

WISCONSIN KNOWLEDGE AND CONCEPTS EXAMINATIONS

FALL 2011 WKCE TECHNICAL REPORT



CTB/MCGRAW-HILL
20 RYAN RANCH ROAD
MONTEREY, CALIFORNIA 93940

Copyright

Developed and published under contract with the Wisconsin Department of Public Instruction by CTB/McGraw-Hill LLC, 20 Ryan Ranch Road, Monterey, California 93940-5703. Copyright © 2012 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, 1999).

Acknowledgments

The following CTB/McGraw-Hill (CTB) and Wisconsin Department of Public Instruction (DPI) staff members are primarily responsible for the content and statistical quality of this report:

CTB:

Christina Schneider, Research Scientist

Jill R. van den Heuvel, Research Scientist

Karla Egan, Research Manager

Soe Aung, Manager, Statistical Analyst

Kristina Wilson, Senior Research Associate

Anne Woods, Research Associate

Russ Wegfehrt, Project Manager, Publishing

Patricia Hildreth, Manager, Content Development

Tracy Keith, Scoring Project Manager

Tammy Bullock, Program Manager

Wisconsin DPI—Office of Educational Accountability:

Lynette Russell, Assistant State Superintendent

Laura Pinsonneault, Director

Duane Dorn, Education Consultant

Justin Meyer, Education Consultant

Alison O'Hara, Education Consultant

Viji Somasundaram, Education Consultant

TABLE OF CONTENTS

Copyright	ii
Foreword	iii
Acknowledgments.....	iv
Part 1: Overview.....	1
Part 2: Test Design and Item Development.....	6
2.1 CONTENT FRAMEWORK AND ASSESSMENT LIMITS.....	7
2.2 TEST BLUEPRINT	7
2.3 READING PASSAGE SELECTION.....	8
2.4 ITEM DEVELOPMENT AND EDITING.....	8
2.5 CONTENT/BIAS REVIEW AND ITEM ALIGNMENT.....	8
Part 3: Test Form Development.....	9
3.1 OVERALL TEST BOOK DEVELOPMENT PROCESS.....	9
3.1.1 WKCE FALL 2011 FORM SELECTION	9
3.1.2 WKCE FIELD TEST ITEM SELECTION	11
3.1.3 QUALITY REVIEWS	11
3.2 DESCRIPTION OF THE WKCE 2011 TESTS	12
3.2.1 READING.....	12
3.2.2 MATHEMATICS	12
3.2.3 LANGUAGE ARTS.....	13
3.2.4 SOCIAL STUDIES	13
3.2.5 SCIENCE.....	13
3.3 CUSTOMER APPROVALS.....	13
3.3.1 ITEM CONTENT AND BIAS REVIEW	14
3.3.2 ITEM SELECTION APPROVAL.....	14
3.3.3 FIRST PAGES APPROVALS	14
3.3.4 SECOND PAGES APPROVALS.....	14
3.3.5 FINAL PAGES SIGN-OFF.....	14
Part 4: Test Administration.....	15
4.1 ACCOMMODATIONS	15
4.2 REPORTING RESULTS OF ASSESSMENTS TAKEN WITH ACCOMMODATIONS	16
4.3 TEST SECURITY.....	16
4.4 TEST ADMINISTRATION	18
Part 5: Scoring.....	20
5.1 SCORING OF MULTIPLE-CHOICE ITEMS.....	20
5.1.1 SCANNING AND CALIBRATION OF SCANNERS	21
5.2 SCORING OF OPEN-ENDED OR CONSTRUCTED-RESPONSE ITEMS	23
5.2.1 DESCRIPTION OF SCORING RUBRICS	23
5.2.2 HANDSCORING PROCESS.....	23
5.2.3 ELECTRONIC HANDSCORING SYSTEM.....	24
5.2.4 ANCHOR PAPERS AND TRAINING PAPERS	24
5.2.5 SCORING PERSONNEL AND QUALIFICATIONS.....	24
5.2.6 READER TRAINING	25
5.2.7 INTER-RATER RELIABILITY	25
5.3 DISTRIBUTION OF CONSTRUCTED-RESPONSE ITEM SCORES	26

Part 6: Characteristics of the Calibration Sample.....	28
Part 7: Calibration, Equating, and Deriving Scale Scores.....	29
7.1 CALIBRATION AND EQUATING METHODS	29
7.1.1 CALIBRATION MODELS	30
7.1.2 CALIBRATION SOFTWARE	31
7.2 CALIBRATION RESULTS	32
7.2.1 IRT ITEM PARAMETERS	32
7.2.2 IRT ITEM FIT	32
7.2.3 EVALUATING ANCHOR ITEMS	34
7.3 DERIVING SCALE SCORES IN THE WKCE	35
7.3.1 STANDARD ERROR OF MEASUREMENT	39
7.3.2 LOSS AND HOSS	40
7.3.3 TEST CHARACTERISTIC CURVES	41
Part 8: Test Results	42
8.1 CLASSICAL ITEM ANALYSIS: ITEM LEVEL STATISTICS.....	42
8.1.1 FLAGGING FOR A POSITIVE DISTRACTOR CORRELATION	47
8.1.2 FLAGGING FOR THE ITEM-TOTAL CORRELATION	47
8.1.3 FLAGGING FOR P-VALUE.....	47
8.1.4 FLAGGING FOR OMIT RATE.....	48
8.1.5 SUPPLEMENTAL TABLES ON CLASSICAL ITEM ANALYSIS	48
8.1.6 SPEEDEDNESS	48
8.2 RAW SCORE RESULTS	49
Reading	50
Mathematics	51
Language Arts	51
Social Studies.....	51
Science	51
Subgroup Performance Patterns in Raw Score Results	52
8.3 SUMMARY STATISTICS FOR SCALE SCORES	53
Reading	53
Mathematics	53
Language Arts	54
Social Studies.....	54
Science	54
Subgroup Performance Patterns in Scale Score Results.....	54
Gender.....	54
Race/Ethnicity.....	55
Socioeconomic Status	55
Disability Status	55
English Language Proficiency	56
8.4 CUT SCORES AND PERFORMANCE LEVEL CLASSIFICATIONS.....	56
Reading	57
Mathematics.....	57
Language Arts.....	57
Social Studies.....	58
Science	58
Subgroup Patterns in Performance Level Results	58
8.5 STANDARD PERFORMANCE INDICATORS FOR CONTENT STANDARDS	59
Reading	60
Mathematics	61
Language Arts.....	61
Social Studies.....	61
Science	61

Summary of Student Performance Indicator Results	62
Summary of Student Achievement Results.....	62
Part 9: Reliability	63
9.1 MEASURES OF INTERNAL CONSISTENCY AND SEM	64
9.2 CLASSIFICATION CONSISTENCY AND ACCURACY	67
9.2.1 KOLEN AND KIM’S METHOD FOR PATTERN SCORING	68
9.3 INTER-RATER RELIABILITY FOR CR ITEMS AND WRITING PROMPTS	71
Reading	74
Mathematics	74
Writing	75
Summary	75
Part 10: Validity	76
10.1 DIFFERENTIAL ITEM FUNCTIONING	78
Results	83
10.2 CONSTRUCT VALIDITY	85
10.3 TEST INTEGRITY: ERASURE ANALYSIS	86
10.4 STANDARDIZED TEST ADMINISTRATION.....	86
Part 11: Summary Recommendations	88
Fall 2011 WKCE Technical Report: References.....	89
Fall 2011 Tables and Figures	93

Appendices

Appendix 1: Fall 2010 Item Selection Check-Off Form	320
Appendix 2: Fall 2011 WKCE Assessment Accommodations Matrix	332
Appendix 3: WKCE Glossary	337

TABLE OF TABLES

PART 2

Table 2-1 Target Reading Test Blueprint: Grades 3–8, 10*	94
Table 2-2 Target Mathematics Test Blueprint: Grades 3–8, 10*	95
Table 2-3 Target Language Arts Test Blueprint: Grades 4, 8, 10	96
Table 2-4 Target Social Studies Test Blueprint: Grades 4, 8, 10	96
Table 2-5 Target Science Test Blueprint: Grades 4, 8, 10*	96
Table 2-6 Actual Reading Test Blueprint: Grades 3–8, 10*	97
Table 2-7 Actual Mathematics Test Blueprint: Grades 3–8, 10*	99
Table 2-8 Actual Language Arts Test Blueprint: Grades 4, 8, 10	100
Table 2-9 Actual Science Test Blueprint: Grades 4, 8, 10*	100
Table 2-10 Actual Social Studies Test Blueprint: Grades 4, 8, 10	100
Table 2-11 Item Development Each Year and Total to Date*	101

PART 3

Table 3-1 Fall 2011 Test Configuration	102
Table 3-2 Unique Items Field Tested Each Year and Total to Date	103

PART 4

Table 4-1 Test Accommodations	104
-------------------------------	-----

PART 5

Table 5-1 Reading Rubric, Grades 3–8 and 10	115
Table 5-2 Mathematics Rubric, Grades 3–8 and 10	116
Table 5-3 Writing Rubric, Conventions of Written English, Grade 4	117
Table 5-4 Writing Rubric, Composing, Grade 4	118
Table 5-5 Writing Rubric, Conventions of Written English, Grade 8	119
Table 5-6 Writing Rubric, Composing, Grade 8	120
Table 5-7 Writing Rubric, Conventions of Written English, Grade 10	122
Table 5-8 Writing Rubric, Composing, Grade 10	123
Table 5-9 Score Distribution for Reading CR Items	125
Table 5-10 Score Distribution for Mathematics CR Items	126
Table 5-11 Score Distribution for Grades 4, 8, and 10 Writing: Composing Rubric	128
Table 5-12 Percentage Distribution of Scores, Grades 4, 8, and 10 Writing: Composing Rubric	128
Table 5-13 Score Distribution, Grades 4, 8, and 10 Writing: Conventions Rubric	129
Table 5-14 Percentage Distribution of Scores for Grades 4, 8, and 10 Writing: Conventions Rubric	129
Table 5-15 Score Distribution for Grades 4, 8, and 10 Writing: Total Score	130
Table 5-16 Percentage Distribution of Scores for Grades 4, 8, and 10 Writing: Total Score	130

PART 7

Table 7-1 Item Flagged Based on Yen’s Q1	131
Table 7-2 Scoring Table for Reading Grade 3	132
Table 7-3 Scoring Table for Reading Grade 4	133
Table 7-4 Scoring Table for Reading Grade 5	134
Table 7-5 Scoring Table for Reading Grade 6	135
Table 7-6 Scoring Table for Reading Grade 7	136
Table 7-7 Scoring Table for Reading Grade 8	137
Table 7-8 Scoring Table for Reading Grade 10	138
Table 7-9 Scoring Table for Mathematics Grade 3	139
Table 7-10 Scoring Table for Mathematics Grade 4	140

TABLE OF TABLES CONTINUED

Table 7-11 Scoring Table for Mathematics Grade 5.....	141
Table 7-12 Scoring Table for Mathematics Grade 6.....	142
Table 7-13 Scoring Table for Mathematics Grade 7.....	143
Table 7-14 Scoring Table for Mathematics Grade 8.....	144
Table 7-15 Scoring Table for Mathematics Grade 10.....	145
Table 7-16 Scoring Table for Language Arts Grade 4.....	146
Table 7-17 Scoring Table for Language Arts Grade 8.....	147
Table 7-18 Scoring Table for Language Arts Grade 10.....	148
Table 7-19 Scoring Table for Social Studies Grade 4.....	149
Table 7-20 Scoring Table for Social Studies Grade 8.....	150
Table 7-21 Scoring Table for Social Studies Grade 10.....	151
Table 7-22 Scoring Table for Science Grade 4.....	152
Table 7-23 Scoring Table for Science Grade 8.....	153
Table 7-24 Scoring Table for Science Grade 10.....	154
Table 7-25 The Number of Students and Percents at LOSS and HOSS.....	155

PART 8

Table 8-1 Item Analysis Grade 3 Reading.....	156
Table 8-2 Item Analysis Grade 4 Reading.....	159
Table 8-3 Item Analysis Grade 5 Reading.....	162
Table 8-4 Item Analysis Grade 6 Reading.....	165
Table 8-5 Item Analysis Grade 7 Reading.....	168
Table 8-6 Item Analysis Grade 8 Reading.....	171
Table 8-7 Item Analysis Grade 10 Reading.....	174
Table 8-8 Item Analysis Grade 3 Mathematics.....	177
Table 8-9 Item Analysis Grade 4 Mathematics.....	180
Table 8-10 Item Analysis Grade 5 Mathematics.....	183
Table 8-11 Item Analysis Grade 6 Mathematics.....	186
Table 8-12 Item Analysis Grade 7 Mathematics.....	189
Table 8-13 Item Analysis Grade 8 Mathematics.....	192
Table 8-14 Item Analysis Grade 10 Mathematics.....	195
Table 8-15 Item Analysis Grade 4 Language Arts.....	198
Table 8-16 Item Analysis Grade 8 Language Arts.....	200
Table 8-17 Item Analysis Grade 10 Language Arts.....	202
Table 8-18 Item Analysis Grade 4 Social Studies.....	204
Table 8-19 Item Analysis Grade 8 Social Studies.....	206
Table 8-20 Item Analysis Grade 10 Social Studies.....	208
Table 8-21 Item Analysis Grade 4 Science.....	210
Table 8-22 Item Analysis Grade 8 Science.....	212
Table 8-23 Item Analysis Grade 10 Science.....	214
Table 8-24 The Number of Items Flagged.....	216
Table 8-25 Raw Score Descriptive Statistics based on Census Data.....	217
Table 8-26 Raw Score Descriptive Statistics by Gender.....	218
Table 8-27 Raw Score Descriptive Statistics for Reading by Race/Ethnicity.....	219
Table 8-28 Raw Score Descriptive Statistics for Mathematics by Race/Ethnicity.....	220
Table 8-29 Raw Score Descriptive Statistics for Language Arts by Race/Ethnicity.....	221
Table 8-30 Raw Score Descriptive Statistics for Social Studies by Race/Ethnicity.....	221
Table 8-31 Raw Score Descriptive Statistics for Science by Race/Ethnicity.....	222
Table 8-32 Raw Score Descriptive Statistics by Socioeconomic Status.....	223
Table 8-33 Raw Score Descriptive Statistics by Disability.....	224
Table 8-34 Raw Score Descriptive Statistics by English Language Proficiency.....	225
Table 8-35 2011 and 2009 Scale Score Mean and Standard Deviation.....	226

TABLE OF TABLES CONTINUED

Table 8-36 Scale Score Descriptive Statistics	227
Table 8-37 Scale Score Descriptive Statistics by Gender	228
Table 8-38 Scale Score Descriptive Statistics for Reading by Race/Ethnicity	229
Table 8-39 Scale Score Descriptive Statistics for Mathematics by Race/Ethnicity	230
Table 8-40 Scale Score Descriptive Statistics for Language Arts by Race/Ethnicity	231
Table 8-41 Scale Score Descriptive Statistics for Social Studies by Race/Ethnicity	232
Table 8-42 Scale Score Descriptive Statistics for Science by Race/Ethnicity	232
Table 8-43 Scale Score Descriptive Statistics by Socioeconomic Status	233
Table 8-44 Scale Score Descriptive Statistics by Disability	234
Table 8-45 Scale Score Descriptive Statistics by English Language Proficiency	235
Table 8-46 Performance Level Cut Scores for all Contents*	236
Table 8-47 Percentage of Students in Each Performance Level by Sub-Group (Reading).....	237
Table 8-48 Percentage of Students in Each Performance Level by Sub-Group (Mathematics)	240
Table 8-49 Percentage of Students in Each Performance Level by Sub-Group (Language Arts)	243
Table 8-50 Percentage of Students in Each Performance Level by Sub-Group (Social Studies)	244
Table 8-51 Percentage of Students in Each Performance Level by Sub-Group (Science)	245
Table 8-52 Cut Scores and Associated Impact Data for WKCE-CRT Reading	246
Table 8-53 Cut Scores and Associated Impact Data for WKCE-CRT Mathematics	246
Table 8-54 Cut Scores and Associated Impact Data for WKCE-CRT Language Arts	247
Table 8-55 Cut Scores and Associated Impact Data for WKCE-CRT Social Studies.....	247
Table 8-56 Cut Scores and Associated Impact Data for WKCE-CRT Science	248
Table 8-57 Summary Statistics for Reading Content Standards Raw and SPI Scores.....	249
Table 8-58 Summary Statistics for Mathematics Content Standards Raw and SPI Scores	251
Table 8-59 Summary Statistics for Language Arts Content Standards Raw and SPI Scores	253
Table 8-60 Summary Statistics for Social Studies Content Standards Raw and SPI Scores	254
Table 8-61 Summary Statistics for Science Content Standards Raw and SPI Scores	255
Table 8-62 SPI Cut Scores.....	256

PART 9

Table 9-1 Reliability for Total Group and Subgroups Using Cronbach’s Alpha.....	259
Table 9-2 Standard Error of Measurement for Total Group and Subgroups	260
Table 9-3 Cronbach’s Alpha Reliability Coefficients for Content Standards.....	261
Table 9-4 Standard Error of Measurement per Content Standard.....	262
Table 9-5 Classification Consistency and Classification Accuracy for Reading Grade 3	263
Table 9-6 Classification Consistency and Classification Accuracy for Reading Grade 4	264
Table 9-7 Classification Consistency and Classification Accuracy for Reading Grade 5	265
Table 9-8 Classification Consistency and Classification Accuracy for Reading Grade 6	266
Table 9-9 Classification Consistency and Classification Accuracy for Reading Grade 7	267
Table 9-10 Classification Consistency and Classification Accuracy for Reading Grade 8	268
Table 9-11 Classification Consistency and Classification Accuracy for Reading Grade 10	269
Table 9-12 Classification Consistency and Classification Accuracy for Mathematics Grade 3	270
Table 9-13 Classification Consistency and Classification Accuracy for Mathematics Grade 4	271
Table 9-14 Classification Consistency and Classification Accuracy for Mathematics Grade 5	272
Table 9-15 Classification Consistency and Classification Accuracy for Mathematics Grade 6	273
Table 9-16 Classification Consistency and Classification Accuracy for Mathematics Grade 7	274
Table 9-17 Classification Consistency and Classification Accuracy for Mathematics Grade 8	275
Table 9-18 Classification Consistency and Classification Accuracy for Mathematics Grade 10	276
Table 9-19 Classification Consistency and Classification Accuracy for Language Arts Grade 4	277
Table 9-20 Classification Consistency and Classification Accuracy for Language Arts Grade 8	278
Table 9-21 Classification Consistency and Classification Accuracy for Language Arts Grade 10	279
Table 9-22 Classification Consistency and Classification Accuracy for Social Studies Grade 4	280
Table 9-23 Classification Consistency and Classification Accuracy for Social Studies Grade 8	281

TABLE OF TABLES CONTINUED

Table 9-24 Classification Consistency and Classification Accuracy for Social Studies Grade 10.....	282
Table 9-25 Classification Consistency and Classification Accuracy for Science Grade 4.....	283
Table 9-26 Classification Consistency and Classification Accuracy for Science Grade 8.....	284
Table 9-27 Classification Consistency and Classification Accuracy for Science Grade 10.....	285
Table 9-28 Inter-Rater Reliability, Reading*.....	286
Table 9-29 Inter-Rater Reliability, Mathematics*.....	287
Table 9-30 Inter-Rater Reliability, Writing Prompts*.....	289

PART 10

Table 10-1 Items Flagged for DIF, By Gender.....	290
Table 10-2 Items Flagged for DIF, By Race/Ethnicity, African American.....	291
Table 10-3 Items Flagged for DIF, By Race/Ethnicity, Hispanic.....	292
Table 10-4 Items Flagged for DIF, By Race/Ethnicity, Asian.....	293
Table 10-5 Items Flagged for DIF, By Race/Ethnicity, American Indian*.....	295
Table 10-6 Items Flagged for DIF, By English Language Proficiency.....	296
Table 10-7 Items Flagged for DIF, By Disability Status.....	298
Table 10-8 Correlations among Reading Objectives.....	299
Table 10-9 Correlations among Mathematics Objectives.....	300
Table 10-10 Correlations among Language Arts Objectives.....	301
Table 10-11 Correlations among Social Studies Objectives.....	301
Table 10-12 Correlations among Science Objectives.....	302
Table 10-13 Principal Components Analysis.....	303

TABLE OF FIGURES

PART 7

Figure 7-1 SEM Curves, Reading Grades 3-6	304
Figure 7-2 SEM Curves, Mathematics Grades 3-6	306
Figure 7-3 SEM Curves, Language Arts Grades 4, 8, 10	308
Figure 7-4 SEM Curves, Social Studies Grades 4, 8, 10	309
Figure 7-5 SEM Curves, Science Grades 4, 8, 10	310
Figure 7-6 TCC Curve for Reading Grades 3-8, 10.....	311
Figure 7-7 TCC Curve for Mathematics Grades 3-8, 10	312
Figure 7-8 TCC Curve for Language Arts Grades 4, 8, 10.....	313
Figure 7-9 TCC Curve for Social Studies Grades 4, 8, 10.....	314
Figure 7-10 TCC Curve for Science Grades 4, 8, 10.....	315

PART 9

Figure 9-1 Reading Indices for Classification Consistency and Classification Accuracy	316
Figure 9-2 Mathematics Indices for Classification Consistency and Classification Accuracy	316
Figure 9-3 Language Arts Indices for Classification Consistency and Classification Accuracy	317
Figure 9-4 Social Studies Indices for Classification Consistency and Classification Accuracy	317
Figure 9-5 Science Indices for Classification Consistency and Classification Accuracy	318

Part 1: Overview

The *Fall 2011 Wisconsin Knowledge and Concepts Examinations (WKCE) Technical Report* documents the processes and procedures applied in the Fall 2011 WKCE and the results. This report also documents processes, procedures, and results of this administration to support validity and reliability evidence for 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], 1999). This report demonstrates that the Fall 2011 WKCE adhered to the appropriate standards and practices of educational assessment. Ultimately, this report serves to document evidence that valid inferences about Wisconsin student performance can be derived from this assessment.

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* spell 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 norm-referenced assessments in the 1997 school year. The assessments used at that time were *TerraNova*TM tests developed by CTB. 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 were adopted. These new standards were the work of the Governor's Commission on Wisconsin Model Academic Standards, chaired by then-current Lieutenant Governor McCallum and the Wisconsin DPI. The Wisconsin Model Academic Standards 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, has been used to determine the Adequate Yearly Progress (AYP) of students at the school, district, and state levels.

Beginning with school year 2007–08, states were also required to administer Science assessments at least once during grades 3–5, grades 6–9, and grades 10–12. Wisconsin students in grades 4, 8, and 10 are, and will continue to be, assessed in Language Arts, Science, and Social Studies as required by state law (§ 118.30 Wisconsin Statutes).

It is 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 Reading and Mathematics tests. The criterion-referenced WKCE is designed specifically for Wisconsin students, and specifically to measure their performance on the Wisconsin Model Academic Standards adopted by the state. These assessments are designed to evaluate students'

knowledge and to measure achievement in the basic skills taught in schools at grades 3–8 and 10. The Fall 2011 WKCE is the seventh administration of these assessments.

The Wisconsin Model Academic Standards describe what students should know and be able to do in grades 4, 8, and 12. To determine what should be tested in grades 3, 5–7, and 10, committees of Wisconsin educators carefully considered what knowledge and skills students should have by the fall of each school year by extrapolating and interpolating the standards for grades 4, 8, and 12. The committees then defined the eligible test content and assessment limits, ensuring that the test framework they designed incorporated the content and performance standards enumerated in the Wisconsin Model Academic Standards. Therefore, the assessment framework, used to define what is tested on the WKCE, reflects what students should have learned by the beginning of the school year in order to be successful in that grade. As a result, the grade 6 test, for example, assesses what students should have learned by the end of grade 5.

The WKCE tests consist of criterion-referenced items written by CTB and edited and reviewed by Wisconsin teachers and items from CTB’s norm-referenced test, *TerraNova*, The Second Edition (*TerraNova*, CTB/McGraw-Hill, 2001). The Fall 2011 WKCE tests include Reading and Mathematics at grades 3–8 and 10 and Science, Social Studies, and Language Arts (including Writing) at grades 4, 8, and 10.

Based on the input of Wisconsin educators and the Wisconsin Model Academic Standards, a design was derived for the development, administration, and scoring of the WKCE. The present Technical Report documents all aspects of the testing cycle in the subsequent chapters. The structure of the present Technical Report mirrors the testing cycle. A brief content summary of the report is provided below.

Test Design and Item Development

- Part 2 of this report describes test design, the item development process, and some aspects of the content-related validity of the WKCE tests.
- More specifically, Part 2 describes how CTB, DPI, and Wisconsin educators collaborated through a series of test development processes to ensure that the appropriate content was included in the WKCE and to ensure that the test items adequately sampled the domain of content knowledge necessary to make legitimate inferences about student performance.
- Wisconsin Model Academic Standards were translated into grade-level content frameworks, which in turn formed the basis for test blueprints and item specifications.
- Wisconsin educators were involved in design at every step to ensure the appropriateness of the test to the standards.
- Test design started in August 2003 with the convention of approximately 35 educators per content area for grades 3–8 and 10 to establish the grade-level content frameworks based on the Wisconsin Model Academic Standards, establish assessment limits, create the test blueprint, and review reading passage and page specifications. The test specifications documents created and later approved by DPI continue to serve as a foundation for item and test development.

Test Form Development

- Part 3 discusses key development tasks and issues related to creating the Fall 2011 WKCE test forms.
- Item development was based on the approved test blueprints, with a sufficient quantity of items written across years to develop multiple operational test forms.
- Part 3 discusses the process of selecting operational test items and the process of obtaining DPI approvals.
- As detailed in Part 3, there have been 5,025 unique multiple-choice (MC) items and 499 unique constructed-response (CR) items field tested to date, that is, through the Fall of 2009, totaling 5,524 unique items.
- Selection of the Fall 2011 operational forms was done using the ITEMWIN (Burket, 2000) software using methods similar to previous administrations for all grades and content areas.

Test Administration

- Part 4 briefly describes test administration and accommodations.
- The test administration window was October 24–November 25, 2011.
- Delivery of materials was handled through the district and school assessment coordinators.
- In 2011, all content area tests in a grade were administered to students using a single test book.

Scoring

- Part 5 documents the scanning and scoring process for the MC and CR items.
- The machine-scanning process and the handscoring process, including the development and review of the scoring rubrics, anchor (sample) papers, and writing prompts, as well as the training of scoring personnel, ongoing quality assurance, the application of an inter-rater reliability assessment, and a systematic review of the resulting score distributions supported reporting of reliable and valid test scores.
- The scoring rubrics used in handscoring are presented in detail for all content areas with handscored items.

Characteristics of the Calibration Sample

- The calibration and equating of the Fall 2011 WKCE tests occurred during the 2009 WKCE administration. The calibration was based on a sample of student response data termed the calibration sample. The calibration sample roughly approximates the census population for minority students and under-represents majority students, as has been the historical practice.

Calibration, Equating, and Deriving Scale Scores

- Part 7 reviews calibration, equating, and scoring methods implemented for WKCE.
- The Fall 2011 WKCE was calibrated and scaled using two different item response theory (IRT) models, one for CR items and one for MC items, which are the item types used for most large-scale standardized testing programs in education. Evaluation of the sufficiency of the IRT model results include model-to-data fit and the standard error of measurement (SEM).
- Item-pattern scoring was applied to the Fall 2011 WKCE. As discussed in Part 7, item-pattern scoring is generally recommended over number-correct scoring because it produces more accurate scores for individual students.
- Part 7 also explains how a student's scale score is derived from the raw score using item-pattern scoring. Examples of a very low-performing student, a very high-performing student, and several students with a 50% correct raw score are provided. Several students with the same 50% correct raw score are provided in order to illustrate how students with the same raw score can have different scale scores.

Test Results

- Part 8 summarizes item analyses, raw scores, scale scores, performance levels, and a standard performance indicator score for content standards.
- Reliability of the WKCE tests are reported using Cronbach's alpha and SEM.
- Summary descriptive statistics for all scores (raw scores, scale scores, standard performance indicator scores, and performance levels) are reported for all students and for subgroups identified by gender, race/ethnicity, socioeconomic status, disability status, and English language proficiency.

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 estimated for all CR items.
- Internal consistency was assessed for all MC and CR items using Cronbach's alpha.
- Classification consistency and accuracy were estimated for performance classification.

Validity

- Part 10 reviews the validity evidence presented in all prior parts and provides additional validity evidence supporting the WKCE tests.
- Factor analysis and correlations among content standards are presented in the context of construct validity.
- An analysis of differential item functioning (DIF) is presented.
- Erasure analysis, a procedure used to identify high erasure rates, is also discussed.

Summary Recommendations

- Key findings of the Fall 2011 WKCE administration are presented in the body of the report. However, some items of a more technical nature, which stand out as key recommendations and summary statements that should be considered in subsequent administrations, are presented in Part 11.
- Recommendations based on the Fall 2011 WKCE administration cover three different phases of the testing cycle: item development, scoring, and psychometric, or measurement-based, research and evaluation.

Part 2: Test Design and Item Development

The purpose of this section is to describe how CTB, DPI, and Wisconsin educators collaborated through a series of test development processes to ensure that appropriate content was included in the WKCE 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 development process for the Fall 2011 WKCE administration.

As described below, the Wisconsin Model Academic Standards were central to the entire test design process. Part 2 of the Technical Report demonstrates the adherence of the WKCE program to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) and specifically to standards 1.2, 1.6, 3.1, 3.2, 3.3, 3.11, 6.4, 6.15, 13.3, and 13.5.

Operational items administered in 2011 adhered to the test specifications documents developed in previous years. The Fall 2006 Technical Report (Parts 2, 3, and 4) provides a detailed account of the development of the test specifications documents during previous years. Interested readers can find these sections of the Fall 2006 Technical Report in Appendix 2 the 2010 WKCE Technical Report available online at: <http://dpi.state.wi.us/oea/pdf/td-2010-techman.pdf>. The assessment frameworks, test design, test blueprints, reading passage specifications, item specifications, art specifications, and style guide were all developed in 2003, the first year of the WKCE program. The role of Wisconsin educators was an essential component of the development of the WKCE. Their professional expertise and judgment were central to approving content that was appropriately rigorous for the grade and content area in which it was presented and that was expected to have been taught to students.

During the first year of the contract, August 2003 to August 2004, the test specifications documents were developed through an extended, collaborative process with DPI and based on the contributions of Wisconsin educators during meetings conducted in 2003 (see the Fall 2006 Technical Report, p. 6, which is provided in Appendix 2 of the 2010 WKCE Technical Report available online at: <http://dpi.state.wi.us/oea/pdf/td-2010-techman.pdf>). Test specification documents include the test blueprints, passage specifications, item specifications, page specifications, and style guide.

According to the most recent edition of the standards (AERA, APA, NCME, 1999), “Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of test scores” (p. 9). Much of the content-related validity evidence is produced during the test development process. The content-related evidence supports inferences from a sample of observations (the test) to a domain of observations (the content area). A substantial source of content-related validity evidence is the expert judgment that the test items are an adequate and representative sample of the domain being measured. Content-related validity evidence can support interpretations of test scores in terms of performance over a performance domain. If the content domain is specified clearly and a representative sample of performance tasks is drawn from the domain, then inferences about expected performance over the domain based on observed performances should be legitimate.

2.1 Content Framework and Assessment Limits

The Assessment Framework documents created by DPI provide information about the content measured at each grade level and explain the relationships among the Wisconsin Model Academic Standards, the Assessment Framework, and classroom instruction. The Framework documents are located on DPI's website at <http://dpi.wi.gov/oea/wkce.html>. The Fall 2006 Technical Report, Section 3.1.1, explains the structure and development of the Assessment Frameworks (see Appendix 2 of the 2010 WKCE Technical Report available online at: <http://dpi.state.wi.us/oea/pdf/td-2010-techman.pdf>).

The Assessment Frameworks specify the broad categories within the content area at which test sub-scores may be reported, for example, “Number Operations and Relations” or “Measurement” for Mathematics and “Understands Text” or “Analyzes Text” for Reading. These broad categories are further delineated into subskills. For example, “Number Operations and Relations” is further subdivided into “Reading, Writing, and Representing Numbers” and “Ordering and Comparing Numbers” and so forth. Assessment limits are bulleted statements that identify the specific content that is eligible for testing for each subskill and may clarify how the content could be assessed. For example, in Mathematics, the size of numbers or the types of plane and solid geometric figures that are appropriate at each grade level would be specified in the assessment limits. For Reading, the assessment limits clarify which prefixes or suffixes or which literary devices are appropriate to assess at each grade level. For the grade 4, 8, and 10 Science assessments, the Wisconsin Model Academic Standards served as the foundation for the creation of the Science Assessment Frameworks. Similarly, the Wisconsin Model Academic Standards for Language Arts and Social Studies provide the content framework for these content area tests at grades 4, 8, and 10.

2.2 Test Blueprint

The test blueprints specify the number of multiple-choice (MC) and constructed-response (CR) items for each reporting category and subskill. The process used for developing the blueprints was described in detail in Parts 2 and 3 of the Fall 2006 Technical Report (see Appendix 2 of the 2010 WKCE Technical Report available online at: <http://dpi.state.wi.us/oea/pdf/td-2010-techman.pdf>). Tables 2-1 through 2-5 present the target blueprints for the Fall 2011 WKCE. Tables 2-6 through 2-10 present the actual test blueprints showing how the items selected for the Fall 2011 WKCE forms were distributed by reporting category and subskill for each item type.

In 2007, some changes were made to the blueprints for Mathematics, Science, and Language Arts grade 8. The Mathematics blueprints were modified to reflect the inclusion of a 2-point CR item and the subsequent reduction of the 3-point CR items from four to three. In addition, the number of MC items for each reporting category was adjusted to reflect the use of MC items for reporting category A. The Science blueprints were modified slightly to show a shift in emphasis among reporting categories A and B and among reporting categories G and H. The Language Arts grade 8 blueprint changes involved shifting two MC items from reporting category D to reporting category B. This change was made in response to Wisconsin educators’

concerns expressed at the 2005 content review that the language test should not require excessive reading. When selecting test forms for 2005 and 2006, CTB made an effort to minimize the number of item sets that use a common stimulus, such as a brief essay or letter. However, when selecting the 2008 forms, the use of two lengthy stimuli would have been necessary in order to meet the blueprint. CTB brought this concern to the attention of DPI and suggested that two items be shifted from category D to category B. DPI approved this change to the blueprint on March 9, 2007.

In addition to the changes above, Depth of Knowledge (DOK) requirements were incorporated into the Reading and Mathematics blueprints to indicate the number of items needed at each DOK level for each reporting category.

2.3 Reading Passage Selection

Reading passages on the 2011 operational¹ forms were selected, reviewed, and approved between 2001 and 2008. The processes used for selecting, reviewing, and approving WKCE Reading passages were detailed in Section 3.1.3 of the Fall 2006 Technical Report (see Appendix 2 of the 2010 WKCE Technical Report available online at: <http://dpi.state.wi.us/oea/pdf/td-2010-techman.pdf>).

2.4 Item Development and Editing

While historically new items have been developed each year for the WKCE, in 2011 new items were not developed. Table 2-11 shows the number of MC, CR, and total items that have been written up to 2011 for the WKCE.

2.5 Content/Bias Review and Item Alignment

Because there were no field test items on the 2011 test forms, content and bias reviews for field test items did not occur.

¹ Operational items are those items that contribute to student scores. Operational items are abbreviated in this report as OP.

Part 3: Test Form Development

Part 3 of the Technical Report focuses on key development tasks and issues related to creating the Fall 2011 WKCE operational test forms. The test specifications 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 CTB test development specialists. The foundation test specifications documents—assessment frameworks, assessment limits, passage specifications, item specifications, test blueprints, art and page specifications, and style guide—developed and approved during the initial phases of the project served as critical guides throughout development and field testing of items. These documents contributed to ensuring that each form of the test accurately measured the content in consistent and stable ways, thus providing evidence supporting the test’s use as an indicator of student achievement of state standards. Information is provided in Part 3 relating to the following topics:

- a general discussion of CTB’s test book creation and editing process
- the process of selecting operational test items
- the process of obtaining DPI approvals

A comprehensive, multi-segment development process guides the development of assessment materials. The following section outlines this process in general terms. The remainder of Part 3 provides details of how these processes were implemented in Wisconsin. This section of the Technical Report addresses the following AERA/APA/NCME (1999) standards: 1.6, 3.1, 3.5, 3.6, 3.7, 3.9, 3.11, 3.16, 6.4, 6.15, 7.3, 7.4, 7.7, 13.3, and 13.5.

3.1 Overall Test Book Development Process

The creation of test book materials involved the expertise of multiple CTB departments, DPI, and Wisconsin educators. The activities that contributed to the creation of the test book materials are described below.

3.1.1 WKCE Fall 2011 Form Selection

The WKCE operational test forms for all content areas and grade levels use the common item non-equivalent groups design in order to equate parallel test forms from year to year. The minimum number of common items (also called anchor items) per content area follows:

- Reading: 16 items
- Mathematics: 18 items
- Language Arts: 15 items
- Social Studies: 15 items
- Science: 15 items

CTB assessment editors selected items for the 2011 operational forms while considering a variety of criteria, including the following:

- Selected items must fully cover the reporting areas of the test blueprint.
- Selected items must represent the diversity of content.
- MC items with p -values below 0.30 should be avoided when possible.
- CR items with p -values below 0.20 should be avoided when possible.
- Items with positive point-biserials on distractors should be avoided when possible.
- Items should represent a range of scale score values.
- Items with differential item functioning (DIF) flags (C flags) should be avoided when possible.
- Items with poor fit flags should be avoided when possible.

CTB content editors used CTB's proprietary software, called ITEMWIN (Burket, 2000), to select items for the Fall 2009 WKCE operational test forms for all content areas and grade levels. These 2009 test forms were re-administered in 2011 with no modifications made to the operational selection. ITEMWIN has two phases. In the first phase, CTB uses ITEMWIN to select a working item pool of manageable size from the larger tryout pool; items clearly inappropriate to the target grade range are eliminated. There is information about each item in the pool, including the item format to which the item is assigned, a descriptive phrase about the item, the association of the item with a stimulus, the item parameters, a fit rating indicating how well the item fits the expectations based on the IRT model used, and a DIF rating indicating whether the probability of answering the item correctly by students of equal ability differed by a particular group or category such as gender, race/ethnicity, socioeconomic status, disability status, or English language proficiency. DIF is discussed further in Part 10.

ITEMWIN shows tables with both the expected number correct and standard error of measurement (SEM) as functions of scale score, as well as statistical and graphical summaries of DIF, fit, and the average standard error of the test as selected. Any fault in the selection, whether the test is too easy or too difficult for the target grade, contains items showing DIF, or does not adequately cover part of the scale score range, becomes apparent as the final statistics are generated. CTB assessment editors and the CTB Research team examined these statistics for each of the WKCE selections against those of the previous operational form to confirm that each new form was parallel in difficulty to the previous operational form. In addition, the vertical properties of tests were assessed by CTB and DPI through a visual inspection of the test characteristic curves for all grades when they are plotted side-by-side, where appropriate. Finally, CTB assessment editors reviewed each selection for content diversity to ensure that no two items were similar in content.

