification Consistency (P)	0.90	0.94	0.97	0.81	
		Probability of Chance	0.53	0.72	0.91	0.46	
		Kappa (k)	0.77	0.77	0.71	0.65	
		Classification Accuracy	0.92	0.96	0.98	0.86	
	Hispanic	Classification Consistency (P)	0.89	0.90	0.94	0.74	
		Probability of Chance	0.51	0.57	0.81	0.30	
		Kappa (k)	0.77	0.78	0.70	0.62	
		Classification Accuracy	0.92	0.93	0.96	0.81	
	Asian	Classification Consistency (P)	0.91	0.89	0.92	0.72	
		Probability of Chance	0.60	0.50	0.66	0.25	
		Kappa (k)	0.76	0.78	0.76	0.63	
		Classification Accuracy	0.93	0.92	0.94	0.80	
	American Indian	Classification Consistency (P)	0.88	0.91	0.95	0.74	
		Probability of Chance	0.50	0.60	0.83	0.32	
		Kappa (k)	0.75	0.77	0.71	0.62	
		Classification Accuracy	0.91	0.93	0.97	0.82	
	Two or More	Classification Consistency (P)	0.91	0.91	0.92	0.74	
		Probability of Chance	0.56	0.51	0.68	0.26	
		Kappa (k)	0.79	0.81	0.75	0.65	
		Classification Accuracy	0.93	0.94	0.94	0.81	
	Limited English Proficiency	Yes	Classification Consistency (P)	0.87	0.95	0.99	0.81
			Probability of Chance	0.59	0.86	0.98	0.55
			Kappa (k)	0.68	0.68	0.63	0.59
			Classification Accuracy	0.90	0.97	0.99	0.86
Disability Status	Yes	Classification Consistency (P)	0.89	0.95	0.97	0.82	
		Probability of Chance	0.55	0.74	0.90	0.48	
		Kappa (k)	0.76	0.80	0.75	0.65	
		Classification Accuracy	0.92	0.96	0.98	0.87	
SES Disadvantaged	Yes	Classification Consistency (P)	0.89	0.91	0.95	0.75	
		Probability of Chance	0.51	0.58	0.81	0.31	
		Kappa (k)	0.78	0.79	0.71	0.64	
		Classification Accuracy	0.92	0.94	0.96	0.82	

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.