CTB assessment editors prepared a detailed document for each selected form that summarized the test and item characteristics, submitted their selections to a content supervisor for review, and in some cases to the Content Development Lead. Appendix 1 shows the Form Selection Summary Document. The supervisor and/or manager requested changes to the selections, as necessary, in order to improve the test characteristic curve or standard error curve. Form selections were then submitted to the CTB Research team for review. Additional revisions

may have been requested at this stage. For the Reading and Mathematics selections, it was important to ensure the test characteristic curves for all grade levels formed a progression. The CTB Research team reviewed the form selections to ensure the test characteristic curves for the 2009 forms that were also used in 2011 were as similar as possible to the 2008 forms and that curves for the anchor items were aligned closely to the test forms.

Upon approval of the selections by the CTB Research team in 2009, the CTB assessment editors submitted the selections to DPI for review. For some selections, DPI requested revisions for content, difficulty, or statistical reasons. Upon making the requested changes and submitting revised selection summary forms, all operational forms were approved by DPI. For 2011, DPI reviewed the 2009 forms and accompanying statistics and approved the re-use of these forms. Table 3-1 shows the structure of operational test forms in the Fall 2011 WKCE.

3.1.2 WKCE Field Test Item Selection

No items were field tested in 2011. Table 3-2 shows the number of items that were field tested up to 2009.

3.1.3 Quality Reviews

A smooth test administration requires that all test materials, including test books, manipulatives, and test administration manuals, align with each other. All items, page numbers, and administration times must be accurate in all components of the test program. When materials are not in alignment, not only can rework and additional costs be incurred, but there is also the possibility of jeopardizing the validity of test results and creating poor publicity. Therefore, to help ensure all documents required for the administration of a test are in alignment with each other, a Materials Integration Review (MIR) is conducted prior to moving the materials on to the Quality Assurance (QA) Department within CTB.

During the MIR, a proctor simulated the test experience by administering the test to two test takers for each grade and content area using the WKCE examiner's manual. The purpose of this review is twofold: to ensure the test materials are in alignment with each other and to verify the answer keys are correct.

In addition, a QA review was conducted on each test book and all ancillary materials. The purpose of the QA review is to ensure all publishable products meet the standards and expectations of DPI. The QA review includes, but is not limited to, the review for page number location/order, header/footer information, "go on" and "stop" signs, item sequence numbering, accuracy of directions, vertical and horizontal alignment, conventions of written English, clarity/accuracy of art, accuracy of cross-references, and that there is only one correct answer to each item. This QA review occurred at the end of the page production cycle and prior to releasing the materials to CTB's Manufacturing Department.

In addition to the MIR and QA review steps, the WKCE test books were reviewed by CTB's Technology Department to verify the scannable test books were constructed to meet CTB's scanning and scoring specifications. With each round of page production, CTB's Production Department staff viewed the position of answer choice bubbles to confirm they were "on grid" and readable by CTB scanners.

3.2 Description of the WKCE 2011 Tests

The 2011 test books contained Reading and Mathematics in a single test book at each grade for grades 3, 5, 6, and 7. The single test books for grades 4, 8, and 10 contained Reading, Mathematics, Science, Language Arts, Writing, and Social Studies. The use of a single test book, rather than multiple test books per student, was first implemented in 2009. This was done to improve data quality because the use of two booklets created problems with matching student records.

The Reading and Mathematics tests for grades 3–8 and 10 consist of custom items developed specifically for the WKCE. Language Arts, Science, and Social Studies at grades 4 and 8 consist primarily of *TerraNova* items. A few custom MC items were added to address content standards not adequately covered by the *TerraNova* items. The grade 10 Language Arts, Science, and Social Studies tests consist of custom items previously developed for Wisconsin.

3.2.1 Reading

Table 3-1 presents the configuration of the operational tests. The Reading tests for grades 3–8 had one operational passage for each of the six types of passages: short literary, long literary, short informational, long informational, poetry, and everyday text.

For grades 3–8 and 10, there was one test form given in three test sessions. Each grade had at least one set of paired reading passages with a few items that required analyzing or synthesizing ideas from the passages. Each of the three sessions had approximately 18 MC items. Two of the three operational sessions included a CR item. One of the CR items was for the reporting category "Analyzing Text," while the other was for the reporting category "Evaluate and Extend Text." Each session was allotted 40 minutes of testing time. The grade 10 test consisted of three sessions: Sessions 1 and 2 were 35 minutes and Session 3 was 40 minutes.

3.2.2 Mathematics

Table 3-1 also shows the operational Mathematics test structure. The Mathematics tests for grades 3, 4, and 5 each had three sessions. Grades 6, 7, 8, and 10 had four sessions.

In each grade, the first session was a "non-calculator" session. Grades 3 and 4 do not permit the use of calculators for any session. For these grades, if a student is provided an

accommodation that allows the use of a calculator, the calculator may not be used to answer the items in Session 1.

3.2.3 Language Arts

The operational test configurations of Language Arts tests for grades 4, 8, and 10 are presented in Table 3-1 as well. The grades 4 and 8 Language Arts tests consisted of 24 *TerraNova* MC items and 6 custom MC items that measure content standard F, “Research and Inquiry.” The session was allotted 30 minutes of testing time. There was a writing session in grades 4 and 8 that presented an operational writing prompt. This session was allotted 30 minutes. The grade 10 test consisted entirely of custom items developed for Wisconsin. The test was administered in two sessions; the first session contained 30 MC items, and the second session contained the writing prompt.

3.2.4 Social Studies

Table 3-1 also presents the operational Social Studies test structure. The Social Studies test at grades 4 and 8 consisted almost entirely of *TerraNova* items, but also included a few custom items previously developed for the WKCE. There was one test session at these grades, which was allotted 40 minutes. The grade 10 test consisted of 50 custom MC items developed for Wisconsin. The test was administered in two sessions. Each session was timed at 25 minutes.

3.2.5 Science

Table 3-1 presents the operational Science test structure as well. The Science test at grades 4 and 8 consisted almost entirely of *TerraNova* items, but also included a few custom items previously developed for the WKCE. There was one test session at these grades, which was allotted 40 minutes. The grade 10 test consisted entirely of custom items developed for Wisconsin. The test was administered in two sessions; each session was allotted 25 minutes.

3.3 Customer Approvals

The development phases where DPI approval was obtained included the following:

- pre-content and bias review of new items
- item content and bias review
- item selection for the Fall 2009 WKCE form that was reused in 2011
- first pages in 2011
- final pages (prior to release to Manufacturing)

More specific information describing DPI’s role during the development phases is overviewed in the following sections.

3.3.1 Item Content and Bias Review

Following the review of items, CTB and DPI staff reviewed the edits recommended by the educator committees. DPI gave final approval of educator recommendations. DPI and CTB each kept a copy of the item review book with the edits marked.

3.3.2 Item Selection Approval

In 2009, CTB submitted item selection summaries to DPI for the 2009 test form, which were subsequently re-administered in 2011. Item selection summaries included test characteristic curves and standard error plots, lists of the items selected, and summary test statistics. DPI approval was obtained using a sign-off form.

3.3.3 First Pages Approvals

CTB assessment editors submitted copies of the test book manuscripts to the CTB Production team. The manuscripts show the items as sequenced within test sessions. The manuscripts for the test administration manuals were also submitted to DPI for review, and content changes were addressed at this stage. DPI approval was obtained using a sign-off form.

The Production team returned the test book pages to CTB style editors as first pages. CTB style editors reviewed first pages to ensure pages followed the proper format. CTB assessment editors reviewed first pages for format and content issues. Assessment editors marked first pages to indicate content changes requested by DPI on the manuscript sign-off form. CTB assessment editors submitted a copy of first pages with correction markup to the Production team, and the edits were incorporated in the InDesign files. CTB editors reviewed the corrected pages before submitting them to DPI for review. If an edit was not incorporated correctly, it was re-marked for correction.

3.3.4 Second Pages Approvals

Because of the re-administration of the 2009 test forms, it was determined that second pages approvals were not needed for this administration.

3.3.5 Final Pages Sign-Off

The final pages represent DPI's last opportunity to review test book and test administration manual pages prior to releasing the materials to CTB's Manufacturing team. At this stage, the materials had been through CTB's quality assurance process and all queries had been resolved. The focus of this review was to verify that previously requested edits had been made and that there were no errors in content or conventions of standard written English. DPI approval was obtained using a sign-off form.

Part 4: Test Administration

In the Fall of 2011, Wisconsin administered assessments in Reading and Mathematics for grades 3–8 and 10 and Language Arts, Social Studies, and Science for grades 4, 8, and 10. The test administration window was October 24–November 25, 2011. Part 4 of the Technical Report describes a set of standardized procedures and policies applied to administer WKCE assessments. The issue of test security in test administration has important implications for the integrity of the results and thus the validity of WKCE scores. Documentation citing the written procedures provided to test administrators and school personnel in order to standardize the administration of the test are also provided in this part. The following AERA, APA, & NCME (1999) standards are addressed in Part 4: 1.13, 3.3, 3.19, 3.20, 3.21, 5.1, 5.2, 5.3, 5.4, 5.5, 5.6, 5.7, 6.11, 6.15, 9.1, 10.1, and 10.2.

DPI is committed to the proposition that all schools, and all students within schools, will be held accountable to a common set of high academic content standards. Students who have an Individualized Education Plan (IEP)—a 504 plan (under Section 504 of the Rehabilitation Act of 1973)—or are identified as limited English proficient (LEP) or formerly limited English proficient (FLEP) may be eligible to receive testing accommodations. Accommodations are changes in the routine conditions under which a student takes an assessment in order to provide the student equal opportunity to demonstrate his or her knowledge. The types of accommodations and guidelines for test administration conditions are described below.

4.1 Accommodations

Accommodations were allowed for eligible individual students participating in the WKCE. 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 WKCE accommodations policy (Appendix 2 and <http://dpi.wi.gov/oea/pdf/accomswd.pdf>). Test administrators indicated which accommodations were used by each student by completing the Student Assessment Report, which is located on the back cover of the student answer document. The following accommodation information was collected from the Student Assessment Report:

Type of Accommodation:

- Used translation
- Signed test questions and content to student
- Used Braille
- Used assistive device (e.g., text-talker, adaptive keyboard, picture symbols)
- Used objects or manipulatives
- Used another DPI-approved accommodation
- Used a non-allowed accommodation resulting in the invalidation of test results

For the Fall 2011 WKCE administration, the State of Wisconsin developed Spanish and Hmong translation scripts for the WKCE. The aim of these scripts is to better help students demonstrate their knowledge on the WKCE without the interference of language. Students whose native language is Spanish or Hmong were given the choice to use all or parts of the translation accommodation, which included a bilingual word list of commonly used content area vocabulary, translation of the test directions, and a written translation script of Mathematics, Science, and Social Studies test items. DPI recommended that educators also consult the list of allowable accommodations in order to create the most appropriate testing situation for their students.

DPI recognizes that approximately 5% of the Wisconsin limited English proficient population speaks a language other than Spanish or Hmong. Districts who serve students who speak languages other than Spanish or Hmong were allowed to use qualified translators to provide oral translation support to students. However, the use of translation support was restricted to Mathematics, Science, and Social Studies tests.

Table 4-1 provides the list of standard accommodations made available for the Fall 2011 WKCE assessments and the number and percent of students provided these accommodations. Table 4-1 also provides a summary view of the accommodations provided, based on all students. The table is split across pages by accommodation, with one accommodation per page. Additional accommodation tables were also delivered to DPI from CTB, which detailed the accommodations provided for subgroup populations of interest, including gender, race/ethnicity, socioeconomic status, disability status, English language proficiency, and migrant status.

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

The primary goal of test security is to protect the integrity of the assessments. To ensure that trends in achievement results can be calculated across years and in order 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. Access to test materials was limited to those educators who required access. DPI ensured that all who had access to test materials understood the critical need for test security. They presented security requirements during the 2011 Pre-Test Workshops and outlined the acceptable and unacceptable test preparation and administration practices (Do's/Don'ts sheet provided in the Test Coordinator Kits). All WKCE tests were administered under secure testing conditions established by DPI.

The following Wisconsin Student Assessment Security Warning Statement was directed by DPI to appear on every test book beginning with the 2004–05 school year:

Test Security

All passages, stimuli, and questions used in the *Wisconsin Knowledge and Concepts Examinations—Criterion-Referenced Test* are CONFIDENTIAL and must be kept SECURE at all times. Unauthorized use, duplication, or reproduction of ANY or ALL portions of the test materials is prohibited. Violation of security can result in district disciplinary action, prosecution, and/or penalties by the Department of Public Instruction or CTB/McGraw-Hill.

Other security measures for WKCE test administrations are described below.

Limited English proficient students and students with disabilities were allowed to use highlighters. Test administrators were instructed to carefully supervise the use of highlighters because they may cause smudging of pencil marks and bubbles, which could affect reliability of scanning and scoring. If highlighters were used, the following guidelines were provided:

Guidelines for Highlighters:

1. Do not allow the highlighting of track marks, litho codes, skunk lines, barcodes, preslugged bubbles, or any carbon black printing. The highlighters cause these black inks to blur and bleed.
2. Do not allow the highlighting of pencil marks of any kind, whether bubbles or handwriting. The highlighters cause pencil marks to blur and bleed.
3. Use only a highlighter from the following list, which were tested and found to have minimal problems:
 - Avery Hi-liter
 - Avery Hi-liter, thin-tipped
 - Bic Brite-Liner
 - Sanford Major Accent
 - Sanford Pocket Accent, thin-tipped

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 test examiner. In addition, students must not be allowed to use any form of wireless communication during these breaks.

Parameters for marking test books with a No. 2 pencil:

- Do not mark in the bubble answer positions.
- Do not mark in the student Pre-ID Barcode on the barcode label.
- Do not mark in the timing tracks (the parallel lines along the side of the test book).
- Do not mark in the skunk lines (the little squares and rectangles across the bottom of each page of the test book).
- Do not mark in the litho codes (the squares and numbers across the bottom of the document on the first and last pages of the test book).
- Do not mark more than one answer bubble as the scanner cannot determine a response.

4.4 Test Administration

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 (DPI, 2011–2012). The guide included the following topics:

- Test Security
- Test Materials and Procedures
- Packaging the Test Materials
- Procedures for Returning Materials
- Test Results
- Responsibilities of District Assessment Coordinators (DACs)
- Responsibilities of School Assessment Coordinators (SACs)
- Checklist for School Assessment Coordinators
- WSAS Policy and Procedure Manual

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

- Test Materials
- Test Security
- Testing Schedules
- Organizing the Classroom
- Preparing Students to Take the Test
- Use of Appropriate Test Procedures
- Filling in the Student Information Page
- Administering the WKCE
- Filling in the Student Assessment Report
- Assembling Materials for Return

For specific information related to test administration, refer to the Test Coordinator's Manual and/or the Test Administration Manuals that are available online at:
<http://www.dpi.wi.gov/oea/publications.html>.

Part 5: Scoring

The purpose of Part 5 is to demonstrate adherence to AERA, APA, & NCME (1999) standards for scoring, including 3.22, 3.23, 3.24, 5.8, and 5.9. Part 5 describes the following:

- The scoring process of MC items
 - The scanning process
 - The calibration of scanners and other quality-control measures

- The scoring of CR items
 - The scoring rubrics
 - The handscoring process
 - The electronic handscoring system
 - The selection of Scoring personnel
 - The selection of anchor papers
 - The distribution of CR item scores

5.1 Scoring of Multiple-Choice Items

At the conclusion of the Fall 2011 WKCE administration window, student test documents were returned to CTB's scoring facility by the districts. Test materials were tracked through the entire scoring process, from the initial retrieval of the student test documents, through all scoring processes, and on to the final document retention period.

CTB's Scoring Operations processes were organized into Lean Processing Scanning Cells. Each cell was a self-contained, cross-functional team made up of the stations, equipment, and personnel skill-sets necessary to efficiently and accurately complete the operational processing cycle for student test documents.

Student answer documents were handled in a series of distinct processes. In order, those processes were as follows:

Receiving—Answer documents were tracked from retrieval to receipt at CTB, checked for damage in shipping, verified for full box counts, registered into an internal tracking system called the On-Hold Tracking System (OHTS), and then passed along to Login.

Login—Answer documents were then removed from the boxes, the pre-work was verified for district accuracy, and stacks of answer documents were aligned and cut for scanning.

Scanning—Stacks of answer documents were fed through optical scanners (see the following section for details) and any scanning problems were monitored and rectified (also detailed below).

Updates—The raw scoring and editing of scanned student data were performed using a system of edits to verify the integrity of each batch of scanned answer documents. The raw scoring and editing of the scanned student data also yielded an error list. Errors were resolved by trained editors using pre-defined guidelines in the Winscore editing system.

Documents were moved directly from process to process or sat momentarily in mini-queues. Once this stepwise process was complete, the student test documents were prepared for secure document retention.

Document Retention—Student test documents were then moved to a staging area where they were caged, warehoused, and ultimately retained for retrieval during the specified retention period. At the end of the 365-day retention period established in the WKCE contract, and upon customer approval, these documents will be loaded into containers provided by a designated NAID-certified² secure destruction company following strict national guidelines. The documents will then be picked up and shredded within 24 hours. Until shredded, the documents are caged and locked in a secure environment.

5.1.1 Scanning and Calibration of Scanners

This section provides a description of the scanning process and quality control processes applied in the scoring process.

Optical scanners captured all MC, ancillary, and student demographic data. An optical scanning technology called Optical Mark Recognition (OMR) detected all pencil marks in the answer section of the scanned document. The student test data was processed through CTB's proprietary Winscore editing system. The Winscore scanning program evaluated detectable marks on both sides of each page, recording the intensity and coordinates of solid marks for resolution in the raw scoring step. The scanner reported intensities in the range 0 (lightest) to 15 (darkest). Winscore scored the darkest mark for each question as the intended response. In this way, completed bubbles were turned into characters of data representing test item responses or other information.

The scanning production systems separated the MC item data from the CR item data. The CR data was handled in a "handscoring" process, as described in Section 5.2. The MC data and the handscoring data were later merged for correction, analysis, and reporting.

CTB's scanning software captured student response data in images called TIFFs. The scanning process also captured data in barcodes and in identification marks (i.e. "skunk marks"), which were used to determine the type of document. Document headers provided customer identification and district, school, and class information. All images were captured during scanning using high-resolution technology, also called "grayscale." Any item determined to be "unclear" was electronically retrieved in grayscale in the Electronic Handscoring System (EHS).

² NAID is the National Association for Information Destruction.

The optical scanners were able to run at a rated speed without any interruptions except for problems with the physical documents. At the beginning of each shift, and after scanning every 5,000 sheets, a diagnostic sheet was used to assess the camera functionality. CTB cell leads also cleaned the scanners at the end of their shifts and ran a “quick check utility” to confirm that the equipment was ready for the next shift. If the scanner did not pass the quick check, or a diagnostic check, a field engineer was then called in to address the problem. If the scanning camera was adjusted in any way, the scanner was recalibrated and the quick check utility was run again. When readied, the scanner was then released for scoring. All scanners were calibrated as scheduled.

No recalibration was necessary during the WKCE Fall 2011 administration. The following processing metrics were obtained:

- Number of sheets scanned: 47,264,236
- Number of books scanned/processed: 431,363

The following checks were used to ensure the integrity of the student response data:

Reliability check—When there were low scores, either among groups or at the individual student level, the reasonability of the low-score ranges was verified.

Biographical data—During the Winscore process, a series of checks were completed on critical Wisconsin fields, such as student name, gender, and date of birth. The system flagged missing, double marked, or invalidly marked data. When a record was flagged for any critical Wisconsin field errors, the document was pulled and the bubbled data was verified and corrected accordingly.

Duplicate barcode and litho code checks—Additional checks were completed in Winscore to ensure that each document was scanned only once. A duplicate checker in Winscore flagged duplicate barcodes and litho codes. If either was flagged, the book was pulled and the barcode or litho code was verified to ensure that it had been accurately scanned, that no document was scanned twice, and that no barcode labels had been incorrectly applied. In addition to checks carried out in Winscore, further checks were carried out in Monarch, a back-end data system that flagged duplications and matched district and school data.

Student counts—The actual book counts generated by the scanners were compared to the book counts provided by the school districts on the School Group List and School Header Sheet. In 2011, 990 discrepancies were identified and resolved by emails and telephone calls placed to the districts. These completeness checks occurred from December 2, 2011, to January 4, 2012.

School name/number—Pre-assigned school numbers and names were verified against data provided by DPI.

The scored student response data were later retrieved by the CTB Research and Technology teams for statistical analyses and for producing reports.

5.2 Scoring of Open-Ended or Constructed-Response Items

Sections 5.2.1 through 5.3 document the scoring processes used for CR items. This documentation forms part of the validity evidence supporting the scoring process used for CR items. Sections 5.2.1 through 5.3 describe the scoring rubrics, the scoring process, the selection of sample (anchor) papers used to train scoring personnel, the process of selecting personnel, inter-rater reliability, and the distributions of scores from CR items.

5.2.1 Description of Scoring Rubrics

In the 2011 administration, the Reading and Mathematics forms in grades 3–8 and 10 contained CR items. A Writing prompt was also administered at grades 4, 8, and 10. The Writing prompts were scored using two holistic rubrics: a 3-point Conventions Rubric and a 6-point Composing Rubric. Tables 5-1 through 5-8 present the scoring rubrics.

5.2.2 Handscoring Process

The Scoring personnel who score CR items are referred to as “readers.” As indicated previously, the process of scoring CR items is referred to as “handscoring.” The handscoring readers were trained using customer-approved training materials, such as the anchor papers described in Section 5.2.4. Once qualified, readers were required to maintain accuracy standards throughout the project. These requirements were assessed at the item level primarily through each reader’s daily “checkset” performance (described below), as well as agreement rates with other readers on the second reads (described below), and targeted read-behinds with team leaders (described below). Data monitors generated reports daily that flagged any readers falling below the established quality standards for any item, providing insight on reader scoring trends (such as difficulty with any particular score point). These reports were shared with handscoring supervisors. Those readers identified in the reports received additional coaching, training, reviews, targeted read-behinds, or additional checksets. Readers who did not meet standards with these initial corrective actions were administered another validation (recalibration) round. Failure to recalibrate resulted in dismissal from the scoring assignment. This process was in place throughout the entire handscoring window.

5.2.3 Electronic Handscoring System

The Electronic Handscoring System (EHS) was used to score CR items. EHS presented images of scanned test books to trained readers, who assigned scores for the CR items. The scanned student responses were viewed on high-quality, 19-inch workstation monitors. Images of each student's responses were automatically routed to two or more readers when required, and images of specific subsets of test items were routed to designated groups of readers trained to score these items.

5.2.4 Anchor Papers and Training Papers

In 2011, all training materials, including scoring guides and Reading and Mathematics rubrics, anchor papers, training papers, qualification round papers, and checksets, were from the 2009 operational administration. Prior to the actual scoring in 2009, the CTB Scoring Center created training materials. A selected group of papers written by WKCE students were selected as models to train raters for scoring. These papers, referred to as "anchor papers," played an important role in deciding which level of writing should receive which score. Range-finding meetings were held with DPI staff and educators to select sample papers for each score point. CTB randomly sampled student answer documents to ensure a representative sample of the possible responses. The sample papers were used to construct scoring guides and training papers. CTB's Scoring team collaborated with DPI to make necessary revisions to the rubrics and in the selection of scoring guides and training papers. This process included several pre-sorting steps and subsequent iterative/consensus processes in order to achieve agreement and precision through a "round robin" scoring process. Once approved by DPI, the scoring guides (consisting of rubrics, anchors, and annotations) served as a constant guide, setting the course for all subsequent training and scoring.

5.2.5 Scoring Personnel and Qualifications

CTB recruited, trained, and managed personnel to complete all of the handscoring operations within the timelines of the contract. This involved extensive consultation between CTB's Scoring and Publishing Departments, Wisconsin educators, and DPI in order to review scoring rubrics, develop the anchor papers and other reader training materials, and to provide analyses of student responses to tryout forms. The characteristics of the readers, team leaders, and scoring supervisors are described in the following sections.

Readers—Many CTB readers had years of classroom teaching experience. The CTB reader pool included many retired and current educators, as well as engineers, editors, published authors, and individuals with advanced degrees. The minimum qualification for all readers was a Bachelor's degree. Readers were required to participate in training and successfully pass at least one of two qualification rounds. Once qualified, readers could start scoring, but throughout the scoring process, reader performance was assessed by a supervisor and data-monitoring staff through the use of checksets, read-behinds, and the review of inter-rater reliability statistics, as described in Sections 5.2.7 and 5.3 and in 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 readers. Team leaders were each responsible for a small team of readers. In addition to performing read-behinds on readers, team leaders also coached readers when needs were identified through data monitoring or otherwise by supervisory staff. Team leaders working on the writing component also resolved discrepant scores.

Scoring Supervisors—Scoring supervisors were the core group at CTB who directed and organized the assessment process and trained team leaders and readers. Scoring supervisors had extensive experience as team leaders prior to their qualification and selection. Scoring supervisors were content area experts in the content areas they supervised and trained. They oversaw all team leaders and readers.

5.2.6 Reader Training

Validation was a critical task in the training process and the final determinant of reader readiness. All readers, 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 dependent on the score point range of an item. For example, where scores were either zero or one point, the level of agreement required was 95%, but where scores could range from zero to two points, the level of agreement required was 90%. Those readers not validating on the first attempt received further training prior to taking an additional qualifying round. Only those who were successfully validated were qualified as readers to score tests. Team leaders were required to complete two validation rounds with at least 80% exact agreement in each round.

5.2.7 Inter-Rater Reliability

Checksets—Throughout the course of the handscoring process, sets of pre-scored papers called “checksets” were administered daily to the team leaders as well as to the readers. The checksets were used to monitor scoring accuracy and to maintain a consistent focus on the established rubric and guidelines. This kind of monitoring occurred without reader knowledge. Readers whose checkset scores fell below the qualifying level were flagged for additional coaching (training review, targeted read-behinds, etc.). Those readers who remained below standard were given another validation (recalibration) round. Readers unable to recalibrate were dismissed.

Read Behinds—The “read-behind” was another valuable monitoring technique used. Each team leader was able to read a random selection of a reader’s scored items. This reading could be targeted at the item and score point level. The scores were compared, and if they agreed, the team leader was able to offer feedback, which enhanced the reader’s confidence and ability to score quickly and accurately. However, if a reader strayed from the standards established in the training and validation samples, the aberrant scoring was detected, and the

team leader was able to offer guidance necessary to refocus the reader’s effort. Readers whose scoring was inconsistent were read behind more frequently by their team leaders, thus correcting any scoring variations.

Final Score—In Writing, Reading, and Mathematics, the first score assigned for each CR item was the final score; however, 5% of the responses per item were double read (in “second reads”) to obtain indices regarding the consistency and accuracy of raters. Inter-rater reliability was monitored throughout the scoring process, as described in Part 9.

5.3 Distribution of Constructed-Response Item Scores

Tables 5-9 through 5-16 show distributions of CR item scores across each score point level (one point, two points, etc.) for each CR item and the Writing prompts. The scoring distributions shown for Reading and Mathematics are the scoring distributions of the first read. As described previously, 5% of the responses to the CR items in Reading, Mathematics, and Writing were double read (in “second reads”) for statistical purposes. These distributions were examined for quality assurance purposes in the scoring process.

These tables use four condition codes. Condition code “A” denotes items with no response or no attempt, code “B” represents an illegible response, code “C” indicates that another language was used in the response, and code “D” denotes a response that was off topic.³

All Reading items had one part and a maximum score of three points. In Mathematics, many CR items in grades 3–8 had two parts: a Part A worth one point and a Part B worth two points. The CR items in grades 3–8 with only one part were worth two points. In grade 10, all Mathematics CR items had one part and were worth two points.

As can be seen in Table 5-9 for Reading, in most cases, most students scored one or two points, and fewer students scored either three points or zero points. Scoring three points was not common in Reading; however, this result may be expected because CR items are often more difficult than MC items.

In Mathematics, although many students scored at the maximum score level for the CR items, many students also obtained a score of zero. This occurred on both Part A and Part B of the two-part CR items.

Students in grades 4, 8, and 10 were administered one Writing prompt. Tables 5-11 through 5-16 present the score distributions for the student responses to the Writing prompts. These tables are split between counts and percentages, and separate tables are provided for the 6-point Composing Rubric and the 3-point Conventions Rubric. The first score assigned for each Writing response on each rubric was the final score; however, 5% of the responses per prompt were double read (“second reads”) to obtain indices regarding the consistency and accuracy of

³ When calculating students’ scores on operational items, CR items receiving these condition codes were given zero score points.

raters. Scores from the first read, the second read, and the difference between the two reads are presented in Tables 5-11 through 5-16. As can be observed in Tables 5-11 through 5-16, the rater scores were very similar. As described previously, inter-rater reliability was also monitored in other ways throughout the scoring process. The full results for inter-rater reliability are presented in Part 9.

As can be seen in Tables 5-11 and 5-12, most scores in the Composing Rubric were in the middle of the 6-point range, and relatively few students were at the low and the high extremes. The Conventions Rubric showed similar results. As can be seen in Tables 5-13 and 5-14, a large proportion of students scored in the middle level of the 3-point range for the Conventions Rubric, and relatively few students scored either 1 point or 3 points.

Tables 5-15 and 5-16 show the total score on the Writing prompt, combining scores from the Composing Rubric and the Conventions Rubric. The combined scores for most students were in the middle or upper-middle range of the 9-point total, from 4 points to 6 points. The highest and lowest levels of scoring were less common, but in every grade, a small proportion of students obtained zero score points and a small proportion obtained the highest possible score.

Part 6: Characteristics of the Calibration Sample

The calibration, equating, and scoring of the Fall 2011 WKCE occurred in 2009 based on student data from a preselected sample of districts in the state. This arrangement was chosen in order to expedite the return of score reports to districts. In accordance with AERA, APA, & NCME (1999) standards 1.5, 1.13, 2.4, 4.7, and 6.1, this section provides a description of how the 2009 calibration sample was selected and how the calibration sample and census data compare in terms of demographic characteristics. Part 6 serves to demonstrate that the 2009 calibration sample was sufficiently representative of the Wisconsin student population for the purposes of calibration. This documentation also serves as validity evidence supporting the WKCE program. Information about the calibration sample can be found in the 2009 WKCE Technical Report available from the DPI at: <http://dpi.state.wi.us/oea/pdf/td-2009-techman.pdf>.

Part 7: Calibration, Equating, and Deriving Scale Scores

Student responses on the WKCE are input into complex mathematic 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). WKCE scores are established through the processes of calibration, equating, and item-pattern scoring. Part 7 of the Technical Report describes these processes as they were applied to the Fall 2011 WKCE administration, as well as the results. This portion of the Technical Report addresses AERA, APA, & NCME (1999) standards 1.13, 4.1, 4.2, 4.3, 4.10, 4.11, 7.1, 7.2, and 7.10.

Readers should note that calibration, equating, and scoring using IRT are mathematically complex and computationally intensive processes and 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 7.3. Additional references are also provided.

Calibration is the mathematical process of estimating characteristics of individual items. These characteristics are termed “item parameters.” Sections 7.1, 7.1.1, and 7.1.2 serve to explain this process beginning with a description of the calibration and equating methods used in 2009 that were applied to the Fall 2011 WKCE, followed by a discussion of the calibration models and the software used. The derivation of scale scores from raw scores is then addressed, with a focus on nontechnical audiences. The results of the calibration process, using model-to-data fit statistics and the standard error of measurement (SEM), are also discussed.

7.1 Calibration and Equating Methods

In the Fall 2009 WKCE, the three-parameter logistic (3PL) IRT model (Lord & Novick, 1968; Lord, 1980) was used for MC items, and the two-parameter partial credit (2PPC) model (Muraki, 1992; Yen, 1993) was used for CR items. Language Arts, Social Studies, and Science items were calibrated using the 3PL model because these three content area tests comprise only MC items. Because the Reading and Mathematics tests consisted of both MC and CR items, a simultaneous calibration with the 3PL and 2PPC models was implemented. A simultaneous calibration was also applied to the Language Arts test in grade 10 because a Writing prompt was included as a component of a student’s scale score at this grade level. The 3PL and 2PPC models are described in detail in the next section.

Simultaneous calibration is used for the mixed format tests in part because a single scale communicates that the measured skills relate to the same underlying qualities and characteristics and that they can be taught and measured using a variety of assessment modes. In considering the simultaneous calibration process, Thissen, Wainer, and Wang (1992) stated that items of diverse types can be scaled together provided that the different types of items assess the same primary characteristics of the content area.

By design, there was a special set of items in each content area and grade level test that were common to both the current (2009) administration and a prior administration. This arrangement is called a *common item non-equivalent group* design. The purpose of this design is to place current operational items on a base scale using the common items. Horizontally equating the current test forms to the previously established scales is necessary in order to obtain results that are comparable across administration years. The equating process also mitigates differences in test difficulty between forms from the current and the previous year, which are built to be similar in difficulty and content (Kolen & Brennan, 1995). The items that were used for equating are called anchor items. In each grade and content area, each set of anchor items was a miniature version of the total test, which adequately represented the test content coverage in terms of item difficulty and the test specifications. The Stocking and Lord (1983) procedure was used to equate the estimated parameters to the scale from which the anchor items were drawn. This procedure estimates the linear transformation constants by minimizing the distance between the test characteristic curves for the calibrated anchor items and the values for the anchor items already on the test scale.

The Reading and Mathematics vertical scales were established in Fall 2005 using a similar plan termed an *adjacent grade common item design*. Based on Fall 2004 data, scores for adjacent grades were linked so that student scores in grades 3–8 and 10 could be expressed on a single scale. Vertical scales were not developed for Language Arts, Social Studies, and Science because these tests were administered only in grades 4, 8, and 10. Instead, the scales for grades 4, 8, and 10 were constructed in such a way so as to show a vertical relationship (i.e., an increase in scale score means) across grades. For additional information on the scaling methods used to establish the WKCE scales, readers can refer to Part 8 and Part 11 of the WKCE Technical Report from the Fall 2005 WKCE administration, which can be found in Appendix 3 of the 2010 Technical Report, which can be found at: <http://dpi.state.wi.us/oea/pdf/td-2010-techman.pdf>. The 2005 Technical Report includes a fairly extensive discussion of the scaling methods.

7.1.1 Calibration Models

The 3PL model defines a MC item in terms of three characteristics, or *item parameters*: (a) item difficulty (or its location on a scale of difficulty/ability), (b) item discrimination (or how well the item differentiates between the low- and high-ability students in relation to its location), and (c) the level of guessing. The 2PPC model defines a CR item in terms of item discrimination and item difficulty for each score point.

In the 3PL model, the probability that a student with scale score θ responds correctly to item i is

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-1.7a_i(\theta - b_i)}},$$

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

The 2PPC model is a special case of Bock's (1972) nominal model. Bock's model states that the probability of an examinee with ability θ having a score at the k th level of the j th item is

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

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

For the special case of the 2PPC model used here, the following constraints were used:

$$A_{jk} = \alpha_j(k-1) \text{ and } C_{jk} = -\sum_{i=0}^{k-1} \gamma_{ji}, \text{ where } \gamma_{j0} = 0,$$

where α_j and γ_{ji} are parameters freely estimated from the data. The first constraint implies that higher item scores reflect higher-ability levels and items can vary in their discriminations. The 2PPC model estimates a total of m_j independent item parameters; for each item, there are $m_j - 1$ independent γ_{ji} parameters and one α_j parameter.

The item calibration process is a process of estimating item parameters. Parameters are estimated in an iterative process using a computer software program called PARDUX (discussed below). The PARDUX program operates by estimating person parameters (ability) and item parameters (e.g., difficulty) through a series of iterations until the change in parameter estimates between iterations is reduced to a given threshold.

7.1.2 Calibration Software

The IRT models and the student response data from the Fall 2009 WKCE administration were used to estimate item parameters for each test. The IRT models were implemented using CTB's PARDUX software (Burket, 1991). Using marginal maximum likelihood procedures implemented with the expected maximum algorithm, PARDUX estimates parameters simultaneously for MC and CR items (Bock & Aitkin, 1981; Thissen, 1982).

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 it 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).

7.2 Calibration Results

The following sections describe the calibration results in terms of the estimation of item parameters, model-to-data fit, and the SEM of the scale scores across content areas and grades.

7.2.1 IRT Item Parameters

At times when calibrating items, items do not converge, meaning the characteristics of the item are not able to be determined. When this occurs, items are suppressed from student scoring and future assessments. In 2009, no convergence issues occurred for any item on the operational tests.

7.2.2 IRT Item Fit

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

CTB 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_j) 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, 1991a)

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

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_j), are computed for ten intervals corresponding to deciles of the ability distribution (Yen, 1984).

Table 7-1 presents items that were flagged for less than optimal fit when the obtained Z -statistic exceeded the critical Z -statistic value. To take an example from the table, in Reading grade 6, item 23 was flagged because the observed Z of 23.29 is larger than the critical Z value of 16.83 based on a sample size of 6,312.

Table 7-1 specifies the item status, content area, grade level, test book form, item number, item type (MC or CR), N size (the number of students), Z , and critical Z , as described previously. For many of the flagged items, the observed Z and the critical Z are not very far apart. For example, in the case of the first item in the table, Reading grade 5 item 18 was flagged because the observed Z of 18.15 is larger than the critical Z value of 17.07. The misfit in this case may be considered small. Although many items in the table show a moderate degree of difference between the obtained Z and the critical Z statistic, others such as the Reading grade 7 item 62 show much larger differences.

In order to evaluate item-to-model fit further, CTB 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, the 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.

MC items flagged for misfit most commonly had empirical (observed) information that differed from the model in the lower-ability range or at the higher-ability range because there are fewer students to provide information at the tails of the distribution. Similarly, for CR items, there are, in general, smaller numbers of students at the lower and higher score levels, which provides less information at the tails of the student distribution. Items that only show misfit at the tails of the distribution provide stable information about the majority of the students—those in the middle range of the distribution. However, if the misfit happens around the middle of the ability range, where there are many students, this may be a concern and may lead to the item being dropped from the test.

In a large-scale assessment such as the WKCE, with 23 grades and content areas, it is expected that some items will be flagged for misfit. The number of items flagged for misfit in the Fall 2009 WKCE is consistent with, though slightly more than, the number flagged in the year prior. 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 CTB Development team

and DPI. As noted in Section 3.1.1, such items are avoided in future selections unless there is a compelling reason that they should be included, such as meeting the test blueprint.

7.2.3 Evaluating Anchor Items

To evaluate whether anchor items are performing differently in the current administration versus a previous administration, differences between the ICCs were computed for each anchor item. Differences between the curves were evaluated using the following statistics:

- Average Signed Difference
- Average Absolute (Unsigned) Difference
- Root Mean Squared Difference

Both unweighted and weighted versions of these statistics were calculated. Unweighted differences gave equal weight to differences across the ability spectrum. Weighted differences assigned weights according to the number of test-takers that are impacted by differences in the curves. For both weighted and unweighted versions of the three statistics listed above, differences greater than + 0.10 were considered large, and differences between + 0.07 and 0.10 were considered moderate. In addition, the Maximum Absolute Difference was identified for each item. Large Maximum Absolute Differences were those greater than + 0.15, and moderate differences were all differences between + 0.125 and 0.15.

Although dropping an anchor item flagged based solely on statistical criteria has its simplicity, this option may change the content coverage and equating constants, shift scale score distributions, and affect the performance level classification of students by moving them into different proficiency levels. Before an anchor item may be dropped from an anchor set, the adequacy of the content coverage must be evaluated and a reason for the anchor item differential performance must be identified. As stated above, an item is removed from the anchor set only if it adversely affects quality of scaling, not desirability of results. As such, CTB does not consider how the removal of an item affects the overall mean scale score or the impact data (percent of students in each performance level) when recommending items for removal.

Items removed from the anchor set are still scored as part of the whole test. Anchor items were considered for exclusion from the WKCE under the following conditions:

1. An item is flagged for large differences on the Average Signed Difference, Average Absolute (Unsigned) Difference, or Root Mean Squared Difference and for moderate or large differences on the Maximum Absolute Difference when examining the differences between the previous versus current ICCs.
2. Alternative explanations have been considered that may explain shifts in performance. For example, performance on the anchor item may improve because of a statewide initiative emphasizing instruction on a particular set of skills. In this case, improved performance on the item represents true growth in that area. Removing the anchor item may artificially lower test scores.

3. Removal of the item may not significantly alter the content distribution of the anchor set. The distribution of the anchor items across the content standards must remain within 10% of the test blueprint.

One grade 4 Social Studies item was flagged based upon the criteria specified above. The item was investigated, and it was found that when it was previously administered the item was part of a testlet whereas in 2009 it was a stand-alone item. Because this is considered a contextual effect that can change the difficulty of an item, it was removed from the anchor set but retained as an operational item for the purpose of scoring students.

7.3 Deriving Scale Scores in the WKCE

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 scale scores account for the characteristics of the items that are in the test (such as item difficulty).

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

Whether looking at an individual item or at a group of items that make up a complete test, IRT uses probability models to describe the relationship between a student's ability and his/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., Language Arts 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 a 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*, and a *pseudo-guessing factor*. Discrimination refers to how well an item sorts students by ability level; difficulty represents the difficulty of the item or its

location on an ability continuum; and the pseudo-guessing factor represents the probability of a low-ability student guessing the correct response.

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

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

The formula indicates that the “estimated maximum likelihood” IRT item-pattern scoring method searches for the ability estimate (θ_0) that maximizes the probability function in (2) and it 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 of the items in a given test.

As indicated, the item-pattern scoring method takes into account not only a student’s total raw score, but also the psychometric characteristics of all items the student responded to, including the items the student responded to incorrectly.

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

Table 7-A. 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 the student response patterns for these six examinees are as follows, where 0 represents an incorrect response, and 1 represents a correct response:

Table 7-B. Item Response Pattern

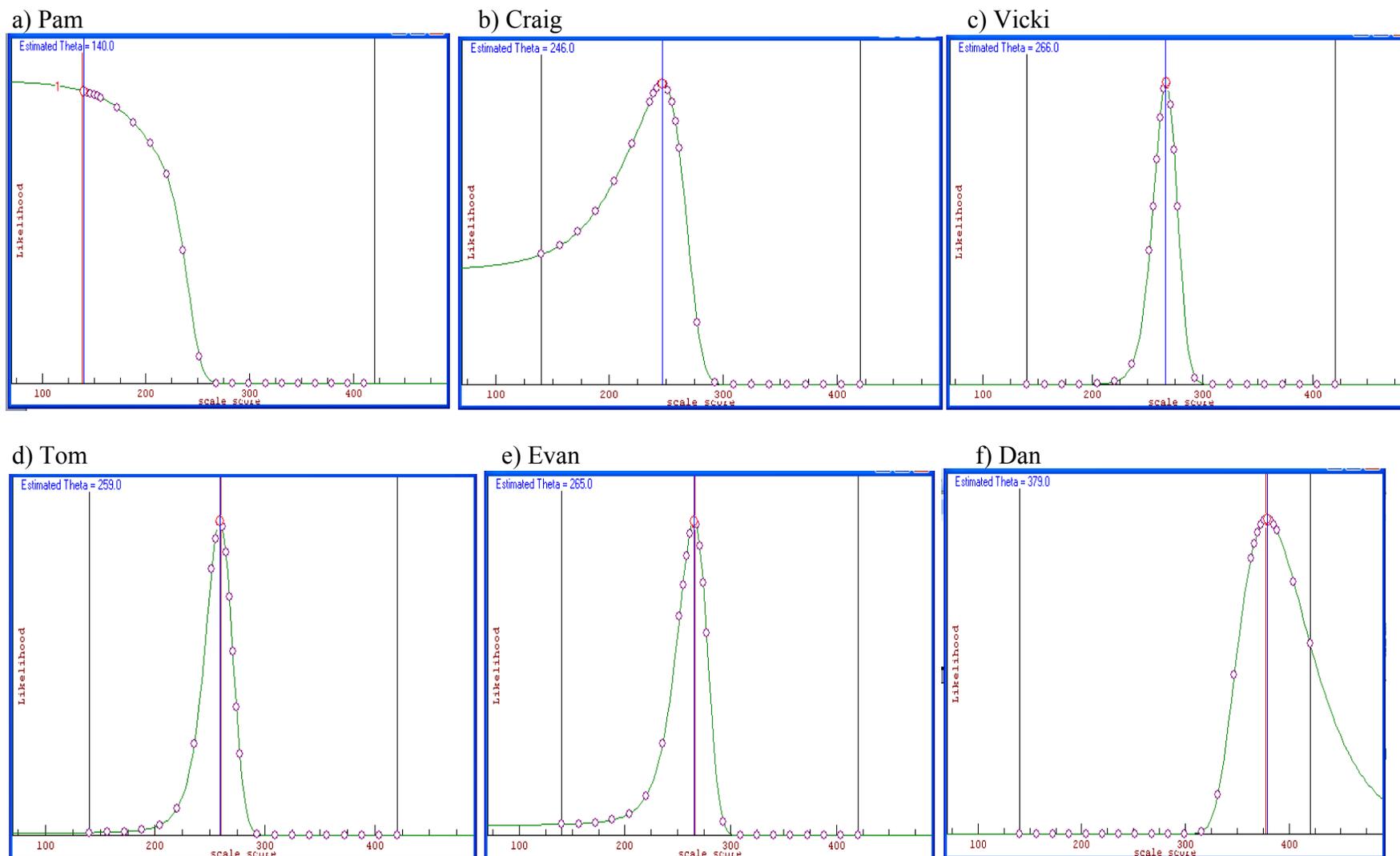
Student	Response Pattern ($u_1u_2 \cdots u_n$)	Raw Score	Item-Pattern Score
Pam	10000110010100000000000000101	7	140
Craig	101010101010101010101010101010	15	246
Vicki	010101010101010101010101010101	15	266
Tom	001100110011001100110011001101	15	259
Evan	110011001100110011001100110010	15	265
Dan	1111111111111111111111111011111	29	379

The first student, Pam, answered seven of the items correctly and obtained a scale score of 140, which is equal to the lowest point on the score range, called the “lowest obtainable scale score,” or LOSS. The next four students each answered 15 out of 30 items correctly, but the response pattern of each of these students is different. The *raw score* of each of these students is 15. However, the maximum likelihood item-pattern scoring method produced a different scale score for each examinee. Scale scores were 246 for Craig, 266 for Vicki, 259 for Tom, and 265 for Evan. These scores can be accounted for by considering the pattern of the student responses on the test together with the properties (or parameters) of the items, as shown in Table 7-A. By referring to Table 7-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 7-A below shows the probability of each ability estimate (or scale score) for the six examinees. The total scale score range for Language Arts 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 the ability estimate actually reflects the examinee’s ability level.

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

Figure 7-A. Likelihood Functions, or the Probability of Each Ability Level Estimate (or Scale Score)*



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

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

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

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

The procedures applied here are similar to those followed in the development of the *TerraNova* test (CTB/McGraw-Hill, 1997), *TerraNova* 2nd Edition (CTB/McGraw-Hill, 2000), and the prior Wisconsin Knowledge and Concepts Examinations developed in conjunction with CTB (1997–2004). Several supplements to this simplified outline of IRT are available. Introductory discussions of IRT can be found in *Educational Measurement* (Linn, 1989) or Chapter 11 in *Introduction to Measurement Theory* (Allen & Yen, 1979). More advanced discussions of partial credit models may be found in Muraki (1990, 1992), Yen (1993), and van der Linden and Hambleton (1997). For additional information on the technical details of the item-pattern scoring, readers can also refer to Yen & Candell (1991) and to *TerraNova* 2nd Edition (CTB/McGraw-Hill, 2000).

7.3.1 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 SEM can be used to obtain the range within which a student's true score is likely to fall, that is, with a certain degree of probability. It is expected that 68% of the time a student's score obtained from a single testing will fall within one SEM of that student's true score and that 95% of the time the obtained score will fall within two standard errors of the true score.

The SEM of the scale scores in the Fall 2011 WKCE, based on the Fall 2009 WKCE, is displayed graphically for each grade and content area in Figures 7-1 through 7-5. The SEM provided is based on item-pattern scoring. Each SEM curve is plotted as a function of the scale scores. These figures show the scale score range within which measurement is most accurate. The figures also show that extreme scale scores have more measurement error than scores in the

middle of the distribution. Scale scores in the high or low extremes of the student distribution are less precise than those in the middle of the distribution because there tends to be fewer test items in these score areas and fewer students. The lower and upper limits of the scale, referred to as the lowest obtainable scale score (LOSS) and highest obtainable scale score (HOSS), are the starting scale score and the last scale score in these figures. LOSS and HOSS are further discussed in the next section.

Because of the nature of item-pattern scoring, a scoring table showing a simple, direct conversion of raw score to scale score cannot be generated for the Fall 2011 WKCE. However, scoring tables showing a rough relationship between raw score, scale score, and SEM can be produced, and they are provided in Tables 7-2 through 7-24.

7.3.2 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, which are set separately by grade, are called the LOSS and the HOSS.

Table 7-25 shows the number and percent of students at the LOSS and the HOSS. In general, there should not be many students clustered at the LOSS or HOSS. An accumulation of a high proportion of students in the LOSS or HOSS may indicate a floor or ceiling effect.

In most grades and content areas, the percentage of students at the LOSS and HOSS was small: less than one percent. However, in some grades and content areas the percentages were larger. In Mathematics grade 10, 2.5% of students were at the LOSS. These percentages at the LOSS can be considered to fall within an acceptable range, although they can still be considered as a point of reference when developing future forms. The percentage at the LOSS in these grades may be reduced in future years by including some additional items that are not difficult. The percentage of students scoring at the HOSS is similar: In most grades and content areas, the percentage was small, although in a few grades and content areas, the percentage was larger. In particular, the percentage at the HOSS in Social Studies grade 4, Science grade 8, and Language Arts grade 8 ranged from 1.69 to 3.82. These tests are shorter than the Mathematics and Reading tests, so it is not unexpected that the percentage of students at the HOSS may be higher on these content areas because a smaller number of items on a test provide less psychometric information about the students. The percentage scoring at the HOSS may be reduced by including some additional difficult items in these grades and content areas or by including more items on the test.

7.3.3 Test Characteristic Curves

Test characteristic curves (TCCs) are provided in Figures 7-6 through 7-10. These curves model the relationship between student ability and expected scoring outcomes at the test level. By following the plotted line for any grade level and content area, one can observe the estimated scoring outcome (the estimated proportion of the maximum correct score) plotted as a function of examinee ability. These curves are based on the IRT models, methods, and scaling processes described above. The vertical relationship across grade levels that can be observed in the TCCs reflects the typical growth pattern: as grade level increases, ability level is also expected to increase across the ability range.

Although the TCCs, overall, show the expected separation across grades, the separation is somewhat less for Reading than for the other content areas. In addition, the Reading curves overlap in grades 4 and 5 and in grades 7 and 8. Although scale overlap is generally not considered the optimal pattern for a vertically scaled assessment, on Reading assessments this is not uncommon. On the WKCE Reading scales, the cut scores for these grades are closer together than the cut scores across grades for the other content areas. For example, the *Proficient* cut score in Reading is only 4 scale score points higher for grade 5 than grade 4, whereas the *Proficient* cut score for Mathematics is 25 points higher for grade 5 than grade 4. Because the item difficulties in the WKCE tests were chosen, in part, to minimize the standard error around the critical *Proficient* cut score, the proximity of the cut scores in grades 4 and 5 would be expected to yield curves with relatively little separation. The proximity of the curves for grades 7 and 8, however, is less easily explained. Given the greater separation between the scales at these two grades, the observed overlap of the TCCs may indicate that the grade 8 assessment would benefit from the addition of some more difficult items. This consideration, however, must be balanced by the need to keep form difficulty comparable each year to meet the assumptions for alternate parallel forms.

Part 8: Test Results

Part 8 presents a classical item analysis and summary of student results for the Fall 2011 WKCE administration. The summary results 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 called “standardized performance indicator” (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 WKCE testing program. The AERA, APA, & NCME (1999) standards addressed in Part 8 include: 1.5, 3.18, 4.3, 4.5, 4.6, 4.7, 4.19, 7.1, 7.10, 13.15, and 13.19.

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 a 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 the p -value were 0.75, this would indicate a much easier item where, on average, students scored 75% of the maximum possible points for the item. As such, the p -value indicates difficulty for CR items as well, with lower p -values indicating more difficult items.

The item-total correlation indicates the extent to which individual test items provide reliable measurement of the construct being measured by the total test, and it is an index of the item’s ability to discriminate between high-ability and low-ability students. For dichotomously scored MC items, the item-total correlations are computed as point-biserial correlations between the score on the item and the score on the remaining items in the test. For CR items, the item-total correlations are computed as Pearson product-moment correlations between the score on the item and the score on the remaining items in the test.⁴ 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

⁴ 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 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 of the items.

For the Fall 2011 WKCE administration, items were flagged for further investigation according to the following rules:

- The p -value was less than 0.30 for MC items. Such a p -value indicates a difficult item, where fewer than 30% 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 (especially easy items) or very low p -values (especially difficult items) in order to provide more reliable measurement at the extreme high and low levels of ability and to fully represent the range of difficulty for particular content standards. In this case, the flagging of p -values is a useful way of verifying that the number of extremely easy or difficult items is relatively small and consistent with the purposes of the test. Thus, flagged items do not necessarily indicate a challenge to test validity because items have been found to be appropriate during item reviews.

Omit rates may reflect a number of different properties, and an item that is omitted by more than 5% of the students (the WKCE 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 in order 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 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 that of 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, 33 items were flagged on the 23 WKCE 2011 operational tests as meeting the investigational criteria bulleted on the previous page. Of the 33 flagged items that were scored, the number flagged for each of the four criteria is consistent with previous administrations.

Table 8-A shows the number of scored items in the Fall 2011 WKCE 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.

Table 8-A. Summary of Flagged Operational Items on the Fall 2011 WKCE

Content	Grade	# of Items Flagged	Number of Flags*			
			Corr <0.15	Distractor Corr >0	Omit >5%	p-value <0.30
RD	3	1		1		
	4	1		1		
	5	3		2		1
	6	2		2		
	7	3	1	1		1
	8	3	2	2		
	10	3		2	1	
MA	3	0				
	4	1				1
	5	2	1	2		
	6	4	1	2		1
	7	1			1	
	8	0				
	10	2			2	1
LA	4	1		1		
	8	0				
	10	1		1		
SS	4	1		1		
	8	1		1		
	10	1	1	1		1
SC	4	1		1		
	8	1		1		
	10	0				
Total		33	6	22	4	6

*Note that number of flags may be greater than number of flagged items.

The flagged items were referred to CTB’s content specialists for further review to ensure that the items were unambiguous and the answer keys correct. As part of this review, CTB’s content experts also evaluated each flagged item against the WKCE depth-of-knowledge (DOK) 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.

Table 8-B. Fall 2011 WKCE Reading Items Flagged for Classical Item Analysis Statistics

Grade	Content	Item	Item Type	p-Value	Corr	Omit Rate	Flags				
							Corr	Distractor	Omit	p-Value	
3	RD	32	MC	0.46	0.42	2.04%		+	0.02		
4	RD	10	MC	0.55	0.23	0.18%		+	0.07		
5	RD	11	MC	0.60	0.35	0.82%		+	0.05		
	RD	33	CR	0.28	0.39	1.35%			.		+
	RD	42	MC	0.39	0.31	0.25%		+	0.04		
6	RD	14	MC	0.57	0.22	1.14%		+	0.05		
	RD	39	MC	0.62	0.18	0.16%		+	0.01		
7	RD	8	MC	0.60	0.12	0.25%	+		.		
	RD	21	MC	0.38	0.31	2.03%		+	0.08		
	RD	56	CR	0.21	0.48	1.44%			.		+
8	RD	1	MC	0.59	0.30	0.05%		+	0.01		
	RD	3	MC	0.77	0.10	0.11%	+		.		
	RD	38	MC	0.66	0.15	0.57%	+	+	0.02		
10	RD	1	MC	0.51	0.37	0.09%		+	0.02		
	RD	10	MC	0.68	0.25	0.25%		+	0.06		
	RD	43	CR	0.52	0.52	6.22%			.	+	

Table 8-C. Fall 2011 WKCE Mathematics Items Flagged for Classical Item Analysis Statistics

Grade	Content	Item	Item Type	p-Value	Corr	Omit Rate	Flags				
							Corr	Distractor	Omit	p-Value	
4	MA	29B	CR	0.28	0.39	2.42%			.		+
5	MA	21	MC	0.48	0.41	0.33%		+	0.04		
	MA	29	MC	0.78	0.09	0.41%	+	+	0.01		
6	MA	11	MC	0.96	0.12	0.20%	+		.		
	MA	20	MC	0.44	0.46	0.20%		+	0.01		
	MA	22B	CR	0.26	0.43	0.76%			.		+
	MA	49	MC	0.59	0.26	0.40%		+	0.05		
7	MA	12	MC	0.72	0.40	9.08%			.	+	
10	MA	27	CR	0.43	0.63	5.10%			.	+	
	MA	38	CR	0.27	0.59	8.84%			.	+	+

Table 8-D. Fall 2011 WKCE Language Arts, Science, & Social Studies Items Flagged for Classical Item Analysis Statistics

Grade	Content	Item	Item Type	<i>p</i> -Value	Corr	Omit Rate	Flags				
							Corr	Distractor	Omit	<i>p</i> -Value	
4	LA	26	MC	0.59	0.32	1.35%		+	0.07		
10	LA	23	MC	0.58	0.24	0.94%		+	0.03		
4	SC	40	MC	0.47	0.16	2.39%		+	0.07		
8	SC	20	MC	0.59	0.35	2.26%		+	0.01		
4	SS	33	MC	0.49	0.29	1.90%		+	0.02		
8	SS	18	MC	0.63	0.20	0.29%		+	0.06		
10	SS	16	MC	0.22	0.12	0.21%	+	+	0.03		+

8.1.1 Flagging for a Positive Distractor Correlation

The distractor correlation coefficients are provided in these tables for items that were flagged because of positive distractor correlations. The distractor correlations tend to be very small and are generally much smaller than the item-total correlations for the correct answer key. All items flagged for a positive distractor had a distractor less than 0.08. These items were judged to be acceptable on the basis of their other statistics and were retained in order to meet the WKCE test blueprints.

8.1.2 Flagging for the Item-Total Correlation

Six items were flagged for item-total correlations <0.15 , and all of the flagged items were 0.12 or above except for two items (Reading grade 8 and Mathematics grade 5 with correlations of 0.10 and 0.09). Although these items, with correlation coefficients ranging from 0.09 to 0.15, are fairly low, the fact that they are positive indicates that the items are contributing information about student ability. These items therefore were retained in order to meet the WKCE blueprints.

8.1.3 Flagging for *p*-Value

Six items were flagged for *p*-values <0.30 , and all six of these items had *p*-values between 0.21 and 0.28. While these statistics indicate items that were very difficult, the number of items flagged for difficulty was very small. None of the test forms had more than one item flagged for difficulty.

8.1.4 Flagging for Omit Rate

Four items were flagged for omit rates greater than 5%. All of the items flagged for omit rates were highly discriminating items. With the exception of one item in Mathematics grade 10 that had a borderline p -value (0.27), all of the other items flagged for high omit rates had consistently good statistics. All were retained to meet the WKCE blueprints.

8.1.5 Supplemental Tables on Classical Item Analysis

Tables 8-1 through 8-23 present more comprehensive results from the classical item analysis for all of the items retained in each grade and content area. Readers may note that the results presented in these tables may differ slightly from testing results presented on DPI's website due to slight differences in the decision rules defining which students are included or excluded from summary results. Official final results are based on the application of detailed inclusion rules, such as whether the student moved into a school and how long he /she was in one school or another over the course of the year.

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

Table 8-24 summarizes the number of flagged items across grade and content areas. 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.

8.1.6 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 items and all test takers should be able to answer at least 75% of the items (Swineford, 1956). However, a more stringent requirement is often applied, considering tests to be unspeeded only if at least 95% of the examinees attempt the final item. As shown in Table 8-E, WKCE tests satisfy this more stringent requirement, with more than 95% of the examinees attempting the final item in each of the five WKCE content areas.

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

Content	Grade						
	3	4	5	6	7	8	10
Reading	98.19%	97.61%	99.16%	99.13%	98.56%	99.54%	99.31%
Mathematics	98.90%	99.47%	99.24%	99.39%	99.69%	99.17%	99.33%
Language Arts		97.68%				98.16%	97.37%
Social Studies		98.95%				98.96%	99.03%
Science		97.61%				99.25%	99.40%

8.2 Raw Score Results

Raw score results based on all students that took the Fall 2011 WKCE assessment are presented in Table 8-25. In order to facilitate interpretation of the raw score results, Table 8-25 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 observed score, the maximum observed score, reliability (Cronbach’s alpha), and the SEM for raw scores. These measurements are further explained below. Readers can refer to Table 3-1 for a count of the number of items in the test and the number of raw score points corresponding to each item.

The mean raw score should be understood by grade and content area and specifically in the context of the maximum possible score points. In Reading, for example, the maximum possible raw score ranges from 56 to 60, and in Mathematics it ranges from 56 to 62.

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: the 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. In Reading grade 5, the test difficulty statistic (0.66) was obtained by taking the mean raw score of 39.68 and dividing it by 60.

Table 8-25 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 8-25 reveals that in most cases the WKCE 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 WKCE 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 Model Academic Standards.

In addition, Table 8-25 shows the minimum observed score is zero where any student failed all items for each test. The maximum observed score is equal to the maximum number of points possible on the test where any student obtained the full scores for all items. For example, as displayed in Table 8-25, in Reading grade 3, there is at least one student who failed all items and at least one student who obtained a perfect raw score of 60.

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: Reliability. 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. The raw score results are discussed with reference to the total student population and in terms of subgroup comparisons based on gender, race/ethnicity, socioeconomic status, disability status, and English language proficiency. These subgroup comparisons draw from Tables 8-26 through 8-34.

Reading

- Test difficulty ranged from 0.64 to 0.70.
- Standard deviations ranged from 9.88 to 12.14 raw score points.
- Alpha was relatively high in every grade (0.90 to 0.94).
- SEM ranged from 3.03 to 3.19.

Mathematics

- Test difficulty ranged from 0.59 to 0.73, with generally lower difficulty in lower grades and higher difficulty in higher grades.
- Standard deviations ranged from 10.13 to 12.59 raw score points.
- Alpha was relatively high in every grade (0.91 to 0.93).
- SEM ranged from 2.98 to 3.48.

Language Arts

- Test difficulty ranged from 0.61 to 0.74.
- Standard deviations ranged from 5.47 to 7.17 raw score points.
- Alpha ranged from 0.82 to 0.87. As discussed in Part 9, alpha is influenced by test length. All else being equal, shorter tests will tend to have lower reliability than longer tests. The reliability levels are consistent with prior years and are within the expected range given the length of the tests.
- SEM ranged from 2.03 to 2.63.

Social Studies

- Test difficulty ranged from 0.65 to 0.77.
- Standard deviations ranged from 5.89 to 9.34 raw score points.
- Alpha ranged from 0.86 to 0.90. This is consistent with prior years and within the expected range for the length of the tests.
- SEM ranged from 2.24 to 2.96.

Science

- Test difficulty ranged from 0.63 to 0.76.
- Standard deviations ranged from 6.87 to 10.11 raw score points.
- Alpha ranged from 0.87 to 0.91. Alpha was lower in grades 4 and 8 and higher in grade 10. As noted previously, alpha is influenced by test length. Grade 10 has more items than grades 4 and 8 so higher reliability is expected. The alpha levels are consistent with prior years and within expected ranges given the lengths of the tests.
- SEM ranged from 2.37 to 3.02.

Subgroup Performance Patterns in Raw Score Results

Overall, the raw score results show some consistent performance patterns by subgroups, that is, in terms of gender, race/ethnicity, socioeconomic status, disability status, and English language proficiency. Results can be seen in Tables 8-26 through 8-34.

- In Reading, female students, as a group, had a slightly higher mean raw score than male students in each grade level, with differences ranging from 1.41 points in grade 4 to 3.16 points in grade 8.
- In Mathematics, the raw score differences between genders were very small, ranging from 0.13 point in grade 5 to 0.93 point in grade 7. Although in some grades male students showed the higher raw score and in other grades female students showed the higher raw score, small differences like these suggest that the two groups may be best understood as showing similar performance in each grade.
- In Language Arts, female students, as a group, had a slightly higher mean raw score than male students in each grade level, with differences ranging from 1.21 points in grade 4 to 2.51 points in grade 10.
- In Social Studies, the raw score differences between genders were very small, ranging from 0.15 point in grade 10 to 0.28 point in grade 4. Small differences like these suggest that the two groups may be best understood as showing similar performance in each grade.
- In Science, male students had a slightly higher mean raw score than female students in each grade level, with differences ranging from 0.23 point in grade 4 to 1.25 points in grade 10.

In all grades and content areas, the raw score results showed consistent performance patterns by ethnicity. In every grade and content area, White students, as a group, had the highest mean raw score, followed by Asian students, American Indian students, Hispanic students, and African American students. American Indian students had a slightly higher mean raw score than Hispanic students. Differences between the mean raw scores of American Indian and Hispanic students were all equal to or less than 0.42 points in Language Arts, 1.56 points in Social Studies, 1.92 points in Mathematics, and 2.38 points in Science and Reading.

In every grade and content area, the mean raw score was higher among those students who were not economically disadvantaged than among those who were economically disadvantaged. The mean raw score difference between the two groups ranged from 3.72 points in Language Arts grade 4 to 9.48 points in Mathematics grade 10.

There were also differences in mean raw scores between students with disabilities and those without disabilities in all grades and content areas. The mean raw score of students without disabilities was consistently higher than the mean raw score of students with disabilities, with differences ranging from 3.86 points in Social Studies grade 4 to 14.42 points in Mathematics grade 7.

In every grade and content area, students who were fully English proficient consistently showed a markedly higher mean raw score than students who were limited English proficient. As might be expected, these differences were largest in Reading, where fully English proficient students scored 8.16 to 13.38 points higher (in grades 3 and 10, respectively) than students who were limited English proficient. Mean raw score differences ranged from 4.63 to 12.67 points in Mathematics, 3.54 to 7.61 points in Language Arts, 3.62 to 10.34 points in Social Studies, and 4.83 to 11.63 points in Science.

8.3 Summary Statistics for Scale Scores

The WKCE 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 achievement in that content area. Higher scale scores indicate higher levels of achievement, and lower scale scores indicate lower levels of achievement. 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. Results for all students are described, as are results based on gender, race/ethnicity, socioeconomic status, disability status, and English language proficiency. Table 8-36 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 observed scale scores, and LOSS and HOSS for all content areas and grades based on the census data. The LOSS and HOSS, as discussed in Part 7, 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. The results for gender, race/ethnicity, socioeconomic status, disability status, and English language proficiency are drawn from Tables 8-37 through 8-45.

Reading

- Mean scale score increased by grade level, ranging from 457.00 to 548.91.
- Standard deviations ranged from 39.29 to 64.92 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 436.91 to 562.60.
- Standard deviations ranged from 46.39 to 50.59 scale score points.
- In each grade level, student scores spanned the full-scale score range, from the LOSS to the HOSS.

Language Arts

- Mean scale score increased by grade level, ranging from 294.35 to 449.73.
- Standard deviations ranged from 30.12 to 41.29 scale score points.
- In each grade level, student scores spanned the full-scale score range, from the LOSS to the HOSS.

Social Studies

- Mean scale score increased by grade level, ranging from 297.94 to 447.77.
- Standard deviations ranged from 26.91 to 46.48 scale score points.
- In each grade level, student scores spanned the full-scale score range, from the LOSS to the HOSS.

Science

- Mean scale score increased by grade level, ranging from 299.67 to 451.03.
- Standard deviations ranged from 31.95 to 49.80 scale score points.
- In each grade level, student scores spanned the full-scale score range, from the LOSS to the HOSS.

Subgroup Performance Patterns in Scale Score Results

The scale score results, like the raw score results, showed some consistent performance patterns in terms of subgroups (gender, race/ethnicity, socioeconomic status, disability status, and English language proficiency).

Gender

- In terms of gender, male students, as a group, showed a slightly lower mean scale score in Reading than female students in each grade level. The difference ranged from 6.06 to 15.06 scale score points.
- In Mathematics, the differences between genders were very small, from 0.02 scale score point to 3.61 scale score points, with male students scoring slightly higher than female students in grades 3, 4, 8, and 10.
- In Language Arts, female students scored from 6.25 to 14.47 scale score points higher than male students.

- There were only small differences between scale scores by gender in Social Studies, from 0.52 scale score point to 1.01 scale score points, and male and female students alternated between the higher and lower score groups.
- In Science, female students, as a group, showed a slightly lower mean scale score than male students in each grade level. The difference ranged from 0.72 to 5.00 scale score points.

Race/Ethnicity

- The scale score results showed some consistent performance differences by ethnicity.
- In almost every grade and content area, White students, as a group, had the highest mean scale score, followed by Asian students, American Indian students, Hispanic students, and African American students, in that order. The only exceptions to this occurred in Mathematics grade 6 where Hispanic students had a slightly higher mean scale score (difference of 0.03) than American Indian students. In all other grades and content areas, American Indian students had a slightly higher mean scale score than Hispanic 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. In about half of the grades and content areas, differences were less than seven scale score points.

Socioeconomic Status

- Economically disadvantaged students, as a group, consistently scored lower than students who were not economically disadvantaged across all grades and content areas. Differences ranged from 17.11 scale score points in Social Studies grade 4 to 45.90 scale score points in Reading grade 10.
- For every grade and content area, the mean scale score of students who were economically disadvantaged was more than one-half 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 score by group. Differences ranged from 15.76 scale score points in Science grade 4 to 82.11 scale score points in Reading grade 10.
- For every grade and content area, the mean scale score of students with disabilities was more than one-half standard deviation lower than the mean scale score of students without disabilities.

English Language Proficiency

- Students who were fully English proficient and students who were limited English proficient showed consistent and large differences in mean scale score by group. Differences ranged from 15.64 scale score points in Social Studies grade 4 to 79.13 scale score points in Reading grade 10.
- For every grade and content area, the mean scale score of limited English proficient students was more than one-half standard deviation lower than the mean scale score of fully English proficient students. These differences increased to approximately one standard deviation at the upper grade levels.

8.4 Cut Scores and Performance Level Classifications

Student performance on the WKCE is reported in terms of four performance categories: *Minimal*, *Basic*, *Proficient*, and *Advanced*. These performance categories are established through “cut scores.”

Standard 4.19 of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) indicates that “[w]hen proposed score interpretation involves one or more cut scores, the rationale and procedures used for establishing cut scores should be clearly documented” (p. 59). In terms of the validity of the WKCE, it is essential to understand that the cut scores were established in a collaborative, participatory process. The two key activities in that process were *standard setting* and *descriptor writing*. Simply speaking, standard setting is a collaborative process of setting cut scores, and descriptor writing is a collaborative process of establishing a plain-language description of what students must know in order to be classified within each of the performance levels established through cut scores.

Performance level descriptors describe the content-based expectations regarding what Wisconsin students should know and be able to do in each grade/content area. Descriptors and cut scores together define, in qualitative and quantitative terms, the differences between a student who is *Proficient* and a student who is not. The Wisconsin Model Academic Standards guided the standard setting and descriptor writing process. These guided participatory processes served to ensure that the performance levels reported for the WKCE reflect the achievement standards and abilities intended by the Wisconsin legislature, teachers, citizens, and DPI.

CTB performed a special study in which the previous WKCE assessments (those that existed until the Fall 2005 WKCE administration) were linked to the current WKCE assessments (those that began with the Fall 2005 WKCE administration) as an important part of setting the cut scores. For details of the linking study, the standard setting activities, and the descriptor writing process, please refer to the Fall 2005 Technical Report (Part 11) and the Fall 2006 Technical Report (Parts 8 and 12), which can be found in Appendices 3 and 2, respectively of the Fall 2010 WKCE Technical Report available from the DPI at: <http://dpi.state.wi.us/oea/pdf/td-2010-techman.pdf>. Interested readers can also refer to the 2005 Standard Setting Technical Manual, which can be located at <http://dpi.state.wi.us/oea/pdf/td-2005-cutscore.pdf>.

Table 8-46 shows the cut scores for each content and grade level. Tables 8-47 through 8-51 show the percentage of all students in each performance category, as well as subgroup comparisons by gender, race/ethnicity, socioeconomic status, disability status, and English language proficiency. The results for each content area and grade are summarized below. For ease of reference, Tables 8-52 through 8-56 provide the scale score ranges that define performance levels together with the percentage of students in each performance level.

Reading

- In terms of the total student population, most students were either *Proficient* or *Advanced* in Reading. Across grade levels, at least 75% of students were either *Proficient* or *Advanced*.
- Approximately 44% of the total student population was classified as *Advanced* in Reading.
- Across all grade levels, less than 25% of students were below *Proficient*. The difference ranged from 13% below *Proficient* in grade 7 to 21% below *Proficient* in grade 10.

Mathematics

- Across grade levels, over 72% of the student population was either *Proficient* or *Advanced* in Mathematics.
- The proportion of students who were *Advanced* climbed by grade, from 38% in grade 3 to 48% in grade 5, and then declined by grade from 38% in grade 6 to 25% in grade 10.
- In grades 3–8, less than 24% of the student population was classified as below *Proficient*, but in grade 10 approximately 28% of students were classified as below *Proficient*.

Language Arts

- At least 65% of the student population was either *Proficient* or *Advanced* in Language Arts.
- In grades 4 and 10, over 71% of students were either *Proficient* or *Advanced*, and in grade 8, 65% of students were either *Proficient* or *Advanced*.
- In grades 4 and 10, 22% and 29% of students, respectively, were below *Proficient*, but in grade 8, 35% of students were below *Proficient*.

Social Studies

- Most of the total student population was either *Proficient* or *Advanced* in Social Studies. The proportion of *Proficient* or *Advanced* students was 94% in grade 4, 82% in grade 8, and 77% in grade 10.
- A large proportion of students were *Advanced*, especially in grade 4: 68% in grade 4, 45% in grade 8, and 46% in grade 10.
- The proportion of students classified as below *Proficient* was 6% in grade 4, 18% in grade 8, and 23% in grade 10.

Science

- Approximately 77% of students were either *Proficient* or *Advanced* in Science.
- The percentage of students classified as *Advanced* increased from 24% in grade 4 and 34% in grade 8 to 40% in grade 10.
- Approximately 23% of students in each tested grade level were below *Proficient*.

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 8-49 through 8-53.

In terms of gender, a higher percentage of female students, as a group, were classified as *Proficient* or above in Reading than male students. Conversely, there were a higher percentage of male students than female students in the lowest performance level category in Reading. In Mathematics, the percentage of both genders was approximately equal in the *Proficient* reporting category and in the lowest performance category. However, the percentage of male students was slightly higher than the percentage of female students in the *Advanced* category for all grades except grade 7. In Language Arts, there was a markedly higher percentage of female students than male students who were *Proficient* or above and a much lower percentage of female students than male students in the lowest performance category. In Social Studies, the percentage of female students who were *Proficient* or above was slightly higher than the percentage of male students. There were also more male than female students in the lowest performance category. In Science, the percentages of both genders were approximately equal in *Proficient* or above, although in every grade there were higher percentages of female students who were classified as *Proficient* and there were higher percentages of male students who were classified as *Advanced*.

There were some consistent patterns in performance by ethnicity across grades and content areas. First, in terms of the *Proficient* or above category, the prevailing tendency was that there were higher percentages of White students, as a group, to be classified as *Proficient* or *Advanced*, followed by Asian students, American Indian students, Hispanic students, and African American students. The inverse sequence was found at the *Minimal* performance level.

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

Performance level results showed a similar pattern in comparisons of students who were fully English proficient with student 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* and much higher percentages of students who were fully English proficient classified as *Advanced*. There were much lower percentages of fully English proficient students who were classified in the lowest performance category in all grades and content areas.

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

8.5 Standard Performance Indicators 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 weaknesses. This kind of information can be derived from scores on subsets of test items that estimate how much a student knows in a clearly defined skill domain. These skill domains are called content standards (or standards, or objectives). Scores on subsets of test items at the content standard level are called standard performance indicator (SPI) scores. The purpose of reporting SPI scores on the WKCE assessments 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 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. These are criterion-referenced scores, in that they estimate how much a student knows in a clearly defined skill domain (i.e., the criterion). Technical readers can refer to *TerraNova 2nd Edition Technical Report* (CTB/McGraw-Hill, 2000) for details of the estimation procedures for SPI.

This approach, identifying student proficiency on each content standard, relates to the Wisconsin Model Academic Standards. The SPI provides a more reliable estimate of student achievement on each content standard than is possible by simply reporting percent correct. However, *the SPI information should be used for low-stakes purposes because the SPI cannot be considered stable for any content standard with a small number of items.*

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

Tables 8-57 through 8-61 identify the content standards, the number of MC and CR items within each standard, the total number of possible points per standard, 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. Table 8-62 identifies the SPI cut scores for each content area reporting category and grade level. The results from Tables 8-57 through 8-61 are summarized below.

Reading

Table 8-57 presents mean *p*-values and SPI scores for Reading across content standards and grades. The mean of the mean Reading SPI scores across grades and content standards was 67.04%, indicating that the items were moderately difficult for examinees. Results show that the mean *p*-values and SPI scores varied across standards in all grades. Mean SPI scores ranged from 46.15% to 83.22%. In general, the difference between the lowest and highest mean SPI scores was greatest in grade 3 (28.72%). The difference was smallest in grade 4 (9.62%), and content standard 4 (Evaluates/Extends Text) and content standard 3 (Analyzes Text) were the most difficult standard at all grades.

Mathematics

Table 8-58 presents Mathematics *p*-values and SPI scores across grades and content standards. The mean of the mean Mathematics SPI scores across grades and content standards was 66.59%, indicating a moderate degree of difficulty. Results show that the mean *p*-values and SPI scores varied across standards in all grades. Mean SPI scores ranged from 48.05% to 81.85%, with the largest difference observed in grade 4 (where SPI scores ranged from 52.68% to 80.19%). Differences between the highest and lowest mean SPI scores ranged from 9.00% (grade 10) to 27.51% (grade 4). Content standard A (Mathematical Processes) was the most difficult standard in grades 3, 4, 6 and 7. In grades 5, 8 and 10, standard E (Statistics/Probability) showed lower mean SPI scores, indicating relatively difficult items in the standards.

Language Arts

Table 8-59 presents Language Arts *p*-values and SPI scores across grades and content standards. The mean of the mean Language Arts SPI scores across grades and content standards was 64.03%, indicating a moderate degree of difficulty. Mean SPI scores ranged from 56.51% to 77.99%, with differences between the highest and lowest mean SPI scores of 8.81% in grade 4, 12.15% in grade 8, and 2.73% in grade 10. The mean *p*-values and SPI scores indicated that content standard F (Research and Inquiry) was the most difficult standard across all grades.

Social Studies

Table 8-60 presents Social Studies *p*-values and SPI scores across grades and content standards. The mean of the mean Social Studies SPI scores across grades and content standards was 71.09%. While this number is somewhat higher than the mean for the other content areas, this is largely the result of the relatively low difficulty of the grade 4 items, with most of the other grades exhibiting more moderate difficulty. Mean SPI scores ranged from 56.92% to 80.26%, with differences between the highest and lowest mean SPI scores of 7.93% in grade 4, 13.60% in grade 8, and 20.64% in grade 10. The mean *p*-values and SPI scores indicated that the most difficult content standard varied between the three Social Studies grades. In grade 4 the most difficult standard was C (Political Science), in grade 8 the most difficult standard was E (Behavioral Science), and in grade 10 the most difficult standard was B (History).

Science

Table 8-61 presents Science *p*-values and SPI scores across grades and content standards. The mean of the mean Science SPI scores across grades and content standards was 69.77%. The results indicate that the content standards in grade 10 were considerably more difficult than in grades 4 and 8. Across all grades and content standards, mean SPI scores ranged from 55.42% to 83.10%, with differences between the highest and lowest mean SPI scores of 6.23% in grade 4, 17.09% in grade 8, and 11.66% in grade 10. The mean *p*-values and SPI scores indicated that

content standard E (Earth and Space) was the most difficult standard in grades 4 and 8, and content standard D (Physical Science) was the most difficult in grade 10.

Summary of Student Performance Indicator Results

Overall, the mean SPI scores across grades and content standards range in difficulty. There are, however, a few instances of high SPI scores:

- Grades 4 and 5 Reading standard 4 (Evaluates/Extends Text)
- Grades 3, 4, 6, and 8 Mathematics standard A (Mathematical Processes)
- Grade 4 Language Arts standards F (Research and Inquiry) and D (Language)
- Grade 10 Science standards E (Earth and Space) and F (Life and Environment)

The mean SPI scores are consistent with those found in previous years, suggesting that some of the differences in mean SPI scores across content standards may reflect the differential difficulty of the standards themselves and not merely variations in the difficulty of the particular items that were selected for the test forms. Nevertheless, 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. For this reason, the SPI scores should be interpreted with caution and should not be used to make comparisons of student performance across testing years or grade levels.

Summary of Student Achievement Results

In the WKCE, the purpose of the Reading, Mathematics, Language Arts, 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 validity evidence, indicate that the scale scores and performance levels reported in the WKCE program are valid and reliable evidence of student achievement in the tested content areas and grades. As such, these test scores 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 can be drawn upon for accountability and reporting purposes associated with *No Child Left Behind* or school improvement initiatives.

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 Fall 2011 WKCE was estimated in four ways:

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

The present chapter addresses AERA, APA, & NCME (1999) standards 2.1, 2.2, 2.10, 2.11, 2.14, and 2.15.

Standard 2.1 advises providing reliability estimates and the SEM for all total scores and subscores reported, standard 2.2 advises reporting SEM in both raw score and scale score units, and standard 2.11 advises that reliability and SEM should be assessed for all population subgroups. To meet these standards, this chapter of the report presents raw score reliability coefficients and SEMs for the five WKCE content areas and for each reported content standard for the total group of examinees and for subgroups identified by gender, race/ethnicity, socioeconomic status, disability status, and English language proficiency. The scale score conditional SEMs are provided in Section 7.3.1.

Standard 2.15 advises that when testing measures are used to make categorical decisions, the reliability of those decisions should be estimated. In the present context, standard 2.15 applies specifically to performance level determinations, such as who is *Proficient* or *Advanced*. As described below, the Fall 2011 WKCE 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 WKCE program. This analysis also addresses standard 2.14 by providing a conditional SEM for the cut scores that separate the performance levels.

Standard 2.10 advises reporting measures of inter-rater consistency where subjective judgment is involved in scoring. As we saw in Part 5, CR items were scored by human raters; the process thus involved subjective judgment. As this section will show, a detailed assessment of inter-rater consistency was applied to the WKCE. The assessment conducted is termed inter-rater reliability; it measures the reliability of human raters as they score CR items.

Combined, Cronbach’s alpha, SEM, classification consistency, classification accuracy, and inter-rater reliability provide several forms of evidence bearing on the reliability of the WKCE. Cronbach’s alpha and the SEM operate at the content level: they provide estimates of reliability for student scores in Reading or Mathematics, for example. Classification consistency and classification accuracy operate on the associated performance level classifications. These are of particular interest in the context of NCLB and the associated AYP requirements. Inter-rater reliability probes further, looking at individual items and evaluating the reliability of the human raters as they assign scores, item by item.

9.1 Measures of Internal Consistency and SEM

Cronbach’s alpha is a frequently used measure of internal consistency for tests consisting of MC and 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 of the raw score distribution and reliability represents Cronbach’s alpha.

Cronbach’s alpha and the SEM are shown in Tables 9-1 and 9-2, respectively. These tables include information for all students and for the subgroup categories of gender, race/ethnicity, socioeconomic status, disability status, and English language proficiency.

As indicated in Table 9-1, reliability was highest in Reading and Mathematics. Looking at all examinees together in the “Total” column, reliability ranges from 0.90 to 0.94 across grades for Reading, from 0.91 to 0.93 for Mathematics, from 0.82 to 0.87 for Language Arts, from 0.86 to 0.90 for Social Studies, and from 0.87 to 0.91 for Science. Ideally, we would like all reliability coefficients to be 0.90 or above. However, for relatively short tests that are designed to measure a fairly broad range of content, this is not always a realistic expectation. If 0.90 is considered a conservative criterion for an acceptable level of reliability, as measured by Cronbach’s alpha, then the grades 4 and 8 Language Arts, Social Studies, and Science tests 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 30-item tests in Language Arts in grades 4, the 29-item test in grade 8, and the 32-item test in grade 10 would need to be increased in length to 60, 36, and 44 items, respectively; the current 40-item tests in Science grades 4 and 8 and the 50-item test in

grade 10 would need to be increased to 64, 58, and 59 items, respectively; and the current 40-item tests in Social Studies in grades 4 and 8 would need to be increased to 52 items.

Table 9-1 shows that many of the subgroup reliability coefficients were similar to, albeit slightly lower than, the total reliability coefficients. Reliability coefficients are particularly sensitive to the score distribution and variance, so this result is consistent with the generally larger standard deviations (as previously discussed in Part 8 of this report and summarized in Tables 8-27 through 8-35) among many of these subgroups.

The differences in reliability among most subgroups on most tests were quite small. Differences between male and female students were within 0.03 of one another for all grades and content areas.

The difference between disabled and not disabled and economically disadvantaged and not disadvantaged students was within 0.05 of one another. Most differences among the five racial/ethnic groups also were quite small, within 0.03 of one another for all grades and content areas except Mathematics grade 10 and Language Arts grade 10 (where the reliability for White students was 0.05 higher than the reliability for African American students); Mathematics grade 8, Language Arts grade 4, and Science grade 10 (where the reliability for Hispanic students was 0.04 higher than the reliability for African American students); Mathematics grade 10 and Language Arts grade 10 (where the reliability for Hispanic students was 0.06 higher than the reliability for African American students); Language Arts grades 4 and 10 (where the reliability for Asian students was 0.05 and 0.04 higher than the reliability for Hispanic students, respectively); and Language Arts grade 4 (where the reliability for Asian students was 0.06 higher than the reliability for American Indian students). The greatest differences were between fully English proficient and limited English proficient students, with consistently lower reliability among limited English proficient 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 we can expect that an individual's obtained score will fall within two standard errors of his or her true score approximately 95% of the time. Although there were some observable differences in SEM for the different subgroups, all differences were within one-half of a score point. The SEMs for Reading and Mathematics were 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 the Reading and Mathematics tests have more raw score points and larger raw score standard deviations than the other content areas. For every grade and content area, the conditional SEM for individual scale scores are provided in the scoring tables previously discussed in Part 7 (Tables 7-2 through 7-24). The SEM at the *Proficient* cut score was low in all grades and content areas. The SEMs are also plotted in Figures 7-1 through 7-5, with the locations of the cut scores shown in each plot so that the associated SEMs can be easily located.

Reliability, as measured by Cronbach's alpha, was also computed for each content standard within each content area. Table 9-3 shows these reliability coefficients by content standard. The last column presents the reliability for the total content area (with all content standards) for all examinees. It is clear that the reliability per content standard is lower than that

for the total test per content area. As discussed above, 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. As discussed in Part 2 of this report, and summarized in Tables 2-1 through 2-5, the targeted number of items per content standard ranged from 5 to 23 items for Reading,⁵ 6 to 15 items for Mathematics, 5 to 20 items for Language Arts, 5 to 13 items for Social Studies, and 4 to 10 items for Science. A lower level of reliability statistics per content standard is therefore expected. The generally lower level of reliability per standard is one of the reasons why the information based on the content standards should be used for low-stakes purposes only (this issue was previously discussed in the context of SPI).

By content standard, the reliability ranges were as follows:

- For Reading, reliability indices by content standard ranged from 0.52 (for standard 4 in grade 3, with 5 items) to 0.87 (for standard 2 in grade 3, with 17 items).
- For Mathematics, reliability indices by content standard ranged from 0.55 (for standards A and C in grade 5, with 6 and 10 items, respectively) to 0.77 (for standard B in grade 3, with 12 items).
- For Language Arts, reliability indices by content standard ranged from 0.35 (for standard D in grade 4, with 5 items) to 0.82 (for standard B in grade 8, with 18 items).
- For Social Studies, reliability indices by content standard ranged from 0.48 (for standard C in grade 4, with 7 items) to 0.70 (for standard B in grade 8, with 13 items).
- For Science, reliability indices by content standard ranged from 0.33 (for standard D in grade 4, with 6 items) to 0.73 (for standards G/H in grade 10, with 10 items).

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 are 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.

⁵ Note that content standard D at grade 3 contains 5 items but is worth 7 points because it includes four MC items and one 3-point CR item. Therefore, the point values for Reading range from 7 to 25 points.

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 achievement at or above each state’s self-defined *Proficient* category, the consistency and accuracy of the classification of students into these performance categories is of particular interest. The following section describes how the consistency and accuracy of these classifications were evaluated and it provides evidence supporting the validity of these classifications.

Conceptually, classification consistency is defined as the extent to which two classifications of a single student agree, either based on two independent administrations of the same test or 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
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.”

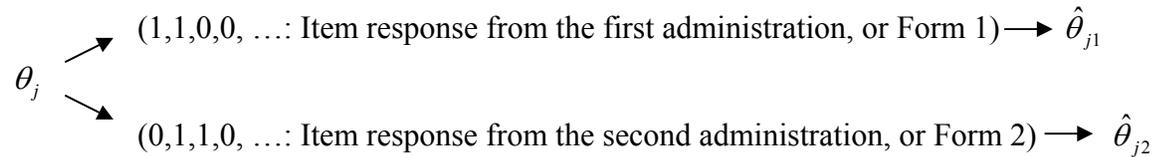
While classification *consistency* refers to the agreement between two observed scores, classification *accuracy* refers to the agreement between the observed score and the true score. Classification accuracy is defined as the extent to which the actual classifications of test takers agree with those 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 corresponding to observed scores. For the WKCE, the method used to estimate classification accuracy and consistency is the Kolen and Kim (2004) method, described in the next section of this report.

9.2.1 Kolen and Kim’s Method for Pattern Scoring

As stated in Part 7, when item response theory (IRT) is applied to score examinees’ responses, two types of scoring are available: number-correct scoring and item-pattern scoring. WKCE uses item-pattern scoring. Many methods of estimating the consistency and accuracy of classification based on number-correct scoring have been suggested in the 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. KKM requires a simulation of item responses as follows:

Step 1: Obtain item parameters (\mathbf{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 using the two ability estimates obtained from Step 2.

Table 9-B
Classification Table for One Cut Point (C_1)⁶

	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 examinees' abilities are estimated at each quadrature point, this quadrature point can be considered the true score. Therefore, classification accuracy is computed using both examinees' estimated abilities (observed scores) and quadrature point (true score). 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.

⁶ This table is constructed for each quadrature point and replication. One, and only one, cell will have a value of one and zeros elsewhere.

As can be seen in Tables 9-5 through 9-27, 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, in Reading grade 3, 0.05 represents the probability of belonging to *Minimal Performance* in both the first and second administrations. The 0.07 represents the probability of belonging to *Proficient* in the first administration and *Advanced* in the second administration. “Sum” is obtained simply by adding the four row values or the four column values. This “Sum” is not always identical to the sum of the values shown in the table because the values displayed have been rounded to two decimal places.

The second table shows indices for classification consistency and classification accuracy. Because there are four performance levels for the WKCE, there are three cut scores. The values in “All Cuts” were obtained by applying all three cuts together. In Table 9-5 for Reading grade 3, when all three cuts were used for the computation, classification consistency (P) is 0.82, chance probability is 0.33, kappa (k) is 0.73, and classification accuracy is 0.87. 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, kappa, 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 AYP 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 0.95, kappa was 0.85, and classification accuracy was 0.97. The interpretation of the values illustrated for Table 9-5 is the same for Tables 9-6 through 9-27.

When only the *Proficient* cut score was applied, P was greater than or equal to 0.90, and kappa was greater than or equal to 0.76 for all Reading and Mathematics tests. For Language Arts, the P associated with the *Proficient* cut was 0.88 for all grades, and the lowest kappa was 0.66. In Social Studies, the lowest P associated with the *Proficient* cut was 0.90, and the lowest kappa was 0.76. For Science, the lowest P was 0.90 and the lowest kappa was 0.73. According to Landis and Koch’s criteria for kappa (presented previously in this report in the discussion of classification consistency), all tests showed excellent agreement based on the cut for the *Proficient* performance level.

Figures 9-1 through 9-5 also show P, kappa, and classification accuracy when students were classified based on “All Cuts.” These values are provided in Tables 9-5 through 9-27, but the results are also provided in the plots for ease of understanding. As can be seen in the plots, all grades and content areas indicated classification consistency (P) based on all cuts over 0.70 for all grades in Reading, Mathematics, Social Studies, and Science. In Language Arts, P was 0.69, 0.68, and 0.72 in grades 4, 8, and 10, respectively. The values of kappa were greater than 0.60

for all grades in Reading and Mathematics. The kappa for Language Arts, Social Studies, and Science were all greater than 0.50. In summary, based on the Landis and Koch criteria all test forms showed good agreement.

9.3 Inter-Rater Reliability for CR Items and Writing Prompts

The reliability of handscoring may be measured in a variety of ways. Two of the most effective ways are 1) tabulations of exact and adjacent agreement and 2) reliability coefficients. Reliability for CR items is typically examined by calculating indices of inter-rater agreement, the degree of reliability with which different human raters assign scores to a given student response. Two indices for inter-rater reliability, intraclass correlation and weighted kappa, are presented here.

Notation. To assess reliability, it is necessary to replicate the scoring process for a subset of papers. This is usually done with “blind double-reads.” Suppose that we have N responses, each of which is scored twice. We denote the two scores of response n by X_{n1} and X_{n2} , where $n=1, 2, \dots, N$. The resulting data may be presented in two ways, enumeration by response and cross-tabulation.

Data Structure 1: Enumeration by Response. Each row represents a single student response:

Response #	Score 1	Score 2	Mean Score
1	X_{11}	X_{12}	$\bar{X}_{1.}$
2	X_{21}	X_{22}	$\bar{X}_{2.}$
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮
N	X_{N1}	X_{N2}	$\bar{X}_{N.}$
Column Mean	$\bar{X}_{.1}$	$\bar{X}_{.2}$	$\bar{X}_{..}$

where

$$\bar{X}_{1.} = (X_{11} + X_{12}) / 2$$

is the mean score for response 1 (similarly for responses 2, 3, ... N),

$$\bar{X}_{.1} = \frac{1}{N} \sum_{n=1}^N X_{n1} = (X_{11} + X_{21} + \dots + X_{N1}) / N$$

is the mean of Score 1 over all responses (similarly for Score 2), and

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

is the overall mean score across both scores of all responses.

Data Structure 2: Cross-Tabulation of Score 1 and Score 2. As an alternative, we may create a square table of counts for each Score 1 by Score 2 (i.e., $X_{n1} \times X_{n2}$) combination:

		Score 2				Row Total
		0	1	...	m	
Score 1	0	n_{00}	n_{01}	...	n_{0m}	n_{0+}
	1	n_{10}	n_{11}	...	n_{1m}	n_{1+}

	m	n_{m0}	n_{m1}	...	n_{mm}	n_{m+}
Column Total		n_{+0}	n_{+1}	...	n_{+m}	n_{++}

where m is the maximum score (for a rubric including zero) obtainable for the item; n_{ij} is the number of responses for which Score 1 = i and Score 2 = j ; n_{i+} is the number of responses for which Score 1 = i , and n_{+j} is the number of responses for which Score 2 = j .

Formulas for the two reliability coefficients of interest are then given:

1. **Intraclass Correlation**, ρ_{IC} , describes the percent 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, $\rho_{IC} = 1$. Always, $0 \leq \rho_{IC} \leq 1$.

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, $k = 0$. Always, $0 \leq k \leq 1$.

Ordinal rating scales (e.g., 0, 1, 2) used in scoring CR 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, but Cohen's kappa does take this into consideration. In general, kappa will have values equal to or smaller than the intraclass correlation. If agreement is perfect, then the value of kappa is 1.0. If agreement is at chance levels, the value of kappa is zero. As noted in Section 9.2.1, 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." Specific criteria for intraclass correlation or weighted kappa are not established.

Tables 9-28 through 9-30 present the rater agreement statistics for CR items and the Writing prompt. The evidence supporting inter-rater reliability is presented in terms of the percentage of agreement between raters, two indices of inter-reliability, and the distributions of scores across score levels. In the table, "Perfect" agreement is defined as scores that are exactly the same. "Adjacent" agreement is defined as scores differing by one point. "Discrepant" cases are those cases where the scores of the two raters differed by more than one raw score point. The column for "Codes" reflects the number of students who received the condition codes A, B, C, or D, which indicate illegible responses, responses that are off-topic, blank responses, or in another language, respectively. "Mean" reflects mean score. "Number of Reads" is exactly two times the number of papers submitted for the purpose of computing inter-rater reliability, as each paper submitted for that purpose is scored twice. The "Frequency" column represents the scoring outcomes for the student responses based on the raw scores given by each of the two raters. For example, as shown in Table 9-28, for Reading grade 4, item 13, the perfect agreement, adjacent agreement, discrepant agreement, and codes are 71%, 25%, 2%, and 2%, respectively.

For Reading, Mathematics, and Writing, student responses were scored by a single rater. To calculate inter-rater reliability, 5% of the responses were scored by a second rater.

The inter-rater reliability results for Reading, Mathematics, and Writing are discussed separately in the following sections. Overall, the results indicate a high degree of reliability for scores on the handscored items in all three content areas.

Reading

Inter-rater reliability results for Reading CR items are shown in Table 9-28. Overall, the rater agreement was very high. The mean percentage of non-discrepant ratings (i.e., perfect agreement plus adjacent scores), averaged across all items, was approximately 95%. As noted in Section 9.2.1, 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.” The mean kappa across all items was approximately 0.77.

Each of the Reading CR items had a maximum possible score of 3. The percentage of discrepant (i.e., nonadjacent) ratings was 3% or less for each of the operational CR items.

The percentages of discrepant ratings for the Reading CR items are summarized below. For these operational CR items, the results were as follows:

- 1% discrepant ratings—3 items (21%)
- 2% discrepant ratings—8 items (57%)
- 3% discrepant ratings—3 items (21%)

The percentage of responses with condition codes ranged from 1% to 8% across all items; the percentage exceeded 4% for only one item (7% of the 14 items). The mean intraclass correlation, averaged across all items, was 0.88. Intraclass correlations ranged from 0.81 to 0.95, and weighted kappa ranged from 0.62 to 0.89.

Mathematics

Table 9-29 provides the inter-rater reliability results for the Mathematics CR items. Overall, the rater agreement was high. The mean percentage of non-discrepant ratings (i.e., perfect agreement plus adjacent scores), averaged across all items, was approximately 97% for operational items. The mean kappa across all items was approximately 0.93.

Treating the two-part CR items as separate items, the maximum possible points per CR item ranges from one to two points. The percentage of discrepant (i.e., nonadjacent) ratings was 4% or less for 45 of the 46 operational CR items. The only exception occurred in Mathematics grade 3 where there was a 6% discrepant rating.

The percentages of discrepant ratings for the Mathematics CR items are summarized below. For these operational CR items, the results were as follows:

- No discrepant ratings—28 items (61%)
- 1% discrepant ratings—11 items (24%)
- 2% discrepant ratings—5 items (11%)
- 3% discrepant ratings—1 item (2%)
- 6% discrepant ratings—1 item (2%)

The percentage of responses with condition codes ranged from 1% to 10% across all items; the percentage exceeded 4% for only 3 items (7% of the 46 items). The mean intraclass correlation, averaged across all items, was 0.97. Intraclass correlations ranged from 0.90 to 1.0, and weighted kappa ranged from 0.81 to 0.99.

Writing

Table 9-30 shows inter-rater reliability results for the Writing prompts. As indicated previously, the Writing prompts were scored on two rubrics, the Composing Rubric (six points) and the Conventions Rubric (three points). Table 9-30 shows that the rate of perfect agreement was lower on the 6-point Composing Rubric than on the 3-point Conventions Rubric. The difference is due to the difference in score points. Perfect agreement is, as discussed previously, less likely with a higher number of possible score points than with a lower number of possible score points. Adjacent and discrepant modes of agreement were, as may also be expected, more common where there were more possible score points. Perfect agreement ranged from 58% to 64% on the Composing Rubric and from 74% to 92% on the Conventions Rubric. Adjacent agreement ranged from 33% to 37% on the Composing Rubric and from 6% to 23% on the Conventions Rubric. The percentage of discrepant (i.e., nonadjacent) ratings for the Writing prompt in grades 4, 8, and 10 ranged from 2% to 3% for the Composing Rubric and was 0% for the Conventions Rubric. Codes were generated in 1% to 3% of the cases. Intraclass correlation ranged from 0.78 to 0.92, and weighted kappa ranged from 0.56 to 0.83.

Summary

Overall, the analyses discussed in this section of the report indicate acceptable levels of reliability for the WKCE assessments. The internal consistency reliability estimates, as measured by Cronbach's alpha coefficient, are 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 low, with acceptably high values for the weighted kappa and intraclass correlations. Finally, the results of the inter-rater reliability analyses indicate a high degree of reliability for scores on the hand-scored items in the WKCE Reading, Mathematics, and Writing assessments.

Part 10: Validity

The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) defines validity as “the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests” (p. 9). The purpose of test score validation is not to validate the test itself, but to validate interpretations of the test scores for particular purposes or actions. Test score validation is not a quantifiable property but an ongoing process, beginning at initial conceptualization and continuing throughout the entire assessment process. Every aspect of an assessment provides evidence in support of (or a challenge to) its validity, including design, content specifications, item development, psychometric quality, and inferences made from the results.

As the Technical Report has progressed, chapter by chapter, it has moved through the phases of the testing cycle. Each part of the Technical Report detailed the procedures and processes applied in the WKCE, as well as their results. Each part also highlighted the meaning and significance of the procedures, processes, and results in terms of validity or a relationship to the *Standards*. Part 10 addresses three final issues in validity: the issues of bias, construct validity, 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 described the involvement of Wisconsin educators, DPI, and CTB in the test development process. As indicated in Part 2, the test development process and the involvement of Wisconsin educators in that process formed an important part of the validity of the entire WKCE. The knowledge, expertise, and professional judgment offered by Wisconsin educators ultimately ensured that the content of the WKCE formed 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 the Technical Report addressed the issue of test form development. Part 3 provided a general discussion of CTB’s test book creation and editing process, the process of selecting operational test items, and the process of obtaining DPI approvals. The test design process and the participation of Wisconsin educators in the process of test selection, including item content and bias reviews, provide a solid rationale for having confidence in the content and design of the WKCE and using it as a tool from which to derive valid inferences about Wisconsin student performance. Parts 2 and 3 together provided evidence to support the content validity of the WKCE and addressed AERA/APA/NCME (1999) standards 1.2, 1.6, 3.1, 3.2, 3.3, 3.5, 3.6, 3.7, 3.9, 3.11, 3.16, 6.4, 6.15, 7.3, 7.4, 7.7, 13.3, and 13.5.

Part 4 of the Technical Report described the process, procedures, and policies that guided the administration of the WKCE, including accommodations, security, and the written procedures provided to test administrators and school personnel. The following AERA, APA, & NCME (1999) standards were addressed: 1.13, 3.3, 3.19, 3.20, 3.21, 5.1, 5.2, 5.3, 5.4, 5.5, 5.6, 5.7, 6.11, 6.15, 9.1, 10.1, and 10.2. The process, procedures, and policies detailed in that section contributed to the validity of the WKCE assessments by reducing the impact of construct-

irrelevant variables (such as non-standardized administration methods, limitations associated with student disabilities, security breaches, etc.) on test performance.

Part 5 of the Technical Report demonstrated adherence to AERA/APA/NCME AERA, APA, & NCME (1999) standards 3.22, 3.23, 3.24, 5.8, and 5.9. It described how MC items and CR items were scored, the handscoring process, the training and selection of readers, the scoring rubrics used for scoring CR items and the resulting score distributions. The procedures described in that section contributed to the validity of the WKCE assessments by preventing hardware- or software-related errors in machine scoring and reducing construct-irrelevant score variance associated with variations in raters' interpretation and application of scoring rubrics.

Part 6 described the sample data used for calibration and scaling, referring the reader to information found in the 2009 WKCE Technical Report.

Part 7 of the Technical Report described the calibration and equating methods, as well as processes and procedures for deriving scale scores from response patterns. Some references to introductory and advanced discussions of IRT were provided. Several axes upon which to evaluate the calibration and equating 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, were all discussed. Part 7 of this report thereby addressed AERA, APA, & NCME (1999) standards 1.13, 4.1, 4.2, 4.3, 4.10, 4.11, 7.1, 7.2, and 7.10. These processes and procedures contributed to the validity of the WKCE by providing the opportunity to identify and eliminate items that were not contributing to the accurate and reliable measurement of the intended constructs and by ensuring that valid comparisons of the WKCE test scores can be made within and across years.

Part 8 presented 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 NCLB policy environment, as well as interests of parents, students, and educators. The SPI scores then probed further, assessing specific skills and abilities. Combined, scale scores, performance levels, and SPI scores provided a comprehensive set of tools to assess Wisconsin student performance by content and grade level and by gender, race/ethnicity, socioeconomic status, disability status, and English language proficiency. Part 8 thus addressed AERA, APA, & NCME (1999) standards 1.5, 3.18, 4.3, 4.5, 4.6, 4.7, 4.19, 7.1, 7.10, 13.15, and 13.19. The analyses addressed in Part 8 contributed to the validity of the WKCE by providing further opportunity to identify and eliminate items that were not contributing to the accurate and reliable measurement of the intended constructs.

Part 9 demonstrated adherence to AERA, APA, & NCME (1999) standards through several analyses of the reliability of the Fall 2011 WKCE. It presented 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. The Fall 2011 WKCE Technical Report thereby addressed AERA, APA, & NCME (1999) standards 2.1, 2.2, 2.10, 2.11, 2.14, and 2.15. Reliability is a prerequisite to score validity, and the analyses in that section

contributed to the WKCE validity evidence by establishing the reliability of the WKCE test scores and proficiency classifications.

In the subsequent pages, Part 10 will, as stated, present additional metrics with which to evaluate the validity of the WKCE program. As described below, the WKCE program formally assessed the issue of test bias through an analysis of differential item functioning (DIF). It is possible for items to function differently among 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 if that possibility occurred and to what degree, item by item, for each of the categories of gender, race/ethnicity, socioeconomic status, disability status, and English language proficiency. This analysis specifically addresses standards 7.1, 7.2, and 7.3.

The present chapter also provides estimations of construct validity. Two measures are provided: correlations among 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 reading ability or mathematics ability. The presence of a single, underlying trait is a fundamental issue when scaling and analyzing results through IRT models. As such, these analyses are essential elements in assessing the validity of the WKCE. Finally, this chapter outlines the erasure analysis procedures that were employed to ensure the integrity of test scores by identifying test papers that may have been fraudulently altered.

10.1 Differential Item Functioning

An empirical differential item functioning (DIF) approach was used to examine potential item bias and to determine if 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, 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.

DIF analyses were conducted based on gender, race/ethnicity, socioeconomic status, disability status, and English language proficiency groups. For the DIF analysis by gender, the reference group is male, meaning that the results for female students are considered with reference to male student performance. In the DIF analysis for race/ethnicity, the reference group is White. This means that the performance of students of each race/ethnicity is considered with reference to the performance of White students. The DIF analysis on socioeconomic status defines students identified as not economically disadvantaged as the reference group, and students identified as economically disadvantaged as the focal group. The DIF analysis for disability status uses students identified as not disabled as a reference group to assess DIF within the student population identified as disabled. The DIF analysis for ELP compares item functioning among students identified as fully English proficient to those identified as limited English proficient. Students identified as fully English proficient are the reference group, and those identified as limited English proficient are the focal group.

Three kinds of DIF statistics were used: Linn-Harnisch, Mantel (or Mantel-Haenszel), and standardized mean difference. Each of these DIF methods can be used to determine if identified groups of examinees with the same underlying level of ability had the same probability of correctly responding to the item. The Mantel-Haenszel method is applied to MC items only. The Linn-Harnisch method is used for both MC and CR items. The Mantel statistic and standardized mean difference are applied to CR items. These DIF statistics and the flagging criteria are described in detail below.

(1) Linn-Harnisch (L-H)

Because the WKCE was built using item response theory (IRT), an appropriate procedure for examining item bias should reflect the IRT model. Several IRT-based procedures are available, such as a procedure that tests the equality of item parameters across groups (Lord, 1980) or any of the procedures that assess the differences in the area between the item characteristic curves (e.g., Linn, Levine, Hastings, & Wardrop, 1981). However, these procedures require a minimum of 800 to 1,000 cases in each group to make reliable comparisons. A procedure that still relies on the predictions of the three-parameter model but does not require as many cases has been suggested by Linn and Harnisch (1981).

To take an example, in the case of gender DIF analyses, item parameters (e.g., discrimination, location, and guessing) and the scale score (θ) for each examinee were estimated using the three-parameter logistic model for MC items and the two-parameter partial credit model for CR items. The sample was then divided into male and female gender subgroups. The members in each group were sorted into ten equal score categories (deciles) based upon their location in the scale score (θ) range. The expected proportion correct for each group based on the model prediction was compared to the observed (actual) proportion correct obtained by the group. The proportion of students in decile g who are expected to answer item i correctly is:

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

where n_g is the number of examinees in decile g . To compute the proportion of students expected to answer item i correctly (over all deciles) for a specific subgroup, the following statistic was computed

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

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

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

where u_{ij} is the dichotomous score for item i for examinee j . The corresponding formula to compute the observed proportion answering each item correctly (over all deciles) for a subgroup is given by

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

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

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

and the overall group difference (D_i) between the observed and expected proportions correct for item i in the complete group (over all deciles) is

$$D_i = O_i - P_i.$$

These indices are indicators of the degree to which subgroup members performed better or worse than expected on each item based on the parameter estimates from all subgroups. Differences for decile groups provide an index for each of the ten regions on the scale score (θ) range. The decile group difference (D_{ig}) can be either positive or negative. Use of the decile group differences as well as the overall group difference allows one to detect items that give a large positive difference in one range of θ and a large negative difference in another range of θ , yet have a small overall difference.

DIF is defined in terms of the decile group and total target subsample differences, the D_{i-} (sum of the negative group differences) and D_{i+} (sum of the positive group differences) values, and the corresponding standardized difference score for the subsample (Linn & Harnisch, 1981, p. 112). The standardized difference score (Z_{ig}) for ability group g is computed as

$$Z_{ig} = \frac{1}{n_g} \sum_{j \in g} \left[\frac{U_{ij} - P_{ij}}{\sqrt{P_{ij}(1 - P_{ij})}} \right],$$

where $U_{ij} = 1$ when student j answers item i correctly, and $U_{ij} = 0$ otherwise. The standardized difference over all the ability groups is

$$Z_i = \frac{\sum_g n_g Z_{ig}}{\sqrt{\sum_g n_g^2}}$$

Items for which $|D_i| \geq 0.10$ and $|Z_i| \geq 2.58$ are flagged for DIF. If D_i is positive, the item is biased in favor of the focal group. If D_i is negative, the item is biased against the focal group.

(2) Mantel and Mantel-Haenszel (M-H)

The Mantel (1963) and Mantel-Haenszel (1959) chi-square statistics are used to evaluate potential bias in individual items by examining item-level differences between different groups of students (e.g., students classified by gender, ethnicity, disability, or other variables of interest), controlling for differences in the relevant ability or abilities measured by the test. In this procedure, subgroups are matched by their raw total test score using a contingency table with K levels. The Mantel statistic is computed by first dividing students into K levels of ability on the total test, then comparing the performance of these matched groups using the formula

$$\text{Mantel } \chi^2 = \frac{(\sum_k F_k - \sum_k E(F_k))^2}{\sum_k \text{Var}(F_k)},$$

where F_k is the sum of scores for the focal group at the k th level of the matching variable, and $E(F_k)$ is the expected sum of scores for the focal group at the k th level of the matching variable.

For dichotomous items, the Mantel statistic is equivalent to the Mantel-Haenszel statistic without the continuity correction (Zwick, Donoghue, & Grima, 1993). With the continuity correction added (Holland & Thayer, 1986), the Mantel-Haenszel statistic has the form

$$\text{Mantel - Haenszel } \chi^2 = \frac{(|\sum_k F_k - \sum_k E(F_k)| - 1/2)^2}{\sum_k \text{Var}(F_k)},$$

with all terms defined as in the prior equation.

In addition to the Mantel-Haenszel chi-square statistic, the delta statistic (Δ_{MH}) was computed for all MC items (Holland & Thayer, 1985). To compute delta, the odds ratio α is first computed as

$$\alpha_{MH} = \frac{\sum_{k=1}^K N_{r1k}N_{f0k} / N_k}{\sum_{k=1}^K N_{f1k}N_{r0k} / N_k},$$

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 .

The Δ_{MH} statistic is then computed as

$$\Delta_{MH} = -2.35 \ln(\alpha_{MH}).$$

Positive values of Δ_{MH} indicate items that favor the focal group, whereas negative values of Δ_{MH} indicate items that favor the reference group. WKCE MC items were flagged for DIF using the following criteria (Zwick, Donoghue, & Grima, 1993):

- A= No DIF: Non-significant Mantel-Haenszel χ^2 or $|\Delta_{MH}| < 1.0$
- B= Weak to moderate DIF: Mantel-Haenszel χ^2 is significantly greater than zero ($p < 0.05$) and $1.0 < |\Delta_{MH}| < 1.5$
- C= Large DIF: Mantel-Haenszel χ^2 is significantly greater than zero ($p < 0.05$) and $|\Delta_{MH}|$ exceeds 1.5.

For CR items, an effect size (ES) statistic based on the Mantel Haenszel χ^2 was used. ES is obtained by dividing the standardized mean difference (SMD) statistics by the standard deviation of the item (detailed description of these procedures can be found in Zwick, et al., 1993). WKCE items are flagged using the same rules that are used in The National Assessment of Educational Progress (NAEP):

- No DIF: Non-significant Mantel χ^2 or $|ES| < 0.17$
- Weak to moderate DIF: Mantel χ^2 is significant ($p < 0.05$) and $0.17 \leq |ES| < 0.25$
- Large DIF: Mantel χ^2 is significant ($p < 0.05$) and $|ES| \geq 0.25$

A positive DIF value indicates that the item favors the focal group, while a negative value indicates that the item disadvantages the focal group.

(3) Standardized Mean Difference (SMD)

A standardized mean difference statistic (SMD) was also computed for CR items. The SMD is an effect size index of DIF which is relatively easy to interpret (Zwick, et al., 1993). The SMD compares the means of the reference and focal groups, adjusting for the distribution of reference and focal group members on the conditioning (i.e., matching) variable (Zwick, et al., 1993). SMD is computed as (Zwick, et al., 1993)

$$ES \ SMD = p_{Fk} \left(\sum_k m_{Fk} - \sum_k m_{Rk} \right),$$

where

p_{Fk} = proportion of the focal group members at the k th level of the matching variable,
 $m_{Fk} = 1/N_{F1k}$, where N_{F1k} is the number of correct responses in the focal group at ability level k , and
 $m_{Rk} = 1/N_{R1k}$, where N_{R1k} is the number of correct responses in the reference group at ability level k .

A negative SMD value indicates an item on which the focal group has a lower mean than the reference group. A positive SMD value indicates an item on which the reference group has a lower mean than the focal group. An item is flagged when

$$|ES - SMD| \geq 0.25.$$

Results

Tables 10-1 through 10-7 show items flagged based on the criteria described previously. Readers may note that some items are flagged by both Linn-Harnisch and Mantel-Haenszel methods and some only by one of the methods. For the Linn-Harnisch, Mantel, and Mantel-Haenszel methods, the summary flag information in the DIF tables is always expressed with reference to the focal group. That means that negative flags (such as -B or -C, as described above) indicate that an item disadvantages the focal group, such as female students, African American students, or economically disadvantaged students. A positive flag indicates that the item favors the focal group. The B flag represents a lower threshold for DIF. Only items that were flagged with a C flag were included in the tables described below. Readers can see

B-flagged items in the tables, but that occurs because those items were also flagged with a C flag.

The DIF results for gender are presented in Table 10-1; results for race/ethnicity are presented in Tables 10-2 through 10-5; English language proficiency (ELP) results are presented in Table 10-6; and results based on disability status are presented in Table 10-7. No items were flagged for DIF for socioeconomic status.

Each DIF table references the grade and content area of the items flagged for DIF, as well as the test form, the item number, and the item type. The tables present Linn-Harnisch statistics (D+, D-, and Z) first, then the SMD, and finally the Mantel or Mantel-Haenszel statistic (Δ_{MH}). MH is only computed for the focal group. After specifying these statistics for each item, two final columns provide a summary flag status. There is a column “LH Flag” to indicate where any of the Linn-Harnisch statistics produced a flag and a “MH Flag” column to indicate where either Δ_{MH} or the SMD produced a flag.

In Table 10-1, looking at all items and all grades and content areas, 8 items were flagged for gender DIF. More items were flagged in the Reading test than in the other content area. This is expected given that more grades are tested in Reading. The number of flagged items in Reading relative to the other content areas should be understood in this context. Note that three of the five items flagged by Linn-Harnisch indicate that the DIF favors (rather than disadvantages) female students.

The other DIF results in Tables 10-2 through 10-7 can be understood in the same fashion. Note that a single item can be flagged for multiple subgroup categories, such as for African American students, Hispanic students, and economically disadvantaged students. Readers should also note that Linn-Harnisch DIF statistics cannot be computed unless the sample sizes are at least 50, with at least five students per group in each decile. In some cases (as is noted in the DIF table for American Indian students) the size of the tested population was too small to include valid Linn-Harnisch DIF statistics. DIF results for focal groups containing fewer than 100 students may be unstable and should be interpreted with caution.

The Fall 2011 WKCE tests were developed using procedures to minimize item and test bias. Expertise in this area is not, however, a substitute for statistical analyses of the items. 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 WKCE. However, in large-scale assessments, such as the WKCE, it is expected that some items will show DIF. All of the items in the Fall 2011 WKCE flagged for DIF were notated as such in the classical item analyses and in the item pool so that content experts would be able to reevaluate these items in future item selection activities. Items with DIF (particularly items flagged for strong DIF) are to be avoided in future selections.

10.2 Construct Validity

Construct validity can be defined as the extent to which tests measure the skills or constructs they intend to measure, and it is the central concept underlying the Fall 2011 WKCE assessment validation process. Evidence for construct validity is comprehensive and integrates evidence from both content- and criterion-related validity. The WKCE test development process included specifications, item writing, review, and test construction.

Threats to construct 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 WKCE blueprint. A test can be said to be unidimensional when all of the items in the test measure the same underlying ability or trait.

Analyses of the internal structure of a test can indicate the extent to which the relationships among 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 among content standards within a test can be expected to be relatively high. Tables 10-8 through 10-12 show the correlations among content standards for each WKCE content area. The correlation coefficients here reflect the degree of linear relationship and direction between any two given content standards. The correlation can range from +1 to -1. A correlation of +1 indicates a perfect positive linear relationship, and a correlation of -1 indicates a perfect negative linear relationship between two content standards. A correlation of zero means there is no linear relationship. In general, the size of the correlation coefficient is influenced by the number of items or score points and by the score variance. Readers are cautioned not to confuse correlation with causation. The presence of a high correlation between two content standards should not be taken as an indication that there is a causal relationship between them.

As may be observed in Tables 10-8 through 10-12, correlations among content standards were generally higher in Reading than in the other content areas. The correlations among content standards ranged from 0.55 to 0.84 in Reading, from 0.52 to 0.75 in Mathematics, from 0.39 to 0.73 in Language Arts, from 0.49 to 0.66 in Social Studies, and from 0.40 to 0.70 in Science. 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 among 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.

WKCE test items are calibrated using unidimensional IRT models, which posit that the test items are measuring an essentially unidimensional construct. To assess the dimensionality of the WKCE assessments, a principal components analysis was conducted for each content area and grade. Principal components analysis is a statistical technique commonly used to evaluate dimensionality by detecting patterns of relationships among items. This method is useful in determining whether the observed scores on a test can be explained largely or entirely in terms of a much smaller number of components. To take an example, if answering the mathematics items

in a mathematics test required a lot of reading ability, the mathematics test would not be only a measure of mathematics ability, it would be a measure of reading ability as well. Such a test would be said to be multidimensional rather than essentially unidimensional. One way of evaluating the dimensions detected in the analysis is by examining the eigenvectors and eigenvalues. In principal components analysis, the eigenvectors correspond to factors, and the eigenvalues correspond to the variance explained by these factors. The sum of the eigenvalues is equal to the number of items in the test. The eigenvalues can be ordered from first to last in terms of the amount of the common variance that each explains. Data are generally considered to be unidimensional if the second eigenvalue is less than or equal to 1.0. Previous research shows that the examination of the ratio of the first two (i.e., the two largest) eigenvalues can be useful in determining the existence of dominant factors. Specifically, where large ratios exist between the first and second eigenvalues, a single dominant factor can be said to exist. Although the definition of large in the present context is subjective, the results in Table 10-13 show that the eigenvalue of the first factor, in almost every case, is at least five times as large as the eigenvalue of the second factor.

As may be seen in Table 10-13, the ratios of the first two eigenvalues range from 3.80 to 8.08. The eigenvalues are proportional to the amount of common variance explained by each component, so these ratios indicate that the variance explained by the first component alone is approximately 4 to 8 times greater than the variance explained by the second component. The eigenvalue ratios ranged from 5.96 to 7.73 in Reading, from 6.12 to 7.38 in Mathematics, from 3.80 to 5.68 in Language Arts, from 4.98 to 8.08 in Social Studies, and from 4.69 to 5.52 in Science. These ratios suggest that the unidimensionality of each of the WKCE 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 WKCE assessments. The correlations among 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 Test Integrity: Erasure Analysis

The Fall 2011 WKCE test results were subjected to a special program that analyzed erasures on MC items. The focus of the analysis was on those cases where an incorrect answer choice was erased and replaced with the correct choice. A high rate of erasures can identify situations in which test integrity needs to be examined further. Separate erasure analyses were performed by grade and content area within schools. A summary erasure report was provided to DPI for evaluation.

10.4 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, or other distractions), test administrator practices that may exacerbate test anxiety in particular students, practices that elicit test wiseness, and security breaches that may result in the exposure of test forms or items. Construct-irrelevant variance may exert a systematic effect on the scores of individual students or groups of students, resulting in an overestimation or underestimation of their true ability.

The standardized WKCE 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.

Part 11: Summary Recommendations

Results and key findings of the Fall 2011 WKCE test administration are presented throughout the body of this report. Test difficulty in comparison to the student ability may warrant further attention in subsequent administrations as explained below.

Table 8-25 reveals that in most grades and content areas the WKCE students are not normally distributed along the raw score scale. This is indicated by the negative values of the skewness statistics, and this occurs because many students are answering most of the test items correctly. From a criterion-referenced testing perspective, the clustering of student scores on the high end of the raw score scale indicates that students on the whole are tending to demonstrate the knowledge, skills, and abilities specified in the Wisconsin Model Academic Standards.

From a measurement perspective, the WKCE may provide limited growth information for students in the highest performance level because large numbers of students are scoring in regions of the scale with the most amount of error. To measure these students with precision more difficult items need to be added to the test. That is, for these students the test serves as a general measure of student skill; however, DPI would expect to see less fluctuation in score for individual students in this highest performing group from year to year if more difficult items were added to the assessment.

For these reasons, the DPI may wish to consider increasing the difficulty of the WKCE tests. This will likely provide more specific information about the higher ability students and allow the opportunity for the students to show growth. Because equating requires that tests maintain a similar level of difficulty from year to year, increasing the test rigor would likely require a cut score review and an examination regarding whether or not a new test scale should be set.

Fall 2011 WKCE Technical Report: 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. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association, Inc.
- Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37, 29–51.
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika*, 46, 443–459.
- Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, 41, 687–699.
- Burket, G. R. (1991). *PARDUX* [Computer program]. Unpublished.
- Burket, G. R. (2000). *ITEMWIN* [Computer program]. Unpublished.
- 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/Thompson Learning.
- Cronbach, L. J., Schönemann, P., & McKie, D. (1965). Alpha coefficients for stratified-parallel tests. *Educational and Psychological Measurement*, 25, 291–312.
- CTB/McGraw-Hill. (1997). *TerraNova* 1st Edition. Monterey, CA: Author.
- CTB/McGraw-Hill. (2000). *TerraNova* 2nd Edition. Monterey, CA: Author.
- CTB/McGraw-Hill. (1997–2001). WKCE. Monterey, CA: Author.
- CTB/McGraw-Hill. (2005). *Wisconsin Knowledge and Concepts Examinations 2005 Standard Setting Technical Manual*. Monterey, CA: Author.
- CTB/McGraw-Hill. (2006). *Wisconsin Knowledge and Concepts Examinations Fall 2005 WKCE-CRT Technical Manual*. Monterey, CA: Author.
- CTB/McGraw-Hill. (2007). *Wisconsin Knowledge and Concepts Examinations Fall 2006 WKCE Technical Manual*. Monterey, CA: Author.

- CTB/McGraw-Hill. (2008). *Wisconsin Knowledge and Concepts Examinations Fall 2007 WKCE Technical Manual*. Monterey, CA: Author.
- 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.
- Hambleton, R. K., & Novick, M. R. (1973). Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement*, 10, 159–96.
- Holland, P. W., & Thayer, D. T. (1985). *An alternative definition of the ETS delta scale of item difficulty*. Princeton, NJ: Educational Testing Service, Research Report RR-85-43.
- Holland, P. W., & Thayer, D. T. (1986). Differential item performance and the Mantel-Haenszel procedure. Paper presented at the American Educational Research Association Annual Meeting, San Francisco, California, April 1986.
- Kolen, M., & Kim, D. (2004). Personal Correspondence.
- Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and Practices*. New York: Springer-Verlag.
- Linn, R. L. (1989). *Educational measurement* (3rd ed). New York: Macmillan.
- Linn, R. L., & Harnisch, D. (1981). Interactions between item content and group membership in achievement test items. *Journal of Educational Measurement*, 18, 109–118.
- Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement*, 5, 159–173.
- Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, 32, 179–197.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems* (71, 179–181). Hillsdale, NJ: Lawrence Erlbaum.
- 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, 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.
- Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement*, 14, 59–71.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159–176.
- Muraki, E., & Bock, R. D. (1991). *PARSCALE: Parameter Scaling of Rating Data* [Computer program]. Chicago, IL: Scientific Software, Inc.
- Patz, R. (1998). *Calculating Handscoring Reliability Coefficients*. Unpublished manuscript.
- Stocking, M.L., & Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201–210.
- Swaminathan, H., Hambleton, R. K., & Algina, J. (1974). Reliability of criterion referenced tests: A decision theoretic formulation. *Journal of Educational Measurement*, 11, 263–268.
- Swineford, F. (1956). Technical manual for users of test analysis. *Statistical Report 56–42*. Princeton, NJ: Educational Testing Service.
- Thissen, D. (1982). Marginal maximum-likelihood estimation for the one-parameter logistic model. *Psychometrika*, 47, 175–186.
- Thissen, D. (1990). *MULTILOG: Multiple categorical item analysis and test scoring*. Version 6 [Computer program]. Mooresville, IN: Scientific Software.
- van der Linden, W.J., & Hambleton, R. (1997). *Handbook of modern item response theory*. New York: Springer.
- Wright, B. D., & Linacre, J. M. (1992). *BIGSTEPS Rasch Analysis* [Computer program]. Chicago, IL: MESA Press.
- Yen, W. M. (1981). Using Simulation results to choose a latent trait model. *Applied Psychological Measurement*, 5, 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, 93–111.

- Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 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*, 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*, 209–228.
- Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*, 233–251.

Fall 2011 Tables and Figures

Table 2-1
Target Reading Test Blueprint: Grades 3–8, 10*

	Category Title	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 10	
		MC	CR	MC	CR										
1	Determines meaning of words or phrases in context	12	0	12	0	11	0	10	0	10	0	10	0	7	0
1.1	Uses context clues to determine meaning of words or phrases	8		8		8		7		7		7		5	
1.2	Uses knowledge of word structure to determine meaning of words	2		3		2		2		2		2		1	
1.3	Uses word reference materials to determine meaning of words and phrases	2		1		1		1		1		1		1	
2	Understands Text	20	0	19	0	17	0	15	0	14	0	14	0	7	0
2.1	Demonstrates understanding of literal meaning by identifying stated information in literary text	9		9		7		7		6		6		2	
2.2	Demonstrates understanding of literal meaning by identifying stated information in informational text	9		8		7		6		6		6		3	
2.3	Demonstrates understanding of explicitly stated sequence of events in literary and informational text	2		2		3		2		2		2		2	
3	Analyzes Text	18	1	18	1	18	1	18	1	19	1	19	1	22	1
3.1	Analyzes literary text	8		8		8		8		8		8		7	
3.2	Analyzes informational text	8		8		8		8		8		8		7	
3.3	Analyzes author's use of language in literary and informational text	2		2		2		2		3		3		8	
4	Evaluates and Extends Text	4	1	5	1	8	1	11	1	11	1	11	1	14	1
4.1	Evaluates and extends literary text	2		2		3		4		4		4		4	
4.2	Evaluates and extends informational text	1		2		3		4		4		4		5	
4.3	Evaluates and extends author's use of language in literary and informational text	1		1		2		3		3		3		5	
	Number of Items	54	2	54	2	54	2	54	2	54	2	54	2	50	2
	Total Score Points for Test	60		60		60		60		60		60		56	

*Note: The CR items do not report out to any single subskill.

Table 2-2
Target Mathematics Test Blueprint: Grades 3–8, 10*

	Category Title	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 10	
		MC	CR	MC	CR										
A	Mathematical Processes	3	3	3	3	3	3	3	3	3	3	3	3	7	1
Aa	Reasoning														
Ab	Communication														
Ac	Connections														
Ad	Representation														
Ae	Problem Solving														
B	Number Operations and Relationships	11	1	11	0	11	0	12	0	12	0	7	0	7	0
Ba	Number Concepts	6		5		5		6		6		4		4	
Bb	Number Computation	5		6		6		6		6		3		3	
C	Geometry	9	1	8	1	9	1	9	1	10	2	8	1	8	1
Ca	Describing Figures	4		3		3		2		3		2		4	
Cb	Spatial Relationships and Transformations	4		4		4		4		4		4		2	
Cc	Coordinate System	1		1		2		3		3		2		2	
D	Measurement	8	0	8	1	9	1	9	1	9	0	11	1	9	1
Da	Measurable Attributes	3		3		4		2		3		2		1	
Db	Direct Measurement	4		4		3		3		3		3		2	
Dc	Indirect Measurement	1		1		2		4		3		6		6	
E	Statistics and Probability	7	1	7	1	9	1	8	1	8	1	8	1	9	0
Ea	Data Analysis and Statistics	5		4		6		5		5		5		4	
Eb	Probability	2		3		3		3		3		3		5	
F	Algebraic Relationships	8	1	9	1	10	1	10	1	9	1	14	1	10	1
Fa	Patterns, Relations, and Functions	4		5		5		5		2		7		5	
Fb	Expressions, Equations, and Inequalities	2		2		3		2		3		6		4	
Fc	Properties	2		2		2		3		4		1		1	
	Number of Items	46	4	46	4	51	4	51	4	51	4	51	4	50	4
	Total Score Points for Test	57		57		62		62		62		62		58	

*Note: The CR items do not report out to any single subskill. The items in “A: Mathematical Processes” also do not report out to any single subskill.

Table 2-3
Target Language Arts Test Blueprint: Grades 4, 8, 10

Content Standard		Grade 4		Grade 8		Grade 10	
		MC	Prompt	MC	Prompt	MC	Prompt
B	Writing	19	1	18	1	15	1
D	Language	5	0	6	0	9	0
F	Research and Inquiry	6	0	6	0	6	0
	Number of Items	30	1	30	1	30	1
	Total Number of Points	30	9	30	9	30	9

Table 2-5
Target Science Test Blueprint: Grades 4, 8, 10*

Content Standard		Grade 4	Grade 8	Grade 10
A	Science Connections	4	4	5
B	Nature of Science	4	3	5
C	Science Inquiry	7	8	10
D	Physical Science	6	6	7
E	Earth and Space	6	6	6
F	Life and Environment	6	6	7
G	Science Applications	4	4	5
H	Personal/Social Perspectives	3	3	5
	Total Number of MC Items	40	40	50

*Note: Standard A, Science Connections, and Standard B, Nature of Science, are combined to form a reporting category; Standard G, Science Applications, and Standard H, Personal/Social Perspectives, are combined to form a reporting category.

Table 2-6
Actual Reading Test Blueprint: Grades 3–8, 10*

	Category Title	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 10	
		MC	CR	MC	CR										
1	Determines meaning of words or phrases in context	12	0	12	0	11	0	10	0	10	0	10	0	7	0
1.1	Uses context clues to determine meaning of words or phrases	10		7		6		6		6		5		5	
1.2	Uses knowledge of word structure to determine meaning of words	0		4		2		3		4		3		0	
1.3	Uses word reference materials to determine meaning of words and phrases	2		1		3		1		0		2		2	
2	Understands Text	20	0	19	0	17	0	14	0	14	0	14	0	7	0
2.1	Demonstrates understanding of literal meaning by identifying stated information in literary text	9		9		7		3		5		7		1	
2.2	Demonstrates understanding of literal meaning by identifying stated information in informational text	9		7		7		8		5		4		6	
2.3	Demonstrates understanding of explicitly stated sequence of events in literary and informational text	2		3		3		3		4		3		0	
3	Analyzes Text	18	1	18	1	18	1	19	1	19	1	19	1	22	1
3.1	Analyzes literary text	8		6	1	6		9	1	8		9		8	
3.2	Analyzes informational text.	8	1	7		8	1	6		8	1	6		9	
3.3	Analyzes author's use of language in literary and informational text.	2		5		4		4		3		4	1	5	1
4	Evaluates and Extends Text	4	1	5	2	11	1	11	1	11	1	11	1	14	1
4.1	Evaluates and extends literary text	2		0		4		4	1	4		4		5	
4.2	Evaluates and extends informational text	1	1	3	1	2	1	5		5		3	1	7	1
4.3	Evaluates and extends author's use of language in literary and informational text	1		2		2		2		2	1	3		2	

Table 2-6 Cont'd
 Actual Reading Test Blueprint: Grades 3–8, 10

	Category Title	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade10	
		MC	CR												
	Number of Items	54	2	54	2	54	2	54	2	54	2	54	2	50	2
	Total Score Points for Test	60		60		60		60		60		60		56	

* Note: The CR items do not report out to any single subskill.

Table 2-4
Target Social Studies Test Blueprint: Grades 4, 8, 10

Content Standard		Grade 4	Grade 8	Grade 10
A	Geography	9	10	10
B	History	8	13	12
C	Political Science	7	6	12
D	Economics	7	6	8
E	Behavioral Science	7	5	8
Total Number of MC Items		38	40	50

Table 2-7
Actual Mathematics Test Blueprint: Grades 3–8, 10*

	Category Title	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 10	
		MC	CR	MC	CR										
A	Mathematical Processes	3	3	3	3	3	3	3	3	3	3	3	3	7	1
Aa	Reasoning														
Ab	Communication														
Ac	Connections														
Ad	Representation														
Ae	Problem Solving														
B	Number Operations and Relationships	11	1	11	0	11	0	12	0	12	0	7	0	7	0
Ba	Number Concepts	6		5		5		6		6		5		4	
Bb	Number Computation	5		6		6		6		6		2		3	
C	Geometry	9	2	9	1	9	1	9	1	10	2	8	1	8	1
Ca	Describing Figures	1		2		2		2		3		1		5	
Cb	Spatial Relationships and Transformations	6		5		5		4		4		4		1	
Cc	Coordinate System	2		2		2		3		3		3		2	
D	Measurement	8	0	8	1	9	1	9	1	9	0	11	1	9	1
Da	Measurable Attributes	3		2		4		2		3		3		1	
Db	Direct Measurement	3		4		3		3		3		3		0	
Dc	Indirect Measurement	2		2		2		4		3		5		8	
E	Statistics and Probability	8	0	7	1	9	1	8	1	8	1	8	1	9	0
Ea	Data Analysis and Statistics	4		5		6		5		5		4		4	
Eb	Probability	4		2		3		3		3		4		5	
F	Algebraic Relationships	7	1	8	1	10	1	10	1	9	1	14	1	10	1
Fa	Patterns, Relations, and Functions	3		3		5		5		2		6		3	

Fb	Expressions, Equations, and Inequalities	2		3		3		2		3		3		6	
Fc	Properties	2		2		2		3		4		5		1	
	Number of Items	46	4	46	4	51	4	51	4	51	4	51	4	50	4
	Total Score Points for Test	57		57		62		62		62		62		58	

*The items in “A: Mathematical Processes” do not report out to any single subskill. Note also that some CR items in Grades 3–8 report out to more than one standard. The total number of CR items is 4 per grade even though some items are associated with more than one standard.

Table 2-8
Actual Language Arts Test Blueprint: Grades 4, 8, 10

Content Standard		Grade 4		Grade 8		Grade 10	
		MC	Prompt	MC	Prompt	MC	Prompt
B	Writing	20	1	17	1	15	1
D	Language	4	0	6	0	9	0
F	Research and Inquiry	6	0	7	0	6	0
Total Number of Items		30	1	30	1	30	1
Total Number of Points		30	9	30	9	30	9

Table 2-9
Actual Science Test Blueprint: Grades 4, 8, 10*

Content Standard*		Grade 4	Grade 8	Grade 10
A	Science Connections	4	4	5
B	Nature of Science	4	3	5
C	Science Inquiry	7	8	10
D	Physical Science	6	6	7
E	Earth and Space	6	6	6
F	Life and Environment	6	6	7
G	Science Applications	4	4	5
H	Personal/Social Perspectives	3	3	5
Total Number of MC Items		40	40	50

*Note: Standard A, Science Connections, and Standard B, Nature of Science, are combined to form a reporting category; Standard G, Science Applications, and Standard H, Personal/Social Perspectives, are combined to form a reporting category.

Table 2-10
Actual Social Studies Test Blueprint: Grades 4, 8, 10

Content Standard		Grade 4	Grade 8	Grade 10
A	Geography	9	10	10
B	History	8	13	12
C	Political Science	7	6	12
D	Economics	7	6	8
E	Behavioral Science	7	5	8
Total Number of MC Items		38	40	50

Table 2-11
Item Development Each Year and Total to Date*

	MC items for 2004	CR items for 2004	MC items for 2005	CR items for 2005	MC items for 2006	CR items for 2006	MC items for 2007	CR items for 2007	MC items for 2008	CR items for 2008	MC items for 2009	CR items for 2009	Total MC to date	Total CR to date
Grade 3														
Reading	411	52	23	2	30	4	40	3	52	4	51	7	607	72
Math	317	36	33	14	18	2	30	4	28	11	52	6	478	73
Total	728	88	56	16	48	6	70	7	80	15	103	13	1085	145
Grade 4														
Reading	380	56	32	3	34	3	25	4	54	4	52	7	577	77
Math	265	35	45	9	29	1	26	4	28	13	54	11	447	73
Language Arts	0	0	0	10	0	0	0	0	0	0	0	0	0	10
Science	0	0	0	0	123	34	0	0	0	0	0	0	123	34
Total	645	91	77	22	186	38	51	8	82	17	106	18	1147	194
Grade 5														
Reading	433	59	36	6	29	5	29	7	44	4	52	7	623	88
Math	305	49	38	11	26	3	30	5	28	13	53	8	480	89
Total	738	108	74	17	55	8	59	12	72	17	105	15	1103	177
Grade 6														
Reading	511	56	32	5	42	5	37	6	46	5	50	7	718	84
Math	310	41	53	16	7	2	28	4	30	12	41	8	469	83
Total	821	97	85	21	49	7	65	10	76	17	91	15	1187	167
Grade 7														
Reading	359	44	35	4	38	4	25	5	50	4	50	7	557	68
Math	305	34	32	23	20	0	28	4	31	10	40	6	456	77
Total	664	78	67	27	58	4	53	9	81	14	90	13	1013	145
Grade 8														
Reading	365	44	30	4	34	4	25	4	44	4	50	7	548	67
Math	289	51	47	25	20	2	28	4	32	17	40	8	456	107
Language Arts	0	0	0	10	0	0	0	0	0	0	0	0	0	10
Science	0	0	0	0	125	34	0	0	0	0	0	0	125	34
Total	654	95	77	39	179	40	53	8	76	21	90	15	1129	218
Grade 10														
Reading	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Math	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Language Arts	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Science	0	0	0	0	18	8	0	0	0	0	0	0	18	8
Total	0	0	0	0	18	8	0	0	0	0	0	0	18	8
TOTALS														
Reading	2,459	311	188	24	207	25	181	29	290	25	305	42	3,630	456
Mathematics	1,791	246	248	98	120	10	170	25	177	76	280	47	2,786	502
Language Arts	0	0	0	20	0	0	0	0	0	0	0	0	0	20
Science	0	0	0	0	266	76	0	0	0	0	0	0	266	76
Grand Total	4,250	557	436	142	593	111	351	54	467	101	585	89	6,682	1,054

*Note: This table includes 17 Fall 2009 Math items rejected by DPI prior to the Content and Bias Review.

Table 3-1
Fall 2011 Test Configuration

Content	Grade	No. of OP MC Items	No. of OP CR Items					Total Score Point	Total OP (MC + CR) Items
			1 point	2 point	3 point	4 point	6 point		
Reading	3	54			2			60	56
	4	53*			2			59	55
	5	54			2			60	56
	6	54			2			60	56
	7	54			2			60	56
	8	54			2			60	56
	10	50			2			56	52
Mathematics**	3	46	3	4				57	53
	4	46	3	4				57	53
	5	51	3	4				62	58
	6	51	3	4				62	58
	7	51	3	4				62	58
	8	51	3	4				62	58
	10	48*	0	4				56	52
Language Arts***	4	30						30	30
	8	29*						29	29
	10	30			1		1	39	32
Social Studies	4	36*						36	36
	8	40						40	40
	10	50						50	50
Science	4	40						40	40
	8	40						40	40
	10	50						50	50

* One item was dropped in Reading grade 4 and in Language Arts grade 8. Two items were dropped in Social Studies grade 4 and Mathematics grade 10. See Part 7 for more information.

** Some Mathematics items include two parts, Part A and Part B. Each part is counted as an item above.

*** For Language Arts grade 10, the two CR items are from the grade 10 Writing prompt. The Writing prompt in grade 10 is part of the scale score for Language Arts in grade 10.

Table 3-2
Unique Items Field Tested Each Year and Total to Date

	MC 2004	CR 2004	MC 2005	CR 2005	MC 2006	CR 2006	MC 2007	CR 2007	MC 2008	CR 2008	MC 2009	CR 2009	Total MC to Date	Total CR to Date
Grade 3														
Reading	242	12	24	2	27	2	40	4	40	4	40	4	413	28
Math	252	24	15	2	32	4	34	5	31	8	40	4	404	47
Total	494	36	39	4	59	6	74	9	71	12	80	8	817	75
Grade 4														
Reading	294	12	24	2	32	3	40	4	40	4	40	4	470	29
Math	231	29	15	2	32	4	34	4	28	8	40	4	380	51
Language Arts	0	0	0	6	0	0	0	0	0	0	0	0	0	6
Science	0	0	0	0	40	0	0	0	0	0	0	0	40	0
Social Studies	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Total	525	41	39	10	104	7	74	8	68	12	80	8	890	86
Grade 5														
Reading	235	14	24	2	28	2	29	6	40	4	40	4	396	32
Math	257	34	15	2	32	4	40	4	34	6	40	4	418	54
Total	492	48	39	4	60	6	69	10	74	10	80	8	814	86
Grade 6														
Reading	259	14	24	1	33	3	35	5	40	4	40	4	431	31
Math	252	33	15	2	32	4	32	4	30	5	32	4	393	52
Total	511	47	39	3	65	7	67	9	70	9	72	8	824	83
Grade 7														
Reading	259	14	24	1	17	2	35	4	40	4	40	4	415	29
Math	243	33	15	2	32	4	32	3	33	4	32	4	387	50
Total	502	47	39	3	49	6	67	7	73	8	72	8	802	79
Grade 8														
Reading	274	14	24	1	33	4	32	5	40	4	40	4	443*	32
Math	234	33	15	2	40	4	32	4	32	5	32	4	385	52
Language Arts	0	0	0	6	0	0	0	0	0	0	0	0	0	6
Science	0	0	0	0	40	0	0	0	0	0	0	0	40	0
Social Studies	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Total	508	47	39	9	113	8	64	9	72	9	72	8	868	90
Grade 10														
Reading	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Math	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Language Arts	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Science	0	0	0	0	10	0	0	0	0	0	0	0	10	0
Social Studies	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Total	0	0	0	0	10	0	0	0	0	0	0	0	10	0
TOTALS														
Grand Totals	3,032	266	234	33	460	40	415	52	428	60	456	48	5025	499

* In the 2008 Technical Report, this subtotal was incorrect by 2 MC items. The totals in 2008, however, were correct. This subtotal error has been corrected in the 2009 report.

Table 4-1
Test Accommodations

Accommodation	Grade	N Count	Content Area	Number of Students	Percent of Total
Used a Scribe	3	59024	Reading	811	1.37
		59352	Mathematics	1068	1.80
	4	58912	Reading	773	1.31
		59128	Mathematics	898	1.52
		59080	Science	371	0.63
		58912	Language Arts	350	0.59
		59200	Writing	461	0.78
		59024	Social Studies	367	0.62
	5	59864	Reading	759	1.27
		60048	Mathematics	783	1.30
	6	60720	Reading	458	0.75
		60816	Mathematics	421	0.69
	7	60344	Reading	313	0.52
		60472	Mathematics	316	0.52
	8	60296	Reading	243	0.40
		60416	Mathematics	225	0.37
		60240	Science	132	0.22
		60184	Language Arts	143	0.24
		60560	Writing	289	0.48
		60232	Social Studies	132	0.22
	10	63624	Reading	137	0.22
		63712	Mathematics	139	0.22
		63384	Science	71	0.11
63296		Language Arts	77	0.12	
64208		Writing	122	0.19	
63296		Social Studies	71	0.11	

Table 4-1
Test Accommodations Cont'd

Accommodation	Grade	N Count	Content Area	Number of Students	Percent of Total
Provided Extra Time	3	59024	Reading	6507	11.02
		59352	Mathematics	6898	11.62
	4	58912	Reading	7288	12.37
		59128	Mathematics	7495	12.67
		59080	Science	7240	12.25
		58912	Language Arts	7093	12.04
		59200	Writing	6964	11.76
		59024	Social Studies	7174	12.15
	5	59864	Reading	7342	12.26
		60048	Mathematics	7564	12.60
	6	60720	Reading	6706	11.04
		60816	Mathematics	6778	11.14
	7	60344	Reading	6465	10.71
		60472	Mathematics	6620	10.95
	8	60296	Reading	6395	10.61
		60416	Mathematics	6461	10.69
		60240	Science	6146	10.20
		60184	Language Arts	6193	10.29
		60560	Writing	6160	10.17
		60232	Social Studies	6167	10.24
10	63624	Reading	4161	6.54	
	63712	Mathematics	4231	6.64	
	63384	Science	3984	6.29	
	63296	Language Arts	4081	6.45	
	64208	Writing	3981	6.20	
	63296	Social Studies	4003	6.32	

Table 4-1
 Test Accommodations Cont'd

Accommodation	Grade	N Count	Content Area	Number of Students	Percent of Total
Read Test Questions and Content to Student	3	59024	Reading	0	0.00
		59352	Mathematics	6038	10.17
	4	58912	Reading	0	0.00
		59128	Mathematics	6531	11.04
		59080	Science	6296	10.66
		58912	Language Arts	5810	9.86
		59200	Writing	5921	10.00
		59024	Social Studies	6261	10.61
	5	59864	Reading	0	0.00
		60048	Mathematics	6361	10.59
	6	60720	Reading	0	0.00
		60816	Mathematics	5291	8.70
	7	60344	Reading	0	0.00
		60472	Mathematics	4916	8.13
	8	60296	Reading	0	0.00
		60416	Mathematics	4574	7.57
		60240	Science	4466	7.41
		60184	Language Arts	4226	7.02
		60560	Writing	4385	7.24
		60232	Social Studies	4440	7.37
	10	63624	Reading	0	0.00
		63712	Mathematics	2657	4.17
		63384	Science	2726	4.30
		63296	Language Arts	2640	4.17
64208		Writing	2751	4.28	
63296		Social Studies	2711	4.28	

Table 4-1
Test Accommodations Cont'd

Accommodation	Grade	N Count	Content Area	Number of Students	Percent of Total
Used DPI-Provided Test Translation	3	59024	Reading	0	0.00
		59352	Mathematics	1076	1.81
	4	58912	Reading	0	0.00
		59128	Mathematics	904	1.53
		59080	Science	878	1.49
		58912	Language Arts	0	0.00
		59200	Writing	671	1.13
		59024	Social Studies	852	1.44
	5	59864	Reading	0	0.00
		60048	Mathematics	718	1.20
	6	60720	Reading	0	0.00
		60816	Mathematics	234	0.38
	7	60344	Reading	0	0.00
		60472	Mathematics	228	0.38
	8	60296	Reading	0	0.00
		60416	Mathematics	146	0.24
		60240	Science	124	0.21
		60184	Language Arts	0	0.00
		60560	Writing	78	0.13
		60232	Social Studies	108	0.18
	10	63624	Reading	0	0.00
		63712	Mathematics	104	0.16
		63384	Science	65	0.10
		63296	Language Arts	0	0.00
		64208	Writing	48	0.07
		63296	Social Studies	67	0.11

Table 4-1
 Test Accommodations Cont'd

Accommodation	Grade	N Count	Content Area	Number of Students	Percent of Total
Used Locally Provided Test Translation	3	59024	Reading	0	0.00
		59352	Mathematics	0	0.00
	4	58912	Reading	0	0.00
		59128	Mathematics	102	0.17
		59080	Science	102	0.17
		58912	Language Arts	0	0.00
		59200	Writing	77	0.13
		59024	Social Studies	97	0.16
	5	59864	Reading	0	0.00
		60048	Mathematics	0	0.00
	6	60720	Reading	0	0.00
		60816	Mathematics	0	0.00
	7	60344	Reading	0	0.00
		60472	Mathematics	0	0.00
	8	60296	Reading	0	0.00
		60416	Mathematics	69	0.11
		60240	Science	66	0.11
		60184	Language Arts	0	0.00
		60560	Writing	54	0.09
		60232	Social Studies	58	0.10
	10	63624	Reading	0	0.00
		63712	Mathematics	55	0.09
		63384	Science	53	0.08
63296		Language Arts	0	0.00	
64208		Writing	42	0.07	
63296		Social Studies	52	0.08	

Table 4-1
Test Accommodations Cont'd

Accommodation	Grade	N Count	Content Area	Number of Students	Percent of Total
Used DPI-Provided Glossary of Terms	3	59024	Reading	0	0.00
		59352	Mathematics	0	0.00
	4	58912	Reading	0	0.00
		59128	Mathematics	295	0.50
		59080	Science	343	0.58
		58912	Language Arts	0	0.00
		59200	Writing	0	0.00
		59024	Social Studies	333	0.56
	5	59864	Reading	0	0.00
		60048	Mathematics	0	0.00
	6	60720	Reading	0	0.00
		60816	Mathematics	0	0.00
	7	60344	Reading	0	0.00
		60472	Mathematics	0	0.00
	8	60296	Reading	0	0.00
		60416	Mathematics	184	0.30
		60240	Science	162	0.27
		60184	Language Arts	0	0.00
		60560	Writing	0	0.00
		60232	Social Studies	147	0.24
	10	63624	Reading	0	0.00
		63712	Mathematics	53	0.08
		63384	Science	12	0.02
		63296	Language Arts	0	0.00
64208		Writing	0	0.00	
63296		Social Studies	40	0.06	

Table 4-1
Test Accommodations Cont'd

Accommodation	Grade	N Count	Content Area	Number of Students	Percent of Total
Used Text Talker	3	59024	Reading	0	0.00
		59352	Mathematics	19	0.03
	4	58912	Reading	0	0.00
		59128	Mathematics	8	0.01
		59080	Science	3	0.01
		58912	Language Arts	3	0.01
		59200	Writing	3	0.01
		59024	Social Studies	4	0.01
	5	59864	Reading	0	0.00
		60048	Mathematics	9	0.02
	6	60720	Reading	0	0.00
		60816	Mathematics	5	0.01
	7	60344	Reading	0	0.00
		60472	Mathematics	2	0.00
	8	60296	Reading	0	0.00
		60416	Mathematics	39	0.06
		60240	Science	33	0.05
		60184	Language Arts	26	0.04
		60560	Writing	3	0.01
		60232	Social Studies	38	0.06
	10	63624	Reading	0	0.00
		63712	Mathematics	13	0.02
		63384	Science	14	0.02
63296		Language Arts	20	0.03	
64208		Writing	10	0.02	
63296		Social Studies	22	0.03	

Table 4-1
 Test Accommodations Cont'd

Accommodation	Grade	N Count	Content Area	Number of Students	Percent of Total
Signed Test Questions and Content to Student	3	59024	Reading	0	0.00
		59352	Mathematics	32	0.05
	4	58912	Reading	0	0.00
		59128	Mathematics	24	0.04
		59080	Science	22	0.04
		58912	Language Arts	20	0.03
		59200	Writing	20	0.03
		59024	Social Studies	18	0.03
	5	59864	Reading	0	0.00
		60048	Mathematics	17	0.03
	6	60720	Reading	0	0.00
		60816	Mathematics	21	0.03
	7	60344	Reading	0	0.00
		60472	Mathematics	17	0.03
	8	60296	Reading	0	0.00
		60416	Mathematics	14	0.02
		60240	Science	13	0.02
		60184	Language Arts	13	0.02
		60560	Writing	13	0.02
		60232	Social Studies	13	0.02
	10	63624	Reading	0	0.00
		63712	Mathematics	20	0.03
		63384	Science	18	0.03
63296		Language Arts	19	0.03	
64208		Writing	20	0.03	
63296		Social Studies	19	0.03	

Table 4-1
 Test Accommodations Cont'd

Accommodation	Grade	N Count	Content Area	Number of Students	Percent of Total
Used Another DPI-Approved Accommodation	3	59024	Reading	1499	2.54
		59352	Mathematics	1781	3.00
	4	58912	Reading	1849	3.14
		59128	Mathematics	1968	3.33
		59080	Science	1752	2.97
		58912	Language Arts	1723	2.92
		59200	Writing	1734	2.93
		59024	Social Studies	1754	2.97
	5	59864	Reading	1959	3.27
		60048	Mathematics	2192	3.65
	6	60720	Reading	1816	2.99
		60816	Mathematics	1826	3.00
	7	60344	Reading	1515	2.51
		60472	Mathematics	1594	2.64
	8	60296	Reading	1398	2.32
		60416	Mathematics	1444	2.39
		60240	Science	1402	2.33
		60184	Language Arts	1389	2.31
		60560	Writing	1422	2.35
		60232	Social Studies	1383	2.30
	10	63624	Reading	700	1.10
		63712	Mathematics	751	1.18
		63384	Science	669	1.06
63296		Language Arts	665	1.05	
64208		Writing	674	1.05	
63296		Social Studies	680	1.07	

Table 4-1
 Test Accommodations Cont'd

Accommodation	Grade	N Count	Content Area	Number of Students	Percent of Total
Used DPI-Provided Braille Test	3	59024	Reading	0	0.00
		59352	Mathematics	0	0.00
	4	58912	Reading	22	0.04
		59128	Mathematics	22	0.04
		59080	Science	22	0.04
		58912	Language Arts	22	0.04
		59200	Writing	22	0.04
		59024	Social Studies	22	0.04
	5	59864	Reading	13	0.02
		60048	Mathematics	13	0.02
	6	60720	Reading	12	0.02
		60816	Mathematics	13	0.02
	7	60344	Reading	6	0.01
		60472	Mathematics	6	0.01
	8	60296	Reading	0	0.00
		60416	Mathematics	0	0.00
		60240	Science	0	0.00
		60184	Language Arts	0	0.00
		60560	Writing	0	0.00
		60232	Social Studies	0	0.00
	10	63624	Reading	9	0.01
		63712	Mathematics	9	0.01
		63384	Science	9	0.01
		63296	Language Arts	9	0.01
64208		Writing	9	0.01	
63296		Social Studies	9	0.01	

Table 4-1
 Test Accommodations Cont'd

Accommodation	Grade	N Count	Content Area	Number of Students	Percent of Total
Used a Non-Allowed Accommodation	3	59024	Reading	0	0.00
		59352	Mathematics	0	0.00
	4	58912	Reading	0	0.00
		59128	Mathematics	0	0.00
		59080	Science	0	0.00
		58912	Language Arts	0	0.00
		59200	Writing	0	0.00
		59024	Social Studies	0	0.00
	5	59864	Reading	0	0.00
		60048	Mathematics	0	0.00
	6	60720	Reading	0	0.00
		60816	Mathematics	0	0.00
	7	60344	Reading	0	0.00
		60472	Mathematics	0	0.00
	8	60296	Reading	0	0.00
		60416	Mathematics	0	0.00
		60240	Science	0	0.00
		60184	Language Arts	0	0.00
		60560	Writing	0	0.00
		60232	Social Studies	0	0.00
	10	63624	Reading	0	0.00
		63712	Mathematics	0	0.00
		63384	Science	0	0.00
		63296	Language Arts	0	0.00
64208		Writing	0	0.00	
63296		Social Studies	0	0.00	

Table 5-1
Reading Rubric, Grades 3–8 and 10

Reading items at all grade levels were scored using item-specific scoring guides that are based on a generic, 0–3 holistic rubric.

3 points

- The response demonstrates *thorough understanding* of the reading concept embodied in the task.
- The response is *accurate, complete, insightful, and fulfills all the requirements* of the task.
- Necessary support and/or examples are included.
- Information is clearly *text-based*.

2 points

- The response demonstrates *partial understanding* of the reading concept embodied in the task.
- The response is *accurate* and *fulfills most of the requirements* of the task.
- Necessary support and/or examples may not be complete or clearly text-based.

1 point

- The response demonstrates *an incomplete understanding* of the reading concept embodied in the task.
- The response provides *some information that is text-based*, but does not fulfill the requirements of the task.
- Information provided is *too general or too simplistic*.
- Necessary support and/or examples may be incomplete or omitted.

0 points

- The response demonstrates *no understanding* of the reading concept embodied in the task.
- The response is *inaccurate, confused, or irrelevant*.
- The student has written a response but *failed to respond to the task*.

Table 5-2
Mathematics Rubric, Grades 3–8 and 10

Generic Rubric for Mathematics for 2-point Constructed Response Items

2 points	<p>The student demonstrates a thorough understanding of the mathematical concepts and/or procedures represented in the problem. The student states appropriate mathematical responses and/or uses procedures and/or concepts to explain or justify a response. The student provides clear and complete explanations and interpretations containing words, calculations, or symbols, when specified in the item stem.</p> <p>The response may contain minor flaws that do <u>not</u> detract from the demonstration of a thorough understanding of the problem.</p>
1 point	<p>The student demonstrates only a partial understanding of the mathematical concepts and/or procedures represented in the problem. The response lacks an appropriate mathematical response or reflects the lack of an essential understanding of the underlying mathematical concepts used in the item.</p> <p>The response contains errors related to the misinterpretation of important aspects of the problem, misuse of mathematical procedures and/or concepts, or misinterpretation of results.</p>
0 points	<p>The student provides completely incorrect responses, explanations, or justifications, or ones that cannot be interpreted, for all responses required in the item.</p>

Generic Rubric for Mathematics for 3-point Constructed Response Items

Mathematics 3-point constructed response items have two parts. Part A is scored as correct/incorrect. Part B is scored using the 2-point holistic rubric below.

2 points	<p>The student demonstrates a thorough understanding of the mathematical concepts and/or procedures represented in the problem. The student uses appropriate mathematical procedures and/or concepts to explain or justify the response to Step A, and provides clear and complete explanations and interpretations containing words, calculations, or symbols, unless otherwise specified in the item stem.</p> <p>The response may contain minor flaws that do <u>not</u> detract from the demonstration of a thorough understanding of the problem.</p>
1 point	<p>The student demonstrates only a partial understanding of the mathematical concepts and/or procedures represented in the problem. The response lacks an essential understanding of the underlying mathematical concepts used to provide the response to Step A.</p> <p>The response contains errors related to the misinterpretation of important aspects of the problem, misuse of mathematical procedures and/or concepts, or misinterpretation of results.</p>
0 points	<p>The student provides a completely incorrect explanation or justification, or one that cannot be interpreted.</p>

Table 5-3
Writing Rubric, Conventions of Written English, Grade 4

3 points Advanced Control

The response demonstrates advanced control of a wide range of conventions identified in the 4th grade Wisconsin Model Academic Standards in English Language Arts:

- Uses parts of speech effectively, including nouns, pronouns, and adjectives
- Uses adverbials effectively, including words and phrases
- Employs principles of agreement related to number, gender, and case
- Capitalizes proper nouns, titles, and initial words of sentences
- Uses punctuation marks and conjunctions, as appropriate, to separate sentences and connect independent clauses
- Uses commas correctly to punctuate appositives and lists
- Spells correctly in general and even on difficult words
- Uses word order and punctuation marks to distinguish statements, questions, exclamations, and commands
- Makes errors that are infrequent and minor

2 points Proficient Control

The response demonstrates proficient control of the essential conventions identified in the 4th grade Wisconsin Model Academic Standards in English Language Arts:

- Generally controls grammar and usage (principles of agreement, noun and verb forms, superlative and comparative forms)
- Capitalizes proper nouns, titles, and initial words of sentences
- Uses end-stop punctuation correctly most of the time; internal punctuation (commas, apostrophes) is sometimes missing or wrong.
- Generally uses correct spelling with common words but more difficult words are problematic
- Makes errors typical of those commonly found in a rough draft; errors do not significantly distract the reader

1 point Minimal Control

The response demonstrates minimal control of the essential conventions identified in the 4th grade Wisconsin Model Academic Standards in English Language Arts:

- Contains numerous serious end-stop punctuation errors, resulting in fragments, comma splices, run-ons
- Shows poor control of subject/verb agreement, possessive forms, capitalization, superlatives, and comparatives
- Spelling errors are frequent, even on common words
- Makes errors that are frequent, varied, and distracting

Table 5-4
Writing Rubric, Composing, Grade 4

Wisconsin Writing Grade 4 Rubric 6-Point Scoring Guide						
Elements of Rubric	Purpose & Focus	Organization & Coherence	Development of Content	Sentence Fluency	Word Choice	
<i>Element Description</i>	Consistently focuses on the topic and maintains a unified purpose Demonstrates understanding of the requirements of the assigned task	Uses a logical plan of development with an effective beginning, middle, and end Keeps relationships among ideas clear Paragraphs logically and uses appropriate transitional devices	Expands and supports main ideas with specific details, examples, and/or reasons that are 1) clearly related to the topic and purpose, and 2) effective for audience	Uses varied sentence structures, creating a fluent, effective, and readable style	Controls word choice with respect to both denotation and connotation Demonstrates attention to context (audience, purpose, situation, tone) Evidences some control over figurative language for rhetorical effect (e.g. metaphors, similes)	
<i>Positive Descriptors</i>	Focused, unified, controlled, relevant	Well organized, integrated, smooth, controlled, coherent	Thorough, specific, well-developed, well-supported, well-illustrated, insightful, convincing	Fluid, varied, controlled, effective	Vivid, precise, concrete, concise	
<i>Negative Descriptors</i>	Rambling, loosely related, redundant, irrelevant, lacks purpose	Disorganized, hard to follow, mechanical, illogical shifts, incoherent	Vague, general, simplistic, superficial, incomplete, illogical, inadequately supported, lacks illustration	Choppy, simple, repetitive, garbled, ineffective, awkward	Awkward, imprecise, vague wordy, repetitive	
Rubric Holistic Scoring Scale						
Scores	6	5	4	3	2	1
<i>Description</i>	Exemplary control of the domain	Advanced control of the domain	Proficient control of the domain	Adequate control of the domain	Basic control of the domain	Minimal control of the domain

Table 5-5
Writing Rubric, Conventions of Written English, Grade 8

3 points Advanced Control

The response demonstrates advanced control of a wide range of conventions identified in the 8th grade Wisconsin Model Academic Standards in English Language Arts:

- Uses words, phrases, and clauses effectively, including coordinate and subordinate conjunctions, relative pronouns, and comparative adjectives
- Uses correct tenses to indicate the relative order of events
- Employs principles of agreement, including subject-verb, pronoun-noun, and preposition-pronoun
- Punctuates compound, complex, and compound-complex sentences correctly
- Employs the conventions of capitalization
- Spells frequently used words correctly and uses effective strategies for spelling unfamiliar words
- Makes errors that are infrequent and minor

2 points Proficient Control

The response demonstrates proficient control of the conventions identified in the 8th grade Wisconsin Model Academic Standards in English Language Arts:

- Generally controls grammar and usage (principles of agreement, noun and verb forms, pronoun reference, superlative, and comparative forms)
- Generally uses phrases, dependent, and independent clauses clearly and correctly
- Capitalizes most words correctly; control over more sophisticated capitalization skills may be spotty
- Uses end-stop punctuation correctly most of the time; internal punctuation (commas, apostrophes, semicolons) is sometimes missing or wrong
- Generally uses correct spelling with grade-level words and reasonable phonetic approaches to more difficult words
- Makes errors typical of those commonly found in a rough draft; errors do not seriously distract the reader

1 point Minimal Control

The response demonstrates minimal control of the conventions identified in the 8th grade Wisconsin Model Academic Standards in English Language Arts:

- Contains numerous serious end-stop or internal punctuation errors, resulting in fragments, comma splices, run-ons
- Shows poor control of grammar and usage (principles of agreement; verb and/or noun forms including possessives; pronoun reference; superlative and comparative forms; appropriate use of phrases/independent, dependent clauses, capitalization)
- Frequently misspells words, even those on grade-level
- Makes errors that are frequent, varied, and distracting

Table 5-6
Writing Rubric, Composing, Grade 8

Wisconsin Writing Grade 8 Rubric 6-Point Scoring Guide					
Elements of Rubric	Purpose & Focus	Organization & Coherence	Development of Content	Sentence Fluency	Word Choice
<i>Element Description</i>	Clearly presents and maintains a unified purpose, focus, and/or thesis Demonstrates understanding of the requirements of the assigned task	Frames the discussion with an effective introduction and conclusion Creates a logical structure of development for the topic, thesis, and purpose Uses transitional strategies (from idea to idea, paragraph to paragraph, and sentence to sentence)	Demonstrates <i>quality</i> of invented content (e.g. of explanations, arguments, rationale, ideas, details, examples, illustrations) Demonstrates <i>thoroughness</i> in the elaboration of content	Demonstrates use of varied syntactic structures including simple, compound, complex, and compound/complex sentences Evidences some control over stylistic effects (e.g. variety, readability)	Controls word choice with respect to both denotation and connotation Demonstrates attention to context (audience, purpose, situation, tone) Evidences some control over figurative language for rhetorical effect (e.g. similes, metaphors, personification)
<i>Positive Descriptors</i>	Focused, unified, controlled, relevant	Well organized, integrated, smooth, controlled, coherent	<u>Quality</u> : clear, convincing, accurate, effective, well-reasoned, insightful <u>Thoroughness</u> : specific, well-developed, well-supported, well-illustrated	Fluid, varied, controlled, effective	Apt, discriminating, vivid, precise, concrete, concise
<i>Negative Descriptors</i>	Rambling, loosely related, redundant, irrelevant, lacks purpose	Disorganized, hard to follow, mechanical, illogical shifts, incoherent	<u>Quality</u> : vague, imprecise, inaccurate, simplistic, poorly reasoned, superficial <u>Thoroughness</u> : incomplete, general, inadequately developed, inadequately supported, lacks illustration	Choppy, monotonous, garbled, ineffective, awkward	Inappropriate, clichéd, awkward, imprecise, vague, wordy

Table 5-6 Cont'd
 Writing Rubric, Composing, Grade 8

Rubric Holistic Scoring Scale						
Scores	6	5	4	3	2	1
<i>Description</i>	Exemplary control of the domain	Advanced control of the domain	Proficient control of the domain	Adequate control of the domain	Basic control of the domain	Minimal control of the domain

Table 5-7
Writing Rubric, Conventions of Written English, Grade 10

3 points **Advanced Control**

The response demonstrates advanced control of a wide range of conventions identified in the 12th grade Wisconsin Model Academic Standards in English Language Arts:

- Uses words, phrases, and clauses effectively, including interrelated clauses in complex sentences
- Uses correct tenses, including conditionals, to indicate the relative order and relationship of events
- Employs principles of agreement, including subject-verb, pronoun-noun, and preposition-pronoun
- Punctuates compound, complex, and compound-complex sentences correctly, including appropriate use of colons, hyphens, dashes, ellipses, and italics; punctuates dialogue correctly; follows citation conventions
- Employs the conventions of capitalization
- Spells frequently used words correctly and uses effective strategies for spelling unfamiliar words
- Makes errors that are infrequent and minor

2 points **Proficient Control**

The response demonstrates proficient control of essential conventions identified in the 12th grade Wisconsin Model Academic Standards in English Language Arts:

- Generally controls grammar and usage (principles of agreement, noun and verb forms, pronoun references, superlative, and comparative forms)
- Generally uses phrases, dependent, and independent clauses clearly and correctly
- Uses end-stop punctuation correctly most of the time; internal punctuation (commas, apostrophes, semicolons, colons) is sometimes missing or wrong; sometimes fails to punctuate dialogue correctly or to accurately follow citation conventions
- Employs the conventions of capitalization
- Generally uses correct spelling with grade-level words and reasonable phonetic approaches to more difficult words
- Makes errors typical of those commonly found in a rough draft; errors do not seriously distract the reader

1 point **Minimal Control**

The response demonstrates minimal control of essential conventions identified in the 12th grade Wisconsin Model Academic Standards in English Language Arts

- Contains numerous serious end-stop or internal punctuation errors, resulting in fragments, comma splices, run-ons
- Shows poor control of grammar and usage (principles of agreement, verb and/or noun forms; pronoun reference; superlative and comparative forms)
- Shows poor control of spelling, even on grade-level words
- Makes errors that are frequent, varied, and distracting

Table 5-8
Writing Rubric, Composing, Grade 10

Wisconsin Writing Grade 10 Rubric 6-Point Scoring Guide					
Elements of Rubric	Purpose & Focus	Organization & Coherence	Development of Content	Sentence Fluency	Word Choice
<i>Element Description</i>	<p>Explicitly states, or strongly implies, a thesis or unifying purpose which firmly guides the paper</p> <p>Demonstrates understanding of the requirements of the assigned task</p>	<p>Frames the discussion with an effective introduction and conclusion</p> <p>Creates a logical structure of development for the topic, thesis, and purpose</p> <p>Uses effective and varied transitional strategies (from idea to idea, paragraph to paragraph, and sentence to sentence)</p>	<p>Demonstrates quality of invented content (e.g. of explanations, arguments, rationale, ideas, details, examples, illustrations)</p> <p>Demonstrates thoroughness in the elaboration of content</p>	<p>Demonstrates syntactic control of simple, compound, complex, and compound/complex sentences</p> <p>Evidences some control over stylistic effects (e.g. flow, cadence, parallelism, variety, readability, judicious use of active and passive voice, effective repetition)</p>	<p>Controls word choice with respect to both denotation and connotation</p> <p>Demonstrates attention to context (audience, purpose, situation, tone)</p> <p>Evidences some control over figurative language for rhetorical effect (e.g. metaphors, similes, hyperbole, analogies)</p>
<i>Positive Descriptors</i>	Focused, unified, controlled, relevant	Well organized, integrated, smooth, controlled, coherent	<p><u>Quality</u>: clear, precise, accurate, effective, well-reasoned, insightful</p> <p><u>Thoroughness</u>: complete, specific, well-developed, well-supported, well-illustrated</p>	Fluid, varied, controlled, effective, skilled	Apt, discriminating, vivid, precise, concrete, concise
<i>Negative Descriptors</i>	Rambling, loosely related, redundant, irrelevant, lacks purpose	Disorganized, hard to follow, mechanical, illogical shifts, incoherent	<p><u>Quality</u>: vague, imprecise, inaccurate, simplistic, poorly reasoned, superficial</p> <p><u>Thoroughness</u>: incomplete, general, inadequately developed, inadequately supported, lacks illustration</p>	Choppy, monotonous, garbled, ineffective, awkward	Inappropriate, clichéd, awkward, imprecise, vague, wordy

Table 5-8 Cont'd
 Writing Rubric, Composing, Grade 10

Rubric Holistic Scoring Scale						
Scores	6	5	4	3	2	1
<i>Description</i>	Exemplary control of the domain	Advanced control of the domain	Proficient control of the domain	Adequate control of the domain	Basic control of the domain	Minimal control of the domain

Table 5-9
Score Distribution for Reading CR Items*

Grade	Test Book Item No.	N	Scores				Condition Codes**			
			0	1	2	3	A	B	C	D
3	31	59007	13230	21716	20119	2400	1418	32	16	76
	56	59007	16411	27265	12647	1540	1070	2	11	61
4	13	58877	11744	31396	13334	1515	829	14	8	37
	56	58877	23544	10888	20619	2250	1408	61	8	99
5	33	59811	16554	35210	6613	561	807	9	5	52
	50	59811	15325	22715	20349	796	609	5	3	9
6	38	60686	11071	18786	24051	5860	857	5	7	49
	56	60686	7768	23102	25437	3796	526	8	7	42
7	36	60311	13512	8460	28206	9354	714	7	2	56
	56	60311	30665	20201	7601	967	869	2	2	4
8	19	60243	8022	21461	22057	7505	1046	1	6	145
	40	60243	15671	15242	25639	2341	1321	4	7	18
10	12	63564	16535	22589	19162	3768	1462	5	3	40
	43	63564	9694	17074	23042	9757	3954	2	5	36

* This is the score distribution of the first read.

** A: No response or no attempt, B: Illegible, C: Another Language, D: Off-topic.

*** Item dropped from scoring.

Table 5-10
Score Distribution for Mathematics CR Items*

Grade	Test Book Item No.	Part	N	Scores			Condition Codes**			
				0	1	2	A	B	C	D
3	10		59343	8372	24055	26465	416	2	31	2
	25	A	59343	29341	29255	0	727	3	5	12
	25	B	59343	23559	17999	16203	1505	9	30	38
	28	A	59343	1561	57196	0	571	5	4	6
	28	B	59343	13433	4180	40675	966	17	43	29
	44	A	59343	16829	41774	0	700	4	8	28
	44	B	59343	36676	4816	16200	1546	13	54	38
4	13		59078	26423	10095	22270	287	0	2	1
	20	A	59078	29837	28947	0	289	3	1	1
	20	B	59078	25431	20797	11367	1447	3	21	12
	29	A	59078	15542	42910	0	623	0	0	3
	29	B	59078	34835	12963	9797	1428	5	24	26
	41	A	59078	17948	40811	0	308	1	5	5
	41	B	59078	13434	14158	30649	784	8	17	28
5	12	A	59997	24275	35329	0	366	1	6	20
	12	B	59997	9648	31292	18487	540	1	16	13
	19		59997	3672	17035	39029	260	0	0	1
	23	A	59997	36295	23364	0	336	0	2	0
	23	B	59997	36182	3680	19671	456	0	7	1
	46	A	59997	7365	52260	0	371	0	1	0
	46	B	59997	3019	2569	53923	474	0	11	1
6	10	A	60778	22445	38085	0	242	4	1	1
	10	B	60778	17582	6574	36271	328	1	17	5
	22	A	60778	4462	56106	0	203	1	4	2
	22	B	60778	30006	29375	906	464	3	21	3
	35	A	60778	41792	18565	0	417	0	4	0
	35	B	60778	27301	18959	13752	738	4	21	3
	53		60778	12269	22152	25938	417	1	1	0
7	4	A	60453	17895	42124	0	430	1	3	0
	4	B	60453	17515	8413	33565	954	2	4	0
	29	A	60453	28705	30952	0	795	0	1	0
	29	B	60453	17024	7077	35159	1169	2	10	12
	32	A	60453	30668	29332	0	453	0	0	0
	32	B	60453	20071	37109	2022	1231	2	16	2
	51		60453	24162	22337	13405	532	0	3	14

* This is the score distribution of the first read.

** A: No response or no attempt, B: Illegible, C: Another Language, D: Off-topic.

Table 5-10 Cont'd
Score Distribution for Mathematics CR Items*

Grade	Test Book Item No.	Part	N	Scores			Condition Codes**			
				0	1	2	A	B	C	D
8	9	A	60348	5980	54073	0	279	5	10	1
	9	B	60348	12111	3914	43762	541	2	12	6
	20	A	60348	32772	26689	0	883	0	0	4
	20	B	60348	18607	15440	24953	1334	1	8	5
	40	A	60348	31214	28528	0	603	0	0	3
	40	B	60348	30136	2836	26232	1125	2	11	6
	53		60348	37496	6385	14636	1823	1	2	5
10	27		63658	25978	16614	17783	3247	5	11	20
	33		63658	10491	18200	33405	1531	5	9	17
	38		63658	40980	2159	14874	5630	1	4	10
	52		63658	24700	15101	21162	2667	0	6	22

* This is the score distribution of the first read.

** A: No response or no attempt, B: Illegible, C: Another Language, D: Off-topic.

Table 5-11
Score Distribution for Grades 4, 8, and 10 Writing: Composing Rubric

Grade	Rater	Total N	Scores						Condition Codes**			
			1	2	3	4	5	6	A	B	C	D
4	Rater 1	3045	47	433	1148	1011	315	32	24	1	0	34
	Rater 2	3045	58	437	1135	999	342	26	23	1	0	24
	Diff*	0	-11	-4	13	12	-27	6	1	0	0	10
8	Rater 1	3043	34	533	1189	1003	167	17	40	0	0	59
	Rater 2	3043	42	510	1247	997	146	18	39	0	0	44
	Diff*	0	-8	23	-58	6	21	-1	1	0	0	15
10	Rater 1	3244	53	382	1015	1143	472	84	83	0	0	12
	Rater 2	3244	52	381	1000	1163	490	66	81	0	0	11
	Diff*	0	1	1	15	-20	-18	18	2	0	0	1

* Diff = N of Rater1 – N of Rater 2.

** A: No response or no attempt, B: Illegible, C: Another Language, D: Off-topic.

Table 5-12
Percentage Distribution of Scores, Grades 4, 8, and 10 Writing: Composing Rubric

Grade	Rater	Total N	Scores						Condition Codes**			
			1	2	3	4	5	6	A	B	C	D
4	Rater 1	3045	1.54	14.22	37.70	33.20	10.34	1.06	0.78	0.04	0.00	1.12
	Rater 2	3045	1.90	14.36	37.28	32.80	11.24	0.86	0.76	0.04	0.00	0.78
8	Rater 1	3043	1.12	17.52	39.08	32.96	5.48	0.56	1.32	0.00	0.00	1.94
	Rater 2	3043	1.38	16.76	40.98	32.76	4.80	0.60	1.28	0.00	0.00	1.44
10	Rater 1	3244	1.64	11.78	31.28	35.24	14.54	2.58	2.56	0.00	0.00	0.36
	Rater 2	3244	1.60	11.74	30.82	35.86	15.10	2.04	2.50	0.00	0.00	0.34

** A: No response or no attempt, B: Illegible, C: Another Language, D: Off-topic.

Table 5-13
Score Distribution, Grades 4, 8, and 10 Writing: Conventions Rubric

Grade	Rater	Total N	Scores			Condition Codes**		
			1	2	3	A	B	C
4	Rater 1	3045	147	2759	114	24	1	0
	Rater 2	3045	172	2736	113	23	1	0
	Diff*	0	-25	23	1	1	0	0
8	Rater 1	3043	68	2871	63	40	0	1
	Rater 2	3043	71	2882	51	39	0	0
	Diff*	0	-3	-11	12	1	0	1
10	Rater 1	3244	45	2504	612	83	0	0
	Rater 2	3244	45	2512	606	81	0	0
	Diff*	0	0	-8	6	2	0	0

* Diff = N of Rater 1 – N of Rater 2.

** A: No response or no attempt, B: Illegible, C: Another Language.

Table 5-14
Percentage Distribution of Scores for Grades 4, 8, and 10 Writing: Conventions Rubric

Grade	Rater	Total N	Scores			Condition Codes**		
			1	2	3	A	B	C
4	Rater 1	3045	4.82	90.60	3.74	0.78	0.04	0.00
	Rater 2	3045	5.64	89.86	3.72	0.76	0.04	0.00
8	Rater 1	3043	2.24	94.34	2.08	1.32	0.00	0.04
	Rater 2	3043	2.34	94.70	1.68	1.28	0.00	0.00
10	Rater 1	3244	1.38	77.18	18.86	2.56	0.00	0.00
	Rater 2	3244	1.38	77.44	18.68	2.50	0.00	0.00

** A: No response or no attempt, B: Illegible, C: Another Language.

Table 5-15

Score Distribution for Grades 4, 8, and 10 Writing: Total Score, Composing and Conventions Combined

Grade	Rater	Total N	Scores									
			0	1	2	3	4	5	6	7	8	9
4	Rater 1	3045	25	6	68	68	408	1105	992	282	73	18
	Rater 2	3045	24	3	75	78	394	1106	963	321	65	16
	Diff*	0	1	3	-7	-10	14	-1	29	-39	8	2
8	Rater 1	3043	41	3	68	57	515	1171	995	137	43	13
	Rater 2	3043	39	1	61	61	484	1236	992	124	30	15
	Diff*	0	2	2	7	-4	31	-65	3	13	13	-2
10	Rater 1	3244	83	3	33	42	371	946	985	450	261	70
	Rater 2	3244	81	2	29	48	369	939	976	493	246	61
	Diff*	0	2	1	4	-6	2	7	9	-43	15	9

* Diff = N of Rater 1 – N of Rater 2.

Table 5-16

Percentage Distribution of Scores for Grades 4, 8, and 10 Writing: Total Score, Composing and Conventions Combined

Grade	Rater	Total N	Scores									
			0	1	2	3	4	5	6	7	8	9
4	Rater 1	3045	0.82	0.20	2.24	2.24	13.40	36.28	32.58	9.26	2.40	0.60
	Rater 2	3045	0.78	0.10	2.46	2.56	12.94	36.32	31.62	10.54	2.14	0.52
8	Rater 1	3043	1.34	0.10	2.24	1.88	16.92	38.48	32.70	4.50	1.42	0.42
	Rater 2	3043	1.28	0.04	2.00	2.00	15.90	40.62	32.60	4.08	0.98	0.50
10	Rater 1	3244	2.56	0.10	1.02	1.30	11.44	29.16	30.36	13.88	8.04	2.16
	Rater 2	3244	2.50	0.06	0.90	1.48	11.38	28.94	30.08	15.20	7.58	1.88

Table 7-1
Item Flagged Based on Yen's Q_1

Status	Content	Grade	Form	Item Number	Type	N	Z	Critical Z
OP	RD	5	All	18	MC	6400	18.15	17.07
	RD	6	All	23	MC	6312	23.29	16.83
	MA	4	All	41B	CR	6407	37.46	17.09
	MA	4	All	49	MC	6434	26.94	17.16
	MA	6	All	35A	CR	6244	22.97	16.65
	MA	7	All	27	MC	6362	18.99	16.97
	MA	10	All	50	MC	6947	21.10	18.53
	MA	10	All	52	CR	6514	17.70	17.37
	LA	10	All	32	CR	6895	18.47	18.39
	SS	8	All	18	MC	6253	18.57	16.67
	SS	10	All	17	MC	6850	20.57	18.27
FT	RD	4	A	67	CR	2440	9.13	6.51
	RD	7	A	62	MC	14848	102.88	39.59
	RD	8	B	57	MC	15138	50.70	40.37
	MA	3	A	61A	CR	2418	25.15	6.45
	MA	3	B	61A	CR	2428	9.87	6.47
	MA	3	B	61B	CR	2399	8.57	6.40
	MA	3	C	61	CR	2308	7.49	6.15
	MA	3	D	61	CR	2373	6.48	6.33
	MA	4	C	58	CR	2440	9.21	6.51
	MA	5	B	66B	CR	2457	12.26	6.55
	MA	5	C	59	MC	14830	42.13	39.55
	MA	6	B	64A	CR	2453	6.57	6.54
	MA	6	C	64B	CR	2419	8.12	6.45
	MA	7	C	62B	CR	2416	7.56	6.44

Table 7-2
Scoring Table for Reading Grade 3

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	270	126	31	434	7
1	270	126	32	436	7
2	270	126	33	438	7
3	270	126	34	440	7
4	270	126	35	442	7
5	270	126	36	444	7
6	270	126	37	446	7
7	270	126	38	448	7
8	270	126	39	450	7
9	270	126	40	452	7
10	270	126	41	455	7
11	337	59	42	457	7
12	362	34	43	459	7
13	374	22	44	462	7
14	382	17	45	464	7
15	388	14	46	467	8
16	393	12	47	470	8
17	397	11	48	473	8
18	401	10	49	476	8
19	404	10	50	479	9
20	407	9	51	483	9
21	410	9	52	488	10
22	413	8	53	493	11
23	416	8	54	499	12
24	418	8	55	506	14
25	420	8	56	516	17
26	423	8	57	529	21
27	425	7	58	551	30
28	427	7	59	588	45
29	429	7	60	640	73
30	431	7			

* **Bold** represents SEM around cut score.

Table 7-3
Scoring Table for Reading Grade 4

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	280	121	30	452	9
1	280	121	31	455	9
2	280	121	32	458	9
3	280	121	33	461	9
4	280	121	34	464	9
5	280	121	35	467	9
6	280	121	36	470	9
7	280	121	37	473	9
8	280	121	38	476	9
9	280	121	39	479	9
10	280	121	40	482	10
11	280	121	41	485	10
12	334	67	42	488	10
13	361	42	43	492	10
14	376	31	44	495	10
15	387	24	45	499	11
16	396	20	46	503	11
17	403	18	47	507	11
18	408	16	48	512	12
19	414	15	49	517	12
20	418	14	50	522	13
21	423	13	51	528	13
22	427	12	52	534	14
23	430	11	53	541	16
24	434	11	54	550	17
25	437	10	55	561	20
26	440	10	56	576	24
27	443	10	57	597	31
28	446	10	58	634	47
29	449	9	59	650	56

* **Bold** represents SEM around cut score.

** A suppressed item in Reading grade 4 reduced the maximum possible score from 60 to 59. See Part 8 for more information.

Table 7-4
Scoring Table for Reading Grade 5

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	290	104	31	448	11
1	290	104	32	451	11
2	290	104	33	455	10
3	290	104	34	458	10
4	290	104	35	461	10
5	290	104	36	464	10
6	290	104	37	467	10
7	290	104	38	471	11
8	290	104	39	474	11
9	290	104	40	477	11
10	290	104	41	481	11
11	290	104	42	484	11
12	331	63	43	488	11
13	353	42	44	492	12
14	367	32	45	496	12
15	377	26	46	501	12
16	385	22	47	505	12
17	392	19	48	510	13
18	398	17	49	515	13
19	404	16	50	521	14
20	408	15	51	527	15
21	413	14	52	534	16
22	417	13	53	543	18
23	421	13	54	552	20
24	425	12	55	564	23
25	428	12	56	580	27
26	432	11	57	600	33
27	435	11	58	629	43
28	439	11	59	678	62
29	442	11	60	690	68
30	445	11			

* **Bold** represents SEM around cut score.

Table 7-5
Scoring Table for Reading Grade 6

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	300	95	31	461	11
1	300	95	32	464	11
2	300	95	33	468	11
3	300	95	34	471	11
4	300	95	35	475	11
5	300	95	36	478	11
6	300	95	37	482	11
7	300	95	38	485	11
8	300	95	39	489	11
9	300	95	40	493	11
10	300	95	41	497	11
11	300	95	42	501	12
12	300	95	43	505	12
13	333	62	44	509	12
14	358	41	45	513	12
15	374	32	46	517	12
16	385	27	47	522	12
17	395	22	48	527	13
18	402	20	49	532	13
19	409	18	50	537	13
20	415	16	51	543	14
21	420	15	52	549	15
22	425	14	53	557	16
23	430	13	54	565	17
24	434	13	55	575	19
25	438	12	56	588	23
26	442	12	57	605	28
27	446	11	58	629	36
28	450	11	59	671	54
29	453	11	60	730	93
30	457	11			

* **Bold** represents SEM around cut score.

Table 7-6
Scoring Table for Reading Grade 7

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	310	106	31	485	12
1	310	106	32	488	11
2	310	106	33	492	11
3	310	106	34	495	11
4	310	106	35	499	11
5	310	106	36	503	11
6	310	106	37	506	11
7	310	106	38	510	11
8	310	106	39	514	12
9	310	106	40	517	12
10	310	106	41	521	12
11	339	77	42	525	12
12	368	48	43	529	12
13	385	35	44	534	12
14	398	27	45	538	13
15	407	23	46	543	13
16	416	20	47	548	13
17	423	19	48	553	14
18	429	17	49	558	14
19	435	16	50	564	15
20	440	15	51	571	16
21	445	15	52	578	17
22	449	14	53	586	18
23	454	14	54	595	19
24	458	13	55	605	21
25	462	13	56	618	24
26	466	13	57	635	29
27	470	12	58	659	37
28	474	12	59	702	57
29	477	12	60	780	121
30	481	12			

* **Bold** represents SEM around cut score.

Table 7-7
Scoring Table for Reading Grade 8

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	330	91	31	482	12
1	330	91	32	486	12
2	330	91	33	490	12
3	330	91	34	493	12
4	330	91	35	497	12
5	330	91	36	501	12
6	330	91	37	505	12
7	330	91	38	509	12
8	330	91	39	513	12
9	330	91	40	517	13
10	330	91	41	521	13
11	330	91	42	526	13
12	330	91	43	530	13
13	336	85	44	535	14
14	369	53	45	540	14
15	388	39	46	545	14
16	401	31	47	550	14
17	412	25	48	556	15
18	420	22	49	561	15
19	428	20	50	568	16
20	434	18	51	574	16
21	440	17	52	582	17
22	445	16	53	590	19
23	450	15	54	600	20
24	455	14	55	611	23
25	459	13	56	625	26
26	463	13	57	643	31
27	467	13	58	670	41
28	471	12	59	717	63
29	475	12	60	790	116
30	479	12			

* **Bold** represents SEM around cut score.

Table 7-8
Scoring Table for Reading Grade 10

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	350	70	29	502	15
1	350	70	30	506	15
2	350	70	31	511	15
3	350	70	32	516	15
4	350	70	33	520	15
5	350	70	34	525	15
6	350	70	35	530	15
7	350	70	36	534	15
8	350	70	37	539	15
9	350	70	38	544	15
10	350	70	39	549	15
11	350	70	40	554	15
12	352	69	41	559	15
13	378	51	42	564	16
14	396	41	43	570	16
15	410	34	44	576	16
16	422	29	45	582	17
17	431	25	46	588	17
18	439	23	47	595	18
19	447	21	48	603	18
20	454	20	49	611	19
21	460	19	50	620	21
22	466	18	51	631	23
23	472	17	52	645	27
24	477	17	53	663	32
25	482	17	54	689	41
26	487	16	55	734	62
27	492	16	56	820	127
28	497	16			

* **Bold** represents SEM around cut score.

Table 7-9
Scoring Table for Mathematics Grade 3

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	220	80	29	388	10
1	220	80	30	391	10
2	220	80	31	395	10
3	220	80	32	398	10
4	220	80	33	401	10
5	220	80	34	404	10
6	220	80	35	408	10
7	220	80	36	411	10
8	220	80	37	414	10
9	220	80	38	418	10
10	243	60	39	421	10
11	275	41	40	424	10
12	294	32	41	428	11
13	307	27	42	432	11
14	317	23	43	435	11
15	325	21	44	439	11
16	333	19	45	443	11
17	339	17	46	448	11
18	345	16	47	452	12
19	350	15	48	457	12
20	355	14	49	462	12
21	359	13	50	467	13
22	363	13	51	474	14
23	367	12	52	481	15
24	371	12	53	489	17
25	375	11	54	500	20
26	378	11	55	516	26
27	382	11	56	542	39
28	385	11	57	630	116

* **Bold** represents SEM around cut score.

Table 7-10
Scoring Table for Mathematics Grade 4

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	240	116	29	427	11
1	240	116	30	431	11
2	240	116	31	434	10
3	240	116	32	437	10
4	240	116	33	440	10
5	240	116	34	443	10
6	240	116	35	447	10
7	240	116	36	450	10
8	240	116	37	453	10
9	240	116	38	456	10
10	271	85	39	459	10
11	307	50	40	463	10
12	327	37	41	466	10
13	341	30	42	470	10
14	352	25	43	473	11
15	361	22	44	477	11
16	368	20	45	481	11
17	375	19	46	486	12
18	381	17	47	490	12
19	387	16	48	495	13
20	392	15	49	501	14
21	397	14	50	507	15
22	401	14	51	514	16
23	405	13	52	522	17
24	409	13	53	532	20
25	413	12	54	545	24
26	417	12	55	563	30
27	420	12	56	595	45
28	424	11	57	650	86

* **Bold** represents SEM around cut score.

Table 7-11
Scoring Table for Mathematics Grade 5

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	270	80	32	461	13
1	270	80	33	465	12
2	270	80	34	469	12
3	270	80	35	472	12
4	270	80	36	476	12
5	270	80	37	479	12
6	270	80	38	483	11
7	270	80	39	486	11
8	270	80	40	490	11
9	270	80	41	493	11
10	270	80	42	496	11
11	270	80	43	500	11
12	294	63	44	504	11
13	324	45	45	507	11
14	343	37	46	511	11
15	358	31	47	515	11
16	370	28	48	518	12
17	380	25	49	523	12
18	388	23	50	527	12
19	396	22	51	531	12
20	403	20	52	536	13
21	410	19	53	541	13
22	416	18	54	547	14
23	421	17	55	553	15
24	426	16	56	560	16
25	431	16	57	568	18
26	436	15	58	578	20
27	441	15	59	590	23
28	445	14	60	608	29
29	449	14	61	640	44
30	453	13	62	680	73
31	457	13			

* **Bold** represents SEM around cut score.

Table 7-12
Scoring Table for Mathematics Grade 6

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	310	72	32	478	11
1	310	72	33	481	11
2	310	72	34	485	11
3	310	72	35	488	11
4	310	72	36	491	11
5	310	72	37	494	11
6	310	72	38	497	11
7	310	72	39	501	11
8	310	72	40	504	11
9	310	72	41	507	11
10	310	72	42	511	11
11	310	72	43	514	11
12	344	48	44	517	11
13	365	37	45	521	11
14	380	31	46	525	11
15	392	26	47	528	11
16	401	24	48	532	11
17	410	21	49	536	11
18	417	19	50	540	11
19	423	18	51	544	12
20	429	17	52	549	12
21	434	16	53	554	12
22	439	15	54	559	13
23	444	14	55	565	14
24	448	14	56	571	14
25	453	13	57	579	16
26	457	13	58	588	18
27	460	12	59	599	21
28	464	12	60	616	26
29	468	12	61	646	39
30	471	11	62	700	78
31	475	11			

* **Bold** represents SEM around cut score.

Table 7-13
Scoring Table for Mathematics Grade 7

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	330	105	32	510	10
1	330	105	33	513	10
2	330	105	34	515	10
3	330	105	35	518	10
4	330	105	36	521	10
5	330	105	37	524	10
6	330	105	38	527	10
7	330	105	39	530	10
8	330	105	40	533	10
9	330	105	41	535	10
10	359	76	42	538	10
11	390	47	43	541	10
12	408	35	44	545	10
13	421	28	45	548	10
14	431	24	46	551	10
15	439	21	47	554	10
16	447	19	48	558	11
17	453	17	49	562	11
18	458	16	50	566	11
19	464	15	51	570	12
20	468	14	52	574	12
21	473	13	53	579	12
22	477	13	54	585	13
23	481	12	55	590	14
24	484	12	56	597	15
25	488	12	57	605	16
26	491	11	58	615	18
27	494	11	59	628	22
28	498	11	60	646	27
29	501	11	61	676	39
30	504	11	62	710	58
31	507	10			

* **Bold** represents SEM around cut score.

Table 7-14
Scoring Table for Mathematics Grade 8

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	350	104	32	534	10
1	350	104	33	537	10
2	350	104	34	540	10
3	350	104	35	543	10
4	350	104	36	545	10
5	350	104	37	548	9
6	350	104	38	551	9
7	350	104	39	554	9
8	350	104	40	556	9
9	350	104	41	559	9
10	350	104	42	562	9
11	350	104	43	565	9
12	350	104	44	567	9
13	376	78	45	570	9
14	416	46	46	573	9
15	437	34	47	576	9
16	451	29	48	579	9
17	462	25	49	582	9
18	471	22	50	585	10
19	478	20	51	589	10
20	485	18	52	593	10
21	491	16	53	597	11
22	496	15	54	601	11
23	501	14	55	606	12
24	505	14	56	612	13
25	510	13	57	618	14
26	514	12	58	626	16
27	517	12	59	636	19
28	521	12	60	651	24
29	524	11	61	678	37
30	528	11	62	730	80
31	531	11			

* **Bold** represents SEM around cut score.

Table 7-15
Scoring Table for Mathematics Grade 10

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	410	87	29	556	10
1	410	87	30	559	10
2	410	87	31	562	9
3	410	87	32	565	9
4	410	87	33	567	9
5	410	87	34	570	9
6	410	87	35	573	9
7	410	87	36	575	9
8	410	87	37	578	9
9	410	87	38	581	9
10	410	87	39	583	9
11	410	87	40	586	9
12	445	53	41	589	9
13	469	34	42	592	9
14	483	27	43	595	9
15	493	23	44	598	9
16	502	19	45	601	9
17	508	18	46	604	9
18	514	16	47	608	10
19	520	15	48	611	10
20	524	14	49	616	11
21	529	13	50	620	11
22	533	12	51	626	12
23	537	12	52	633	14
24	540	11	53	641	16
25	544	11	54	653	20
26	547	10	55	674	29
27	550	10	56	750	96
28	553	10			

* **Bold** represents SEM around cut score.

** Two items were suppressed in Mathematics grade 10. This reduced the maximum possible score from 58 to 56. See Part 8 for more information.

Table 7-16
Scoring Table for Language Arts Grade 4

Raw Score	Scale Score	SEM
0	140	118
1	140	118
2	140	118
3	140	118
4	140	118
5	140	118
6	140	118
7	223	35
8	238	21
9	248	15
10	255	13
11	260	12
12	266	11
13	270	10
14	275	10
15	279	10
16	283	10
17	288	10
18	292	9
19	296	9
20	300	9
21	304	9
22	308	9
23	313	9
24	317	9
25	322	9
26	328	10
27	335	12
28	345	15
29	362	22
30	420	73

* **Bold** represents SEM around cut score.

Table 7-17
Scoring Table for Language Arts Grade 8

Raw Score	Scale Score	SEM
0	250	90
1	250	90
2	250	90
3	250	90
4	250	90
5	250	90
6	253	87
7	303	37
8	319	23
9	328	18
10	336	15
11	342	13
12	348	12
13	353	11
14	357	11
15	362	10
16	366	10
17	370	10
18	375	10
19	379	10
20	384	10
21	388	10
22	393	10
23	399	11
24	405	11
25	411	12
26	420	14
27	431	16
28	449	23
29	520	86

* Bold represents SEM around cut score.

** A suppressed item in Language Arts grade 8 reduced the maximum possible score from 30 to 29. See Part 8 for more information.

Table 7-18
Scoring Table for Language Arts Grade 10

Raw Score	Scale Score	SEM
0	290	63
1	290	63
2	290	63
3	290	63
4	290	63
5	290	63
6	290	63
7	330	31
8	349	23
9	361	21
10	371	19
11	379	18
12	387	17
13	394	16
14	400	15
15	406	14
16	411	14
17	416	13
18	421	13
19	426	12
20	431	12
21	436	12
22	440	12
23	445	11
24	449	11
25	454	12
26	458	12
27	463	12
28	468	12
29	474	13
30	480	13
31	486	14
32	493	15
33	501	16
34	511	17
35	523	20
36	538	23
37	560	29
38	597	42
39	630	57

* **Bold** represents SEM around cut score.

Table 7-19
Scoring Table for Social Studies Grade 4

Raw Score	Scale Score	SEM
0	170	75
1	170	75
2	170	75
3	170	75
4	170	75
5	170	75
6	170	75
7	170	75
8	210	35
9	224	21
10	233	15
11	239	12
12	244	11
13	248	10
14	252	9
15	256	8
16	259	8
17	262	8
18	265	8
19	268	7
20	271	7
21	274	7
22	276	7
23	279	7
24	282	7
25	285	7
26	288	7
27	291	7
28	294	7
29	297	8
30	301	8
31	306	9
32	311	10
33	318	12
34	328	15
35	346	24
36	400	70

* **Bold** represents SEM around cut score.

** Two items were suppressed in Social Studies grade 4. This reduced the maximum possible score from 38 to 36. See Part 8 for more information.

Table 7-20
Scoring Table for Social Studies Grade 8

Raw Score	Scale Score	SEM
0	230	98
1	230	98
2	230	98
3	230	98
4	230	98
5	230	98
6	230	98
7	230	98
8	230	98
9	230	98
10	277	51
11	299	30
12	311	22
13	320	17
14	327	15
15	333	14
16	338	13
17	343	12
18	348	11
19	352	11
20	357	11
21	361	11
22	365	11
23	369	11
24	373	11
25	378	11
26	382	11
27	386	11
28	391	11
29	395	11
30	400	11
31	405	11
32	410	11
33	416	12
34	422	13
35	429	14
36	437	15
37	447	17
38	461	20
39	485	30
40	530	66

* **Bold** represents SEM around cut score.

Table 7-21
Scoring Table for Social Studies Grade 10

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	240	125	26	423	11
1	240	125	27	427	11
2	240	125	28	431	11
3	240	125	29	435	11
4	240	125	30	438	11
5	240	125	31	442	11
6	240	125	32	446	11
7	240	125	33	450	11
8	240	125	34	453	11
9	240	125	35	457	11
10	240	125	36	461	11
11	240	125	37	466	11
12	307	58	38	470	11
13	335	35	39	474	11
14	351	27	40	479	11
15	363	23	41	484	11
16	372	21	42	489	11
17	380	18	43	494	12
18	386	17	44	500	12
19	392	15	45	506	13
20	398	14	46	514	14
21	403	13	47	523	16
22	407	13	48	537	20
23	412	12	49	563	34
24	416	12	50	620	85
25	420	11			

* **Bold** represents SEM around cut score.

Table 7-22
Scoring Table for Science Grade 4

Raw Score	Scale Score	SEM
0	170	69
1	170	69
2	170	69
3	170	69
4	170	69
5	170	69
6	170	69
7	170	69
8	170	69
9	170	69
10	199	40
11	216	26
12	227	20
13	235	17
14	242	15
15	247	14
16	252	12
17	257	11
18	261	11
19	265	10
20	268	10
21	272	9
22	275	9
23	278	9
24	281	8
25	284	8
26	287	8
27	290	8
28	293	8
29	297	8
30	300	9
31	304	9
32	308	9
33	312	10
34	317	10
35	323	11
36	330	13
37	338	14
38	350	18
39	372	28
40	440	88

* **Bold** represents SEM around cut score.

Table 7-23
Scoring Table for Science Grade 8

Raw Score	Scale Score	SEM
0	230	89
1	230	89
2	230	89
3	230	89
4	230	89
5	230	89
6	230	89
7	230	89
8	230	89
9	230	89
10	268	51
11	290	33
12	304	25
13	314	21
14	323	18
15	330	17
16	336	15
17	342	14
18	347	13
19	352	12
20	356	12
21	361	11
22	365	11
23	369	11
24	373	10
25	376	10
26	380	10
27	384	10
28	388	10
29	392	10
30	396	10
31	401	10
32	405	11
33	410	11
34	416	12
35	422	12
36	429	13
37	438	15
38	450	19
39	470	27
40	560	108

* **Bold** represents SEM around cut score.

Table 7-24
Scoring Table for Science Grade 10

Raw Score	Scale Score	SEM	Raw Score	Scale Score	SEM
0	240	155	26	435	11
1	240	155	27	438	11
2	240	155	28	441	10
3	240	155	29	445	10
4	240	155	30	448	10
5	240	155	31	451	10
6	240	155	32	454	10
7	240	155	33	457	10
8	240	155	34	461	10
9	240	155	35	464	10
10	240	155	36	468	10
11	305	90	37	471	10
12	344	51	38	475	11
13	362	35	39	479	11
14	374	27	40	483	11
15	383	22	41	488	12
16	391	19	42	493	12
17	397	17	43	498	13
18	402	16	44	504	14
19	407	15	45	511	15
20	412	14	46	520	17
21	416	13	47	530	19
22	420	12	48	545	23
23	424	12	49	569	33
24	428	11	50	610	62
25	431	11			

* **Bold** represents SEM around cut score.

Table 7-25
The Number of Students and Percents at LOSS and HOSS

Content	Grade	LOSS	N	Percent	HOSS	N	Percent
RD	3	270	669	1.13	640	33	0.06
	4	280	759	1.29	650	8	0.01
	5	290	453	0.76	690	4	0.01
	6	300	381	0.63	730	18	0.03
	7	310	224	0.37	780	6	0.01
	8	330	388	0.64	790	27	0.04
	10	350	838	1.32	820	66	0.10
MA	3	220	93	0.16	630	346	0.58
	4	240	62	0.10	650	250	0.42
	5	270	148	0.25	680	192	0.32
	6	310	96	0.16	700	28	0.05
	7	330	179	0.30	710	32	0.05
	8	350	377	0.62	730	164	0.27
	10	410	1588	2.50	750	260	0.41
LA	4	140	588	1.00	420	209	0.36
	8	250	526	0.87	520	2297	3.82
	10	290	375	0.59	630	63	0.10
SS	4	170	188	0.32	400	996	1.69
	8	230	419	0.70	530	647	1.07
	10	240	346	0.55	620	207	0.33
SC	4	170	208	0.35	440	388	0.66
	8	230	232	0.39	560	1329	2.21
	10	240	826	1.30	610	187	0.30

Table 8-1
Item Analysis Grade 3 Reading

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
1	MC	0.89	0.40	0.04%				
2	MC	0.93	0.47	0.08%				
3	MC	0.92	0.43	0.09%				
4	MC	0.73	0.33	0.26%				
5	MC	0.89	0.50	0.08%				
6	MC	0.92	0.40	0.14%				
7	MC	0.79	0.53	0.19%				
8	MC	0.87	0.52	0.47%				
9	MC	0.79	0.50	0.64%				
10	MC	0.72	0.47	1.35%				
11	MC	0.63	0.47	0.20%				
12	MC	0.79	0.48	0.38%				
13	MC	0.61	0.42	0.61%				
14	MC	0.83	0.53	0.62%				
15	MC	0.83	0.50	0.78%				
16	MC	0.83	0.47	0.89%				
17	MC	0.73	0.45	1.07%				
18	MC	0.66	0.24	1.18%				
19	MC	0.82	0.46	0.16%				
20	MC	0.82	0.49	0.24%				
21	MC	0.79	0.56	0.49%				
22	MC	0.74	0.54	0.94%				
23	MC	0.76	0.42	0.23%				
24	MC	0.77	0.52	0.69%				
25	MC	0.63	0.49	0.32%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-1 Cont'd
Item Analysis Grade 3 Reading

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
26	MC	0.79	0.45	0.83%				
27	MC	0.55	0.28	1.23%				
28	MC	0.59	0.44	0.20%				
29	MC	0.76	0.55	0.32%				
30	MC	0.53	0.34	0.80%				
31	CR	0.40	0.40	2.40%				
32	MC	0.46	0.42	2.04%		+		
33	MC	0.62	0.43	2.42%				
34	MC	0.72	0.50	2.95%				
35	MC	0.75	0.53	3.02%				
36	MC	0.57	0.49	3.52%				
37	MC	0.63	0.56	3.50%				
38	MC	0.77	0.52	3.82%				
39	MC	0.44	0.31	4.10%				
40	MC	0.88	0.49	0.13%				
41	MC	0.50	0.43	0.25%				
42	MC	0.72	0.50	0.40%				
43	MC	0.69	0.37	0.26%				
44	MC	0.68	0.45	0.34%				
45	MC	0.69	0.40	0.98%				
46	MC	0.77	0.47	3.53%				
47	MC	0.63	0.41	0.17%				
48	MC	0.53	0.34	0.27%				
49	MC	0.68	0.53	0.40%				
50	MC	0.66	0.54	0.23%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-1 Cont'd
Item Analysis Grade 3 Reading

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
51	MC	0.75	0.50	0.56%				
52	MC	0.64	0.47	0.56%				
53	MC	0.64	0.40	1.44%				
54	MC	0.77	0.47	0.29%				
55	MC	0.70	0.56	0.54%				
56	CR	0.33	0.47	1.81%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-2
Item Analysis Grade 4 Reading

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
1	MC	0.61	0.25	0.06%				
2	MC	0.79	0.36	0.08%				
3	MC	0.75	0.42	0.28%				
4	MC	0.69	0.34	0.10%				
5	MC	0.77	0.46	0.24%				
6	MC	0.38	0.36	0.40%				
7	MC	0.72	0.47	0.12%				
8	MC	0.53	0.31	0.22%				
9	MC	0.82	0.40	0.78%				
10	MC	0.55	0.23	0.18%		+		
11	MC	0.61	0.36	0.26%				
12	MC	0.71	0.32	0.55%				
13	CR	0.36	0.53	1.41%				
14	MC	0.73	0.31	1.04%				
15	MC	0.78	0.29	1.17%				
16	MC	0.76	0.50	1.32%				
17	MC	0.81	0.50	1.39%				
18	MC	0.90	0.41	1.50%				
19	MC	0.55	0.44	0.22%				
20	MC	0.68	0.34	0.22%				
21	MC	0.44	0.40	0.31%				
22	MC	0.48	0.42	0.26%				
23	MC	0.58	0.34	0.43%				
24	MC	0.74	0.56	0.33%				
25	MC	0.73	0.54	0.74%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-2 Cont'd
Item Analysis Grade 4 Reading

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
26	MC	0.66	0.41	0.20%				
27	MC	0.57	0.41	0.42%				
28	MC	0.64	0.36	1.19%				
29	MC	0.78	0.46	0.24%				
30	MC	0.51	0.40	0.40%				
31	MC	0.40	0.20	0.69%				
32 **	MC							
33	MC	0.69	0.48	0.79%				
34	MC	0.66	0.43	0.79%				
35	MC	0.75	0.46	1.12%				
36	MC	0.76	0.52	0.83%				
37	MC	0.81	0.45	0.91%				
38	MC	0.83	0.52	0.95%				
39	MC	0.84	0.35	0.22%				
40	MC	0.78	0.53	0.24%				
41	MC	0.78	0.59	0.32%				
42	MC	0.74	0.50	0.62%				
43	MC	0.89	0.49	0.96%				
44	MC	0.72	0.55	1.96%				
45	MC	0.63	0.39	0.26%				
46	MC	0.63	0.40	0.64%				
47	MC	0.78	0.53	0.42%				
48	MC	0.71	0.28	0.39%				
49	MC	0.61	0.35	0.37%				
50	MC	0.74	0.49	0.77%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

** Item dropped from scoring.

Table 8-2 Cont'd
Item Analysis Grade 4 Reading

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
51	MC	0.68	0.52	0.61%				
52	MC	0.70	0.42	1.38%				
53	MC	0.50	0.29	0.43%				
54	MC	0.58	0.34	0.66%				
55	MC	0.61	0.38	1.61%				
56	CR	0.34	0.47	2.39%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-3
Item Analysis Grade 5 Reading

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
1	MC	0.89	0.30	0.06%				
2	MC	0.76	0.46	0.05%				
3	MC	0.49	0.27	0.16%				
4	MC	0.91	0.41	0.47%				
5	MC	0.74	0.49	0.67%				
6	MC	0.86	0.49	0.08%				
7	MC	0.67	0.44	0.55%				
8	MC	0.59	0.36	0.15%				
9	MC	0.90	0.49	0.17%				
10	MC	0.82	0.46	0.52%				
11	MC	0.60	0.35	0.82%		+		
12	MC	0.74	0.50	1.33%				
13	MC	0.80	0.54	0.34%				
14	MC	0.78	0.49	0.51%				
15	MC	0.85	0.39	1.49%				
16	MC	0.84	0.46	1.72%				
17	MC	0.87	0.48	1.94%				
18	MC	0.81	0.33	2.23%				
19	MC	0.71	0.42	2.20%				
20	MC	0.76	0.46	2.90%				
21	MC	0.57	0.47	2.54%				
22	MC	0.77	0.39	0.12%				
23	MC	0.75	0.40	0.12%				
24	MC	0.76	0.52	0.16%				
25	MC	0.57	0.36	0.15%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-3 Cont'd
Item Analysis Grade 5 Reading

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
26	MC	0.66	0.40	0.48%				
27	MC	0.75	0.48	0.16%				
28	MC	0.75	0.53	0.34%				
29	MC	0.44	0.36	0.95%				
30	MC	0.57	0.31	1.61%				
31	MC	0.50	0.35	0.22%				
32	MC	0.82	0.32	1.30%				
33	CR	0.28	0.39	1.35%				+
34	MC	0.62	0.39	0.51%				
35	MC	0.79	0.29	0.62%				
36	MC	0.81	0.36	0.68%				
37	MC	0.61	0.34	0.65%				
38	MC	0.60	0.34	0.74%				
39	MC	0.60	0.37	0.13%				
40	MC	0.76	0.33	0.17%				
41	MC	0.57	0.26	0.24%				
42	MC	0.39	0.31	0.25%		+		
43	MC	0.81	0.39	0.22%				
44	MC	0.48	0.32	0.40%				
45	MC	0.84	0.42	0.25%				
46	MC	0.71	0.50	0.70%				
47	MC	0.83	0.48	0.75%				
48	MC	0.75	0.41	0.28%				
49	MC	0.70	0.45	0.35%				
50	CR	0.37	0.46	1.02%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-3 Cont'd
Item Analysis Grade 5 Reading

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
51	MC	0.59	0.34	0.68%				
52	MC	0.73	0.31	0.57%				
53	MC	0.80	0.41	1.14%				
54	MC	0.50	0.30	1.44%				
55	MC	0.65	0.43	0.77%				
56	MC	0.59	0.45	0.84%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-4
Item Analysis Grade 6 Reading

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
1	MC	0.75	0.31	0.07%				
2	MC	0.84	0.47	0.48%				
3	MC	0.60	0.32	0.08%				
4	MC	0.56	0.45	0.14%				
5	MC	0.91	0.45	0.16%				
6	MC	0.69	0.41	0.30%				
7	MC	0.77	0.40	0.87%				
8	MC	0.37	0.20	1.32%				
9	MC	0.58	0.22	0.16%				
10	MC	0.79	0.51	0.36%				
11	MC	0.55	0.39	0.32%				
12	MC	0.82	0.38	0.32%				
13	MC	0.76	0.35	0.45%				
14	MC	0.57	0.22	1.14%		+		
15	MC	0.45	0.36	1.39%				
16	MC	0.59	0.38	2.03%				
17	MC	0.79	0.49	0.50%				
18	MC	0.71	0.47	0.57%				
19	MC	0.75	0.27	0.08%				
20	MC	0.81	0.50	0.14%				
21	MC	0.84	0.41	0.26%				
22	MC	0.62	0.29	0.17%				
23	MC	0.85	0.37	0.16%				
24	MC	0.85	0.41	0.21%				
25	MC	0.67	0.37	0.23%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-4 Cont'd
Item Analysis Grade 6 Reading

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
26	MC	0.89	0.46	0.19%				
27	MC	0.79	0.45	0.33%				
28	MC	0.61	0.34	0.52%				
29	MC	0.82	0.55	0.26%				
30	MC	0.94	0.47	0.26%				
31	MC	0.80	0.43	0.27%				
32	MC	0.87	0.57	0.29%				
33	MC	0.89	0.52	0.37%				
34	MC	0.76	0.31	0.48%				
35	MC	0.85	0.50	0.30%				
36	MC	0.88	0.46	0.33%				
37	MC	0.80	0.32	0.34%				
38	CR	0.47	0.56	1.41%				
39	MC	0.62	0.18	0.16%		+		
40	MC	0.92	0.32	0.26%				
41	MC	0.86	0.44	0.52%				
42	MC	0.76	0.33	0.24%				
43	MC	0.88	0.32	0.27%				
44	MC	0.74	0.35	0.31%				
45	MC	0.52	0.33	2.48%				
46	MC	0.54	0.41	3.50%				
47	MC	0.73	0.46	0.24%				
48	MC	0.66	0.45	0.21%				
49	MC	0.54	0.43	0.22%				
50	MC	0.72	0.41	0.22%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-4 Cont'd
Item Analysis Grade 6 Reading

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
51	MC	0.64	0.43	0.32%				
52	MC	0.51	0.41	0.30%				
53	MC	0.76	0.51	0.95%				
54	MC	0.94	0.38	0.23%				
55	MC	0.43	0.32	0.46%				
56	CR	0.47	0.44	0.87%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-5
Item Analysis Grade 7 Reading

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
1	MC	0.64	0.58	0.07%				
2	MC	0.69	0.50	0.08%				
3	MC	0.81	0.44	0.07%				
4	MC	0.93	0.32	0.15%				
5	MC	0.74	0.48	0.15%				
6	MC	0.89	0.43	0.14%				
7	MC	0.39	0.24	0.19%				
8	MC	0.60	0.12	0.25%	+			
9	MC	0.43	0.41	0.10%				
10	MC	0.64	0.26	0.37%				
11	MC	0.95	0.37	0.15%				
12	MC	0.83	0.32	0.25%				
13	MC	0.67	0.38	0.60%				
14	MC	0.52	0.41	0.36%				
15	MC	0.71	0.42	0.45%				
16	MC	0.59	0.33	0.50%				
17	MC	0.62	0.43	0.54%				
18	MC	0.43	0.20	0.63%				
19	MC	0.77	0.31	0.64%				
20	MC	0.65	0.43	0.57%				
21	MC	0.38	0.31	2.03%				
22	MC	0.46	0.35	0.22%		+		
23	MC	0.82	0.28	0.54%				
24	MC	0.78	0.49	0.22%				
25	MC	0.57	0.34	0.60%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-5 Cont'd
Item Analysis Grade 7 Reading

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
26	MC	0.66	0.40	0.21%				
27	MC	0.85	0.43	0.30%				
28	MC	0.79	0.35	0.45%				
29	MC	0.49	0.31	1.97%				
30	MC	0.86	0.46	0.24%				
31	MC	0.74	0.35	0.45%				
32	MC	0.78	0.36	0.61%				
33	MC	0.65	0.30	0.34%				
34	MC	0.79	0.43	0.54%				
35	MC	0.46	0.31	0.82%				
36	CR	0.52	0.54	1.18%				
37	MC	0.83	0.49	0.13%				
38	MC	0.75	0.20	0.22%				
39	MC	0.83	0.48	0.20%				
40	MC	0.84	0.39	0.87%				
41	MC	0.94	0.38	0.17%				
42	MC	0.85	0.35	0.28%				
43	MC	0.89	0.43	0.87%				
44	MC	0.66	0.44	1.37%				
45	MC	0.74	0.55	1.86%				
46	MC	0.79	0.43	0.31%				
47	MC	0.52	0.30	1.87%				
48	MC	0.49	0.29	0.27%				
49	MC	0.66	0.32	0.89%				
50	MC	0.69	0.47	0.39%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-5 Cont'd
Item Analysis Grade 7 Reading

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
51	MC	0.73	0.31	0.71%				
52	MC	0.36	0.19	0.29%				
53	MC	0.61	0.38	0.37%				
54	MC	0.79	0.43	0.51%				
55	MC	0.65	0.38	1.23%				
56	CR	0.21	0.48	1.44%				+

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-6
Item Analysis Grade 8 Reading

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
1	MC	0.59	0.30	0.05%				
2	MC	0.80	0.36	0.07%				
3	MC	0.77	0.10	0.11%				
4	MC	0.75	0.39	0.15%				
5	MC	0.82	0.41	0.10%				
6	MC	0.72	0.32	0.18%				
7	MC	0.52	0.33	0.17%				
8	MC	0.60	0.36	0.32%				
9	MC	0.88	0.39	0.20%				
10	MC	0.52	0.17	0.80%				
11	MC	0.91	0.36	0.12%				
12	MC	0.85	0.43	0.17%				
13	MC	0.92	0.28	0.32%				
14	MC	0.88	0.41	0.48%				
15	MC	0.51	0.28	0.45%				
16	MC	0.84	0.39	1.14%				
17	MC	0.63	0.24	0.50%				
18	MC	0.59	0.19	0.54%				
19	CR	0.50	0.45	1.74%				
20	MC	0.88	0.36	0.16%				
21	MC	0.84	0.45	1.26%				
22	MC	0.75	0.46	0.16%				
23	MC	0.72	0.41	0.66%				
24	MC	0.52	0.26	0.28%				
25	MC	0.79	0.47	0.26%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-6 Cont'd
Item Analysis Grade 8 Reading

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
26	MC	0.78	0.34	0.22%				
27	MC	0.81	0.42	0.28%				
28	MC	0.62	0.32	0.19%				
29	MC	0.49	0.21	0.27%				
30	MC	0.41	0.29	0.38%				
31	MC	0.55	0.29	0.53%				
32	MC	0.54	0.33	0.25%				
33	MC	0.62	0.23	0.28%				
34	MC	0.82	0.44	0.38%				
35	MC	0.85	0.45	0.47%				
36	MC	0.77	0.41	0.40%				
37	MC	0.51	0.35	0.50%				
38	MC	0.66	0.15	0.57%	+	+		
39	MC	0.56	0.29	0.81%				
40	CR	0.42	0.43	2.19%				
41	MC	0.67	0.34	0.26%				
42	MC	0.74	0.28	0.31%				
43	MC	0.64	0.38	0.41%				
44	MC	0.75	0.38	0.79%				
45	MC	0.86	0.44	0.26%				
46	MC	0.40	0.35	0.37%				
47	MC	0.82	0.52	0.35%				
48	MC	0.56	0.43	0.58%				
49	MC	0.88	0.49	0.37%				
50	MC	0.82	0.54	0.37%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-6 Cont'd
Item Analysis Grade 8 Reading

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
51	MC	0.81	0.53	0.31%				
52	MC	0.44	0.26	0.45%				
53	MC	0.90	0.46	0.34%				
54	MC	0.59	0.26	0.47%				
55	MC	0.94	0.39	0.37%				
56	MC	0.81	0.47	0.46%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-7
Item Analysis Grade 10 Reading

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
1	MC	0.51	0.37	0.09%		+		
2	MC	0.95	0.25	0.02%				
3	MC	0.79	0.47	0.09%				
4	MC	0.38	0.44	0.18%				
5	MC	0.76	0.43	0.09%				
6	MC	0.93	0.44	0.10%				
7	MC	0.71	0.36	0.18%				
8	MC	0.82	0.52	1.95%				
9	MC	0.69	0.42	0.09%				
10	MC	0.68	0.25	0.25%		+		
11	MC	0.87	0.50	0.23%				
12	CR	0.39	0.50	2.30%				
13	MC	0.75	0.37	0.29%				
14	MC	0.67	0.45	0.38%				
15	MC	0.76	0.42	0.60%				
16	MC	0.83	0.48	0.40%				
17	MC	0.74	0.47	0.42%				
18	MC	0.61	0.37	0.24%				
19	MC	0.62	0.46	0.25%				
20	MC	0.82	0.47	0.18%				
21	MC	0.51	0.24	0.54%				
22	MC	0.65	0.43	0.21%				
23	MC	0.64	0.42	0.25%				
24	MC	0.79	0.46	0.27%				
25	MC	0.66	0.47	0.31%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-7 Cont'd
Item Analysis Grade 10 Reading

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
26	MC	0.78	0.42	0.74%				
27	MC	0.58	0.48	0.32%				
28	MC	0.85	0.35	0.41%				
29	MC	0.79	0.34	0.24%				
30	MC	0.65	0.31	0.29%				
31	MC	0.75	0.38	0.30%				
32	MC	0.57	0.35	0.33%				
33	MC	0.68	0.24	0.29%				
34	MC	0.75	0.44	0.28%				
35	MC	0.73	0.40	0.56%				
36	MC	0.75	0.41	0.42%				
37	MC	0.59	0.46	0.41%				
38	MC	0.78	0.47	0.41%				
39	MC	0.85	0.50	0.33%				
40	MC	0.86	0.45	2.49%				
41	MC	0.84	0.36	0.34%				
42	MC	0.68	0.40	0.37%				
43	CR	0.52	0.52	6.22%				+
44	MC	0.63	0.35	0.40%				
45	MC	0.81	0.36	0.42%				
46	MC	0.77	0.34	0.57%				
47	MC	0.46	0.27	0.57%				
48	MC	0.76	0.42	0.49%				
49	MC	0.77	0.39	0.55%				
50	MC	0.66	0.38	0.63%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-7 Cont'd
Item Analysis Grade 10 Reading

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
51	MC	0.47	0.33	0.55%				
52	MC	0.64	0.41	0.69%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-8
Item Analysis Grade 3 Mathematics

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
1	MC	0.91	0.28	0.26%				
2	MC	0.71	0.37	0.34%				
3	MC	0.79	0.46	0.18%				
4	MC	0.95	0.24	0.29%				
5	MC	0.52	0.46	0.25%				
6	MC	0.74	0.36	1.02%				
7	MC	0.88	0.36	0.64%				
8	MC	0.45	0.30	0.34%				
9	MC	0.69	0.41	0.87%				
10	CR	0.65	0.37	0.70%				
11	MC	0.81	0.54	1.24%				
12	MC	0.98	0.26	1.24%				
13	MC	0.85	0.47	1.51%				
14	MC	0.84	0.43	2.17%				
15	MC	0.76	0.49	0.13%				
16	MC	0.67	0.31	0.21%				
17	MC	0.74	0.46	0.35%				
18	MC	0.88	0.48	1.37%				
19	MC	0.86	0.47	0.20%				
20	MC	0.83	0.41	0.31%				
21	MC	0.85	0.39	1.05%				
22	MC	0.84	0.26	0.28%				
23	MC	0.48	0.37	0.35%				
24	MC	0.89	0.35	0.40%				
25A	CR	0.50	0.52	1.23%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-8 Cont'd
Item Analysis Grade 3 Mathematics

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
25B	CR	0.44	0.50	2.54%				
26	MC	0.85	0.48	0.79%				
27	MC	0.94	0.35	0.69%				
28A	CR	0.97	0.19	0.96%				
28B	CR	0.73	0.30	1.63%				
29	MC	0.54	0.46	1.10%				
30	MC	0.72	0.35	1.30%				
31	MC	0.84	0.49	0.15%				
32	MC	0.89	0.45	0.19%				
33	MC	0.62	0.46	0.35%				
34	MC	0.87	0.39	0.21%				
35	MC	0.60	0.33	0.85%				
36	MC	0.86	0.43	0.28%				
37	MC	0.85	0.37	0.33%				
38	MC	0.87	0.47	0.80%				
39	MC	0.63	0.43	1.51%				
40	MC	0.88	0.43	0.88%				
41	MC	0.75	0.30	0.37%				
42	MC	0.56	0.25	0.75%				
43	MC	0.91	0.42	1.09%				
44A	CR	0.71	0.52	1.18%				
44B	CR	0.32	0.43	2.61%				
45	MC	0.74	0.47	0.88%				
46	MC	0.76	0.46	0.80%				
47	MC	0.86	0.27	0.82%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-8 Cont'd
Item Analysis Grade 3 Mathematics

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
48	MC	0.58	0.46	0.82%				
49	MC	0.78	0.43	1.15%				
50	MC	0.74	0.50	1.10%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-9
Item Analysis Grade 4 Mathematics

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
1	MC	0.81	0.25	0.03%				
2	MC	0.85	0.36	0.42%				
3	MC	0.83	0.39	0.40%				
4	MC	0.71	0.45	0.08%				
5	MC	0.74	0.42	0.13%				
6	MC	0.82	0.32	0.35%				
7	MC	0.77	0.39	0.32%				
8	MC	0.90	0.39	0.23%				
9	MC	0.71	0.54	0.14%				
10	MC	0.77	0.51	0.76%				
11	MC	0.75	0.53	2.15%				
12	MC	0.94	0.24	0.35%				
13	CR	0.46	0.47	0.49%				
14	MC	0.74	0.48	0.77%				
15	MC	0.85	0.53	0.18%				
16	MC	0.94	0.29	0.26%				
17	MC	0.86	0.32	0.17%				
18	MC	0.78	0.23	0.21%				
19	MC	0.87	0.33	0.28%				
20A	CR	0.49	0.55	0.49%				
20B	CR	0.38	0.45	2.45%				
21	MC	0.65	0.42	0.24%				
22	MC	0.82	0.35	0.26%				
23	MC	0.80	0.38	0.33%				
24	MC	0.73	0.29	0.42%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-9 Cont'd
Item Analysis Grade 4 Mathematics

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
25	MC	0.82	0.51	0.75%				
26	MC	0.91	0.30	1.11%				
27	MC	0.81	0.39	1.12%				
28	MC	0.74	0.48	1.49%				
29A	CR	0.73	0.28	1.05%				
29B	CR	0.28	0.39	2.42%				+
30	MC	0.79	0.37	1.26%				
31	MC	0.77	0.25	0.16%				
32	MC	0.84	0.36	0.36%				
33	MC	0.76	0.29	0.92%				
34	MC	0.70	0.48	0.26%				
35	MC	0.85	0.48	0.28%				
36	MC	0.66	0.38	0.40%				
37	MC	0.67	0.42	3.01%				
38	MC	0.83	0.38	0.48%				
39	MC	0.54	0.36	0.41%				
40	MC	0.47	0.39	0.43%				
41A	CR	0.69	0.47	0.52%				
41B	CR	0.65	0.47	1.33%				
42	MC	0.74	0.37	0.24%				
43	MC	0.61	0.49	0.60%				
44	MC	0.72	0.39	0.76%				
45	MC	0.68	0.50	0.50%				
46	MC	0.91	0.42	0.58%				
47	MC	0.75	0.54	0.42%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-9 Cont'd
Item Analysis Grade 4 Mathematics

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
48	MC	0.96	0.22	0.47%				
49	MC	0.89	0.32	0.87%				
50	MC	0.91	0.33	0.53%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-10
Item Analysis Grade 5 Mathematics

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
1	MC	0.88	0.40	0.02%				
2	MC	0.45	0.47	0.22%				
3	MC	0.73	0.37	0.27%				
4	MC	0.69	0.37	0.39%				
5	MC	0.90	0.34	0.12%				
6	MC	0.64	0.32	0.46%				
7	MC	0.51	0.38	0.33%				
8	MC	0.46	0.51	0.38%				
9	MC	0.56	0.41	0.16%				
10	MC	0.77	0.44	0.20%				
11	MC	0.76	0.34	0.29%				
12A	CR	0.59	0.51	0.61%				
12B	CR	0.57	0.49	0.90%				
13	MC	0.62	0.50	0.89%				
14	MC	0.52	0.42	1.12%				
15	MC	0.75	0.23	0.13%				
16	MC	0.51	0.26	0.16%				
17	MC	0.44	0.35	0.15%				
18	MC	0.90	0.38	0.15%				
19	CR	0.80	0.29	0.43%				
20	MC	0.96	0.22	0.13%				
21	MC	0.48	0.41	0.33%		+		
22	MC	0.67	0.54	0.59%				
23A	CR	0.39	0.48	0.56%				
23B	CR	0.36	0.48	0.76%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-10 Cont'd
Item Analysis Grade 5 Mathematics

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
24	MC	0.40	0.36	0.35%				
25	MC	0.67	0.53	0.52%				
26	MC	0.58	0.38	0.23%				
27	MC	0.71	0.46	0.33%				
28	MC	0.95	0.25	0.44%				
29	MC	0.78	0.09	0.41%	+	+		
30	MC	0.47	0.42	0.43%				
31	MC	0.85	0.40	0.57%				
32	MC	0.74	0.37	0.61%				
33	MC	0.71	0.48	0.48%				
34	MC	0.84	0.47	0.73%				
35	MC	0.71	0.47	0.81%				
36	MC	0.84	0.39	0.19%				
37	MC	0.54	0.42	0.19%				
38	MC	0.71	0.41	0.26%				
39	MC	0.72	0.43	0.35%				
40	MC	0.70	0.32	0.73%				
41	MC	0.86	0.34	0.35%				
42	MC	0.71	0.35	0.31%				
43	MC	0.49	0.36	0.45%				
44	MC	0.65	0.43	0.36%				
45	MC	0.93	0.33	0.39%				
46A	CR	0.88	0.25	0.62%				
46B	CR	0.93	0.25	0.79%				
47	MC	0.85	0.39	0.56%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-10 Cont'd
Item Analysis Grade 5 Mathematics

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
48	MC	0.63	0.32	0.50%				
49	MC	0.85	0.28	1.48%				
50	MC	0.75	0.37	1.03%				
51	MC	0.53	0.30	0.72%				
52	MC	0.69	0.43	0.85%				
53	MC	0.81	0.40	0.68%				
54	MC	0.55	0.41	0.85%				
55	MC	0.77	0.38	0.76%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-11
Item Analysis Grade 6 Mathematics

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
1	MC	0.52	0.47	0.28%				
2	MC	0.83	0.41	0.14%				
3	MC	0.77	0.46	0.20%				
4	MC	0.92	0.41	0.34%				
5	MC	0.52	0.36	0.50%				
6	MC	0.86	0.32	0.60%				
7	MC	0.90	0.41	0.09%				
8	MC	0.75	0.42	0.25%				
9	MC	0.77	0.54	0.40%				
10A	CR	0.63	0.49	0.40%				
10B	CR	0.65	0.48	0.54%				
11	MC	0.96	0.12	0.20%	+			
12	MC	0.80	0.42	0.49%				
13	MC	0.91	0.27	0.42%				
14	MC	0.88	0.46	0.47%				
15	MC	0.93	0.29	0.46%				
16	MC	0.74	0.43	0.19%				
17	MC	0.52	0.39	0.24%				
18	MC	0.91	0.21	0.16%				
19	MC	0.84	0.43	0.89%				
20	MC	0.44	0.46	0.20%		+		
21	MC	0.79	0.45	0.49%				
22A	CR	0.93	0.34	0.33%				
22B	CR	0.26	0.43	0.76%				+
23	MC	0.75	0.39	0.37%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-11 Cont'd
Item Analysis Grade 6 Mathematics

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
24	MC	0.58	0.42	0.45%				
25	MC	0.82	0.42	0.37%				
26	MC	0.38	0.30	0.51%				
27	MC	0.82	0.49	0.36%				
28	MC	0.55	0.38	0.61%				
29	MC	0.77	0.44	0.77%				
30	MC	0.85	0.31	0.15%				
31	MC	0.66	0.48	0.18%				
32	MC	0.82	0.36	0.15%				
33	MC	0.58	0.33	0.19%				
34	MC	0.63	0.45	0.21%				
35A	CR	0.31	0.47	0.69%				
35B	CR	0.39	0.57	1.21%				
36	MC	0.97	0.18	0.36%				
37	MC	0.78	0.29	0.29%				
38	MC	0.96	0.28	0.33%				
39	MC	0.52	0.38	0.94%				
40	MC	0.71	0.49	0.23%				
41	MC	0.72	0.41	0.33%				
42	MC	0.60	0.42	0.50%				
43	MC	0.78	0.42	0.16%				
44	MC	0.37	0.42	0.20%				
45	MC	0.73	0.34	0.22%				
46	MC	0.80	0.27	0.40%				
47	MC	0.71	0.39	0.27%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-11 Cont'd
Item Analysis Grade 6 Mathematics

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
48	MC	0.84	0.36	0.31%				
49	MC	0.59	0.26	0.40%		+		
50	MC	0.70	0.37	0.32%				
51	MC	0.79	0.42	0.37%				
52	MC	0.80	0.22	0.45%				
53	CR	0.61	0.49	0.69%				
54	MC	0.68	0.40	0.56%				
55	MC	0.63	0.44	0.61%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-12
Item Analysis Grade 7 Mathematics

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
1	MC	0.61	0.48	0.16%				
2	MC	0.69	0.38	0.09%				
3	MC	0.90	0.34	0.08%				
4A	CR	0.70	0.50	0.71%				
4B	CR	0.63	0.58	1.58%				
5	MC	0.79	0.40	0.13%				
6	MC	0.65	0.24	0.10%				
7	MC	0.82	0.45	0.17%				
8	MC	0.74	0.43	0.25%				
9	MC	0.48	0.37	0.44%				
10	MC	0.84	0.35	0.52%				
11	MC	0.82	0.30	0.32%				
12	MC	0.72	0.40	9.08%				
13	MC	0.73	0.41	0.51%				
14	MC	0.60	0.35	0.74%				
15	MC	0.61	0.43	0.80%				
16	MC	0.65	0.47	0.32%				
17	MC	0.75	0.29	0.32%				
18	MC	0.93	0.17	0.39%				
19	MC	0.41	0.26	0.83%				
20	MC	0.77	0.32	0.20%				
21	MC	0.63	0.40	0.26%				
22	MC	0.74	0.35	0.36%				
23	MC	0.88	0.38	0.33%				
24	MC	0.74	0.43	0.33%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-12 Cont'd
Item Analysis Grade 7 Mathematics

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
25	MC	0.59	0.48	0.23%				
26	MC	0.90	0.37	0.49%				
27	MC	0.88	0.48	0.21%				
28	MC	0.85	0.30	0.29%				
29A	CR	0.52	0.36	1.32%				
29B	CR	0.65	0.52	1.93%				
30	MC	0.75	0.36	0.17%				
31	MC	0.72	0.50	1.37%				
32A	CR	0.49	0.46	0.75%				
32B	CR	0.35	0.49	2.04%				
33	MC	0.48	0.41	0.75%				
34	MC	0.68	0.44	0.25%				
35	MC	0.68	0.39	0.25%				
36	MC	0.83	0.39	0.38%				
37	MC	0.83	0.46	0.37%				
38	MC	0.86	0.23	0.26%				
39	MC	0.60	0.48	0.30%				
40	MC	0.46	0.31	0.23%				
41	MC	0.45	0.38	0.41%				
42	MC	0.53	0.43	0.31%				
43	MC	0.89	0.34	0.22%				
44	MC	0.40	0.39	0.29%				
45	MC	0.73	0.28	0.25%				
46	MC	0.58	0.46	0.64%				
47	MC	0.88	0.34	0.74%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-12 Cont'd
Item Analysis Grade 7 Mathematics

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
48	MC	0.83	0.31	0.24%				
49	MC	0.54	0.23	0.37%				
50	MC	0.71	0.42	0.41%				
51	CR	0.41	0.58	0.88%				
52	MC	0.71	0.45	0.43%				
53	MC	0.77	0.45	0.32%				
54	MC	0.62	0.47	0.41%				
55	MC	0.52	0.43	0.31%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-13
Item Analysis Grade 8 Mathematics

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
1	MC	0.53	0.36	0.20%				
2	MC	0.72	0.39	0.05%				
3	MC	0.52	0.44	0.15%				
4	MC	0.71	0.48	0.12%				
5	MC	0.70	0.44	0.42%				
6	MC	0.54	0.29	0.12%				
7	MC	0.67	0.43	0.24%				
8	MC	0.74	0.40	0.24%				
9A	CR	0.90	0.36	0.46%				
9B	CR	0.76	0.45	0.90%				
10	MC	0.80	0.28	0.10%				
11	MC	0.82	0.43	0.21%				
12	MC	0.65	0.42	0.33%				
13	MC	0.60	0.46	0.44%				
14	MC	0.80	0.42	0.64%				
15	MC	0.47	0.38	1.03%				
16	MC	0.90	0.30	0.16%				
17	MC	0.55	0.52	0.14%				
18	MC	0.70	0.29	0.22%				
19	MC	0.39	0.28	0.25%				
20A	CR	0.45	0.57	1.46%				
20B	CR	0.55	0.64	2.21%				
21	MC	0.77	0.37	0.23%				
22	MC	0.49	0.23	0.44%				
23	MC	0.48	0.31	0.25%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-13 Cont'd
Item Analysis Grade 8 Mathematics

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
24	MC	0.63	0.36	0.30%				
25	MC	0.47	0.37	0.97%				
26	MC	0.70	0.48	0.87%				
27	MC	0.53	0.45	0.32%				
28	MC	0.70	0.45	0.44%				
29	MC	0.91	0.33	0.31%				
30	MC	0.56	0.36	0.48%				
31	MC	0.63	0.34	0.25%				
32	MC	0.65	0.27	0.38%				
33	MC	0.50	0.35	0.73%				
34	MC	0.52	0.38	0.92%				
35	MC	0.42	0.23	0.59%				
36	MC	0.52	0.43	0.44%				
37	MC	0.84	0.38	0.43%				
38	MC	0.54	0.53	0.66%				
39	MC	0.68	0.35	0.91%				
40A	CR	0.48	0.62	1.00%				
40B	CR	0.47	0.61	1.86%				
41	MC	0.53	0.35	0.66%				
42	MC	0.84	0.45	0.69%				
43	MC	0.47	0.31	0.43%				
44	MC	0.68	0.56	0.30%				
45	MC	0.48	0.45	0.35%				
46	MC	0.61	0.35	0.51%				
47	MC	0.40	0.59	0.32%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-13 Cont'd
Item Analysis Grade 8 Mathematics

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
48	MC	0.91	0.20	0.31%				
49	MC	0.64	0.45	0.64%				
50	MC	0.68	0.34	0.44%				
51	MC	0.45	0.31	0.91%				
52	MC	0.84	0.43	1.77%				
53	CR	0.30	0.58	3.02%				
54	MC	0.68	0.33	0.77%				
55	MC	0.61	0.24	0.83%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-14
Item Analysis Grade 10 Mathematics

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
1	MC	0.46	0.44	0.14%				
2	MC	0.64	0.32	0.19%				
3	MC	0.56	0.43	0.09%				
4	MC	0.53	0.40	0.11%				
5	MC	0.50	0.42	0.13%				
6 **	MC							
7	MC	0.68	0.48	0.16%				
8	MC	0.64	0.42	0.14%				
9	MC	0.51	0.53	0.14%				
10	MC	0.53	0.37	0.21%				
11	MC	0.47	0.43	0.17%				
12	MC	0.53	0.38	0.35%				
13	MC	0.50	0.50	0.54%				
14	MC	0.72	0.42	0.26%				
15	MC	0.84	0.42	0.21%				
16	MC	0.45	0.40	0.29%				
17	MC	0.53	0.53	0.50%				
18	MC	0.79	0.30	0.36%				
19	MC	0.71	0.32	0.19%				
20	MC	0.50	0.41	0.36%				
21	MC	0.72	0.53	0.32%				
22	MC	0.59	0.44	0.97%				
23	MC	0.73	0.47	0.50%				
24	MC	0.72	0.42	0.51%				
25	MC	0.58	0.36	0.62%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

** Item dropped from scoring.

Table 8-14 Cont'd
Item Analysis Grade 10 Mathematics

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
26 **	MC							
27	CR	0.43	0.63	5.10%			+	
28	MC	0.65	0.33	0.29%				
29	MC	0.83	0.34	0.22%				
30	MC	0.73	0.31	0.25%				
31	MC	0.76	0.44	0.53%				
32	MC	0.74	0.38	0.36%				
33	CR	0.68	0.44	2.41%				
34	MC	0.51	0.32	0.36%				
35	MC	0.68	0.27	0.31%				
36	MC	0.61	0.22	0.41%				
37	MC	0.50	0.48	0.38%				
38	CR	0.27	0.59	8.84%			+	+
39	MC	0.80	0.45	0.58%				
40	MC	0.35	0.41	0.34%				
41	MC	0.52	0.41	0.43%				
42	MC	0.32	0.31	0.83%				
43	MC	0.57	0.42	0.43%				
44	MC	0.53	0.52	0.39%				
45	MC	0.55	0.44	0.62%				
46	MC	0.60	0.42	0.60%				
47	MC	0.72	0.48	0.72%				
48	MC	0.59	0.43	0.40%				
49	MC	0.66	0.42	0.71%				
50	MC	0.73	0.48	0.41%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

** Item dropped from scoring.

Table 8-14 Cont'd
Item Analysis Grade 10 Mathematics

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
51	MC	0.63	0.58	0.70%				
52	CR	0.47	0.55	4.19%				
53	MC	0.67	0.42	0.66%				
54	MC	0.74	0.53	0.67%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-15
Item Analysis Grade 4 Language Arts

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
1	MC	0.48	0.34	0.13%				
2	MC	0.91	0.37	0.05%				
3	MC	0.78	0.31	0.15%				
4	MC	0.41	0.34	0.27%				
5	MC	0.67	0.28	0.04%				
6	MC	0.57	0.24	0.29%				
7	MC	0.82	0.33	0.16%				
8	MC	0.73	0.26	0.50%				
9	MC	0.89	0.33	0.23%				
10	MC	0.38	0.27	0.28%				
11	MC	0.85	0.32	0.19%				
12	MC	0.78	0.36	0.23%				
13	MC	0.38	0.28	0.46%				
14	MC	0.75	0.36	0.83%				
15	MC	0.65	0.38	0.46%				
16	MC	0.70	0.46	0.27%				
17	MC	0.50	0.34	0.85%				
18	MC	0.46	0.30	1.44%				
19	MC	0.65	0.36	0.52%				
20	MC	0.70	0.41	0.66%				
21	MC	0.50	0.26	1.10%				
22	MC	0.55	0.31	0.88%				
23	MC	0.42	0.21	1.13%				
24	MC	0.81	0.38	1.09%				
25	MC	0.85	0.38	1.25%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-15 Cont'd
Item Analysis Grade 4 Language Arts

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
26	MC	0.59	0.32	1.35%		+		
27	MC	0.55	0.31	3.78%				
28	MC	0.52	0.34	1.86%				
29	MC	0.70	0.50	2.35%				
30	MC	0.35	0.26	2.32%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-16
Item Analysis Grade 8 Language Arts

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
1	MC	0.77	0.39	0.19%				
2	MC	0.34	0.30	0.21%				
3	MC	0.83	0.36	0.13%				
4	MC	0.78	0.43	0.27%				
5	MC	0.79	0.41	0.12%				
6	MC	0.81	0.49	0.14%				
7	MC	0.92	0.42	0.44%				
8	MC	0.70	0.37	0.20%				
9	MC	0.66	0.43	0.21%				
10	MC	0.92	0.35	0.80%				
11	MC	0.83	0.41	0.22%				
12	MC	0.56	0.41	0.26%				
13	MC	0.87	0.34	0.22%				
14	MC	0.80	0.50	0.30%				
15	MC	0.63	0.40	0.25%				
16	MC	0.89	0.45	0.39%				
17	MC	0.83	0.44	0.28%				
18	MC	0.57	0.49	0.80%				
19	MC	0.92	0.44	0.37%				
20	MC	0.81	0.50	0.42%				
21	MC	0.77	0.51	0.53%				
22	MC	0.86	0.41	0.68%				
23	MC	0.78	0.50	0.55%				
24	MC	0.62	0.48	1.56%				
25	MC	0.71	0.41	0.84%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-16 Cont'd
Item Analysis Grade 8 Language Arts

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
26	MC	0.53	0.29	1.13%				
27	MC	0.66	0.35	1.29%				
28	MC	0.71	0.36	1.59%				
29	MC	0.68	0.45	1.68%				
30 **	MC	.	.					

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

** Item dropped from scoring.

Table 8-17
Item Analysis Grade 10 Language Arts

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
1	MC	0.44	0.31	0.10%				
1A	CR	0.60	0.54	1.06%				
1B	CR	0.73	0.41	1.06%				
2	MC	0.85	0.39	0.14%				
3	MC	0.60	0.43	0.25%				
4	MC	0.59	0.29	0.39%				
5	MC	0.58	0.29	0.19%				
6	MC	0.81	0.40	0.48%				
7	MC	0.67	0.42	0.26%				
8	MC	0.62	0.51	0.24%				
9	MC	0.58	0.30	1.03%				
10	MC	0.64	0.51	0.35%				
11	MC	0.78	0.43	0.47%				
12	MC	0.75	0.38	0.23%				
13	MC	0.54	0.39	0.32%				
14	MC	0.65	0.34	0.28%				
15	MC	0.42	0.23	0.38%				
16	MC	0.62	0.43	0.26%				
17	MC	0.56	0.22	0.36%				
18	MC	0.51	0.42	0.26%				
19	MC	0.49	0.36	0.36%				
20	MC	0.59	0.30	0.48%				
21	MC	0.63	0.41	1.88%				
22	MC	0.61	0.43	0.73%				
23	MC	0.58	0.24	0.94%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-17 Cont'd
Item Analysis Grade 10 Language Arts

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
24	MC	0.58	0.37	1.35%				
25	MC	0.51	0.35	1.55%				
26	MC	0.54	0.36	1.62%				
27	MC	0.49	0.34	2.15%				
28	MC	0.61	0.47	2.29%				
29	MC	0.55	0.51	2.26%				
30	MC	0.79	0.48	2.63%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

** Writing prompt items are included here. The Writing raw score contributes to the scale score for Language Arts in grade 10.

Table 8-18
Item Analysis Grade 4 Social Studies

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
1	MC	0.82	0.33	0.13%				
2	MC	0.87	0.32	2.04%				
3	MC	0.80	0.37	0.16%				
4 **	MC							
5	MC	0.82	0.46	0.22%				
6	MC	0.95	0.30	2.18%				
7	MC	0.83	0.29	0.27%				
8	MC	0.87	0.34	0.69%				
9 **	MC							
10	MC	0.95	0.38	0.52%				
11	MC	0.79	0.27	0.32%				
12	MC	0.94	0.27	0.72%				
13	MC	0.89	0.40	0.73%				
14	MC	0.66	0.27	0.60%				
15	MC	0.92	0.32	0.18%				
16	MC	0.97	0.30	0.53%				
17	MC	0.72	0.47	0.37%				
18	MC	0.68	0.42	1.66%				
19	MC	0.67	0.46	0.33%				
20	MC	0.90	0.36	0.24%				
21	MC	0.78	0.41	0.51%				
22	MC	0.86	0.37	0.33%				
23	MC	0.62	0.35	2.26%				
24	MC	0.82	0.46	0.40%				
25	MC	0.75	0.29	0.69%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-18 Cont'd
Item Analysis Grade 4 Social Studies

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
26	MC	0.80	0.38	0.34%				
27	MC	0.78	0.45	1.54%				
28	MC	0.71	0.26	0.42%				
29	MC	0.47	0.18	0.70%				
30	MC	0.71	0.42	0.53%				
31	MC	0.91	0.45	0.66%				
32	MC	0.63	0.43	1.28%				
33	MC	0.49	0.29	1.90%		+		
34	MC	0.79	0.38	0.77%				
35	MC	0.52	0.26	1.16%				
36	MC	0.75	0.47	1.09%				
37	MC	0.79	0.27	0.73%				
38	MC	0.57	0.35	1.05%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-19
Item Analysis Grade 8 Social Studies

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
1	MC	0.87	0.35	0.09%				
2	MC	0.91	0.40	0.08%				
3	MC	0.89	0.45	0.19%				
4	MC	0.83	0.44	0.24%				
5	MC	0.88	0.35	0.19%				
6	MC	0.81	0.40	1.39%				
7	MC	0.34	0.22	0.19%				
8	MC	0.85	0.35	0.17%				
9	MC	0.73	0.46	0.34%				
10	MC	0.95	0.33	0.15%				
11	MC	0.81	0.43	0.29%				
12	MC	0.83	0.46	0.19%				
13	MC	0.87	0.50	2.30%				
14	MC	0.92	0.42	0.22%				
15	MC	0.73	0.42	0.99%				
16	MC	0.74	0.39	0.49%				
17	MC	0.73	0.39	0.19%				
18	MC	0.63	0.20	0.29%		+		
19	MC	0.83	0.35	0.40%				
20	MC	0.60	0.47	0.59%				
21	MC	0.49	0.27	0.88%				
22	MC	0.74	0.35	0.30%				
23	MC	0.84	0.46	0.32%				
24	MC	0.64	0.37	0.46%				
25	MC	0.85	0.44	0.88%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-19 Cont'd
Item Analysis Grade 8 Social Studies

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
26	MC	0.61	0.51	1.06%				
27	MC	0.77	0.42	1.16%				
28	MC	0.65	0.41	0.41%				
29	MC	0.55	0.48	0.49%				
30	MC	0.76	0.45	0.82%				
31	MC	0.76	0.48	0.52%				
32	MC	0.85	0.44	0.81%				
33	MC	0.58	0.34	0.64%				
34	MC	0.53	0.22	1.29%				
35	MC	0.56	0.33	0.68%				
36	MC	0.65	0.19	0.80%				
37	MC	0.60	0.33	0.93%				
38	MC	0.44	0.28	0.97%				
39	MC	0.63	0.40	1.70%				
40	MC	0.36	0.33	1.04%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-20
Item Analysis Grade 10 Social Studies

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
1	MC	0.64	0.45	0.14%				
2	MC	0.67	0.42	0.08%				
3	MC	0.39	0.18	0.15%				
4	MC	0.49	0.16	0.22%				
5	MC	0.57	0.33	0.34%				
6	MC	0.77	0.45	0.30%				
7	MC	0.59	0.22	0.24%				
8	MC	0.68	0.33	0.16%				
9	MC	0.73	0.36	0.18%				
10	MC	0.74	0.45	0.18%				
11	MC	0.76	0.34	0.20%				
12	MC	0.83	0.42	0.41%				
13	MC	0.96	0.32	0.16%				
14	MC	0.73	0.37	0.24%				
15	MC	0.80	0.41	0.20%				
16	MC	0.22	0.12	0.21%	+	+		+
17	MC	0.74	0.38	0.24%				
18	MC	0.70	0.49	0.41%				
19	MC	0.40	0.25	0.32%				
20	MC	0.69	0.39	0.91%				
21	MC	0.52	0.42	0.55%				
22	MC	0.35	0.40	0.71%				
23	MC	0.68	0.44	0.71%				
24	MC	0.75	0.52	0.94%				
25	MC	0.64	0.39	1.06%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-20 Cont'd
Item Analysis Grade 10 Social Studies

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
26	MC	0.70	0.44	0.36%				
27	MC	0.51	0.35	0.35%				
28	MC	0.53	0.39	0.27%				
29	MC	0.74	0.42	0.41%				
30	MC	0.56	0.40	0.38%				
31	MC	0.85	0.50	1.16%				
32	MC	0.52	0.23	0.30%				
33	MC	0.53	0.26	0.68%				
34	MC	0.64	0.41	0.59%				
35	MC	0.78	0.54	0.38%				
36	MC	0.61	0.37	1.55%				
37	MC	0.63	0.45	0.38%				
38	MC	0.69	0.39	0.40%				
39	MC	0.98	0.22	0.31%				
40	MC	0.45	0.30	0.43%				
41	MC	0.58	0.23	0.42%				
42	MC	0.83	0.42	0.36%				
43	MC	0.67	0.37	0.50%				
44	MC	0.79	0.38	0.42%				
45	MC	0.81	0.51	0.52%				
46	MC	0.66	0.39	0.47%				
47	MC	0.48	0.28	0.58%				
48	MC	0.60	0.37	0.90%				
49	MC	0.65	0.46	1.03%				
50	MC	0.74	0.49	0.97%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

** Item dropped from scoring.

Table 8-21
Item Analysis Grade 4 Science

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
1	MC	0.70	0.39	0.07%				
2	MC	0.48	0.29	0.11%				
3	MC	0.73	0.45	0.21%				
4	MC	0.83	0.33	0.04%				
5	MC	0.95	0.32	0.16%				
6	MC	0.83	0.32	0.16%				
7	MC	0.94	0.33	0.30%				
8	MC	0.80	0.42	2.02%				
9	MC	0.77	0.23	0.21%				
10	MC	0.92	0.20	0.20%				
11	MC	0.78	0.41	0.24%				
12	MC	0.84	0.31	0.29%				
13	MC	0.69	0.46	0.31%				
14	MC	0.94	0.19	0.36%				
15	MC	0.60	0.37	1.11%				
16	MC	0.70	0.34	1.58%				
17	MC	0.36	0.21	0.35%				
18	MC	0.71	0.44	0.59%				
19	MC	0.80	0.30	0.58%				
20	MC	0.66	0.36	0.71%				
21	MC	0.80	0.46	0.28%				
22	MC	0.75	0.33	0.40%				
23	MC	0.63	0.36	0.30%				
24	MC	0.85	0.32	0.38%				
25	MC	0.63	0.36	0.64%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-21 Cont'd
Item Analysis Grade 4 Science

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
26	MC	0.78	0.40	0.40%				
27	MC	0.54	0.36	0.43%				
28	MC	0.79	0.47	0.60%				
29	MC	0.79	0.46	0.72%				
30	MC	0.84	0.28	1.69%				
31	MC	0.58	0.32	0.73%				
32	MC	0.67	0.35	0.85%				
33	MC	0.40	0.32	1.69%				
34	MC	0.81	0.43	0.65%				
35	MC	0.65	0.45	1.11%				
36	MC	0.75	0.46	0.89%				
37	MC	0.81	0.46	1.71%				
38	MC	0.65	0.42	1.17%				
39	MC	0.74	0.46	1.41%				
40	MC	0.47	0.16	2.39%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-22
Item Analysis Grade 8 Science

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
1	MC	0.87	0.36	0.02%				
2	MC	0.92	0.41	0.06%				
3	MC	0.82	0.34	0.05%				
4	MC	0.91	0.30	0.03%				
5	MC	0.77	0.41	0.12%				
6	MC	0.77	0.30	0.17%				
7	MC	0.94	0.34	0.16%				
8	MC	0.86	0.43	0.18%				
9	MC	0.87	0.38	0.60%				
10	MC	0.96	0.32	0.63%				
11	MC	0.86	0.47	0.86%				
12	MC	0.84	0.48	0.17%				
13	MC	0.79	0.48	0.23%				
14	MC	0.89	0.47	0.37%				
15	MC	0.88	0.32	0.22%				
16	MC	0.93	0.22	0.24%				
17	MC	0.56	0.28	0.36%				
18	MC	0.70	0.44	0.27%				
19	MC	0.72	0.39	0.31%				
20	MC	0.59	0.35	2.26%		+		
21	MC	0.73	0.30	0.44%				
22	MC	0.74	0.33	0.28%				
23	MC	0.74	0.40	0.16%				
24	MC	0.91	0.24	0.14%				
25	MC	0.75	0.49	0.24%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-22 Cont'd
Item Analysis Grade 8 Science

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
26	MC	0.66	0.48	0.34%				
27	MC	0.66	0.43	0.48%				
28	MC	0.46	0.36	0.40%				
29	MC	0.83	0.47	0.52%				
30	MC	0.64	0.34	0.34%				
31	MC	0.62	0.50	0.20%				
32	MC	0.76	0.27	0.28%				
33	MC	0.76	0.40	0.35%				
34	MC	0.47	0.28	0.47%				
35	MC	0.61	0.36	0.38%				
36	MC	0.62	0.36	0.62%				
37	MC	0.77	0.43	0.41%				
38	MC	0.66	0.35	0.69%				
39	MC	0.86	0.48	0.40%				
40	MC	0.82	0.24	0.75%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-23
Item Analysis Grade 10 Science

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
1	MC	0.68	0.48	0.08%				
2	MC	0.72	0.43	0.08%				
3	MC	0.66	0.35	0.10%				
4	MC	0.60	0.41	0.09%				
5	MC	0.81	0.42	0.09%				
6	MC	0.72	0.35	0.14%				
7	MC	0.84	0.32	0.14%				
8	MC	0.82	0.38	0.22%				
9	MC	0.61	0.35	0.29%				
10	MC	0.78	0.45	0.19%				
11	MC	0.67	0.33	0.21%				
12	MC	0.50	0.27	0.29%				
13	MC	0.74	0.46	0.59%				
14	MC	0.60	0.46	0.69%				
15	MC	0.72	0.42	0.61%				
16	MC	0.60	0.37	0.15%				
17	MC	0.53	0.47	0.30%				
18	MC	0.78	0.49	0.29%				
19	MC	0.52	0.34	0.20%				
20	MC	0.52	0.32	0.25%				
21	MC	0.65	0.49	2.52%				
22	MC	0.57	0.36	0.24%				
23	MC	0.69	0.48	0.34%				
24	MC	0.55	0.45	0.49%				
25	MC	0.57	0.37	0.39%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-23 Cont'd
Item Analysis Grade 10 Science

Test Book Item	Item Statistic Fields				Flag			
	Item Type	<i>p</i> -value	Corr	Omit Rate	Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
26	MC	0.51	0.40	0.21%				
27	MC	0.77	0.54	0.21%				
28	MC	0.71	0.45	0.22%				
29	MC	0.66	0.47	0.43%				
30	MC	0.76	0.30	0.29%				
31	MC	0.40	0.26	0.36%				
32	MC	0.46	0.21	0.38%				
33	MC	0.47	0.38	0.89%				
34	MC	0.57	0.44	0.84%				
35	MC	0.61	0.31	0.36%				
36	MC	0.65	0.44	0.42%				
37	MC	0.61	0.41	0.60%				
38	MC	0.69	0.45	0.44%				
39	MC	0.71	0.41	0.54%				
40	MC	0.76	0.45	0.32%				
41	MC	0.57	0.36	0.37%				
42	MC	0.58	0.35	0.44%				
43	MC	0.41	0.39	0.56%				
44	MC	0.43	0.33	0.45%				
45	MC	0.38	0.24	0.64%				
46	MC	0.78	0.34	0.43%				
47	MC	0.60	0.37	0.51%				
48	MC	0.46	0.40	0.53%				
49	MC	0.70	0.44	0.56%				
50	MC	0.73	0.42	0.60%				

* Note: The correlation is flagged when it falls below 0.15, the distractor is flagged when it has a positive correlation with the correct answer, the omit rate is flagged when it is above 5%, the *p*-value is flagged when it is below 0.30.

Table 8-24
The Number of Items Flagged

Content	Grade	OP Items			
		Flag Corr	Flag Distractor	Flag Omit	Flag <i>p</i> -value
RD	3		1		
	4		1		
	5		2		1
	6		2		
	7	1	1		1
	8	2	2		
	10		2	1	
MA	3				
	4				1
	5	1	2		
	6	1	2		1
	7			1	
	8				
	10			2	1
LA	4		1		
	8				
	10		1		
SS	4		1		
	8		1		
	10				
SC	4		1		
	8		1		
	10	1	1		1

Table 8-25
Raw Score Descriptive Statistics based on Census Data

Content	Grade	N Count	Mean	Test Difficulty	SD	Skewness	Kurtosis	Min Obtained	Max Obtained	Max Possible	Alpha	SEM
Reading	3	59004	40.59	0.68	12.14	-0.73	-0.42	1	60	60	0.94	3.03
	4	58867	37.89	0.64	11.52	-0.52	-0.62	1	59	59	0.92	3.18
	5	59811	39.68	0.66	10.96	-0.67	-0.32	2	60	60	0.92	3.10
	6	60683	41.79	0.70	10.55	-0.84	0.06	2	60	60	0.92	3.07
	7	60297	39.21	0.65	10.50	-0.62	-0.31	0	60	60	0.91	3.19
	8	60248	40.73	0.68	9.88	-0.76	0.12	0	60	60	0.90	3.18
	10	63547	38.04	0.68	10.60	-0.67	-0.29	0	56	56	0.92	3.07
Mathematics	3	59340	41.66	0.73	10.13	-0.82	0.05	1	57	57	0.91	2.98
	4	59069	41.17	0.72	10.17	-0.78	-0.10	0	57	57	0.91	3.00
	5	59998	42.14	0.68	11.12	-0.47	-0.50	3	62	62	0.92	3.23
	6	60776	42.94	0.69	11.00	-0.61	-0.27	1	62	62	0.92	3.13
	7	60440	41.27	0.67	11.76	-0.50	-0.52	0	62	62	0.92	3.28
	8	60356	37.93	0.61	12.59	-0.10	-0.95	2	62	62	0.92	3.48
	10	63647	32.88	0.59	12.23	-0.04	-1.03	0	56	56	0.93	3.29
Language Arts	4	58848	18.75	0.62	5.47	-0.27	-0.61	1	30	30	0.82	2.32
	8	60186	21.42	0.74	5.73	-0.87	0.05	0	29	29	0.87	2.03
	10	63286	23.74	0.61	7.17	-0.24	-0.73	0	39	39	0.87	2.63
Social Studies	4	59006	27.59	0.77	5.89	-1.02	0.72	3	36	36	0.86	2.24
	8	60230	28.42	0.71	7.22	-0.66	-0.16	0	40	40	0.88	2.48
	10	63252	32.39	0.65	9.34	-0.38	-0.61	0	50	50	0.90	2.96
Science	4	59017	28.74	0.72	7.03	-0.68	-0.26	1	40	40	0.87	2.50
	8	60207	30.40	0.76	6.87	-0.92	0.31	0	40	40	0.88	2.37
	10	63319	31.31	0.63	10.11	-0.28	-0.86	0	50	50	0.91	3.02

Table 8-26
Raw Score Descriptive Statistics by Gender

Content	Grade	Male					Female				
		N Count	Mean	Test Difficulty	SD	Alpha	N Count	Mean	Test Difficulty	SD	Alpha
Reading	3	30224	39.49	0.66	12.48	0.94	28773	41.75	0.70	11.67	0.93
	4	30014	37.20	0.63	11.80	0.93	28843	38.61	0.65	11.17	0.92
	5	30689	38.67	0.64	11.22	0.92	29118	40.74	0.68	10.57	0.92
	6	31022	40.60	0.68	10.88	0.92	29660	43.03	0.72	10.03	0.91
	7	30859	38.33	0.64	10.81	0.91	29437	40.14	0.67	10.10	0.90
	8	30732	39.18	0.65	10.16	0.90	29513	42.34	0.71	9.30	0.89
	10	32374	37.21	0.66	10.85	0.92	31170	38.90	0.69	10.27	0.91
Mathematics	3	30413	41.80	0.73	10.17	0.91	28920	41.51	0.73	10.09	0.91
	4	30116	41.51	0.73	10.16	0.91	28943	40.82	0.72	10.17	0.91
	5	30800	42.07	0.68	11.21	0.92	29194	42.20	0.68	11.02	0.91
	6	31068	42.87	0.69	11.20	0.92	29706	43.02	0.69	10.78	0.92
	7	30937	40.82	0.66	11.97	0.92	29502	41.75	0.67	11.51	0.92
	8	30796	38.00	0.61	12.73	0.93	29557	37.86	0.61	12.44	0.92
	10	32430	33.11	0.59	12.57	0.93	31214	32.64	0.58	11.87	0.92
Language Arts	4	30005	18.15	0.61	5.45	0.81	28833	19.36	0.65	5.41	0.82
	8	30696	20.62	0.71	6.01	0.88	29487	22.24	0.77	5.30	0.86
	10	32240	22.51	0.58	7.22	0.86	31043	25.02	0.64	6.88	0.86
Social Studies	4	30087	27.45	0.76	6.03	0.86	28909	27.74	0.77	5.73	0.85
	8	30721	28.50	0.71	7.54	0.89	29506	28.33	0.71	6.87	0.87
	10	32228	32.46	0.65	9.82	0.91	31021	32.32	0.65	8.82	0.89
Science	4	30101	28.85	0.72	7.10	0.88	28906	28.62	0.72	6.95	0.87
	8	30717	30.63	0.77	7.16	0.89	29487	30.16	0.75	6.55	0.87
	10	32253	31.92	0.64	10.46	0.92	31063	30.68	0.61	9.69	0.90

Table 8-27
Raw Score Descriptive Statistics for Reading by Race/Ethnicity

Content	Race/Ethnicity	Grade	N Count	Mean	Test Difficulty	SD	Alpha
Reading	W	3	42963	42.72	0.71	11.18	0.93
		4	42944	40.04	0.68	10.70	0.92
		5	43953	41.59	0.69	10.11	0.91
		6	45043	43.82	0.73	9.48	0.90
		7	45089	41.10	0.69	9.66	0.90
		8	45395	42.31	0.71	9.06	0.88
		10	49268	39.81	0.71	9.67	0.90
	AA	3	6478	32.40	0.54	12.83	0.93
		4	6411	29.57	0.50	11.46	0.91
		5	6306	31.83	0.53	11.44	0.91
		6	6537	33.20	0.55	11.40	0.91
		7	6380	31.09	0.52	10.60	0.90
		8	6353	33.40	0.56	10.68	0.90
		10	5994	29.02	0.52	10.92	0.91
	H	3	6166	35.11	0.59	12.35	0.93
		4	6151	32.41	0.55	11.22	0.91
		5	6118	34.77	0.58	10.92	0.91
		6	5738	36.83	0.61	10.54	0.90
		7	5538	34.24	0.57	10.53	0.90
		8	5294	36.61	0.61	10.12	0.89
		10	4827	32.59	0.58	11.06	0.91
	A	3	2412	40.00	0.67	12.05	0.93
		4	2344	37.24	0.63	11.63	0.92
		5	2448	39.22	0.65	11.34	0.92
		6	2391	40.47	0.67	10.75	0.92
		7	2309	38.21	0.64	10.73	0.91
		8	2205	40.52	0.68	10.09	0.90
		10	2448	36.76	0.66	11.29	0.92
	AI	3	980	37.45	0.62	11.99	0.93
		4	1005	34.31	0.58	10.91	0.91
5		981	36.35	0.61	10.74	0.91	
6		974	37.96	0.63	10.51	0.91	
7		979	35.63	0.59	10.71	0.90	
8		998	37.49	0.62	9.94	0.89	
10		1003	34.41	0.61	10.60	0.91	

Table 8-28
Raw Score Descriptive Statistics for Mathematics by Race/Ethnicity

Content	Race/Ethnicity	Grade	N Count	Mean	Test Difficulty	SD	Alpha
Mathematics	W	3	42984	43.60	0.76	9.06	0.90
		4	42946	43.18	0.76	9.05	0.90
		5	43977	44.10	0.71	10.29	0.90
		6	45050	44.98	0.73	10.02	0.91
		7	45115	43.39	0.70	10.87	0.91
		8	45382	40.02	0.65	12.03	0.92
		10	49286	35.10	0.63	11.59	0.92
	AA	3	6489	33.05	0.58	11.24	0.92
		4	6412	32.20	0.56	11.04	0.91
		5	6318	33.18	0.54	11.21	0.91
		6	6547	33.35	0.54	11.29	0.91
		7	6388	31.23	0.50	11.55	0.91
		8	6362	28.11	0.45	11.02	0.89
		10	6006	21.24	0.38	9.32	0.87
	H	3	6421	37.66	0.66	10.07	0.90
		4	6303	36.99	0.65	10.25	0.90
		5	6233	37.44	0.60	10.73	0.90
		6	5793	38.29	0.62	10.59	0.90
		7	5620	36.00	0.58	11.32	0.91
		8	5366	32.13	0.52	11.39	0.90
		10	4882	25.77	0.46	10.68	0.90
	A	3	2462	42.22	0.74	9.84	0.91
		4	2389	41.64	0.73	10.01	0.91
		5	2482	43.32	0.70	11.05	0.92
		6	2412	44.01	0.71	11.00	0.92
		7	2334	42.59	0.69	11.38	0.92
		8	2247	39.58	0.64	12.97	0.93
		10	2465	33.05	0.59	12.46	0.93
	AI	3	979	38.47	0.67	10.14	0.90
		4	1007	37.76	0.66	9.85	0.90
5		983	38.63	0.62	10.78	0.90	
6		974	38.36	0.62	10.65	0.90	
7		981	36.52	0.59	11.38	0.91	
8		996	33.02	0.53	11.62	0.90	
10		1001	27.69	0.49	10.72	0.90	

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

Content	Race/Ethnicity	Grade	N Count	Mean	Test Difficulty	SD	Alpha
Language Arts	W	4	42922	19.68	0.66	5.17	0.80
		8	45373	22.33	0.77	5.27	0.86
		10	49169	24.88	0.64	6.80	0.86
	AA	4	6392	15.01	0.50	5.33	0.79
		8	6318	17.20	0.59	6.27	0.87
		10	5880	17.81	0.46	6.42	0.81
	H	4	6164	16.48	0.55	5.21	0.78
		8	5296	19.04	0.66	5.80	0.86
		10	4788	20.13	0.52	6.63	0.83
	A	4	2357	18.54	0.62	5.63	0.83
		8	2206	21.21	0.73	5.59	0.87
		10	2447	23.57	0.60	7.29	0.87
	AI	4	1001	16.91	0.56	5.04	0.77
		8	990	19.42	0.67	5.93	0.87
		10	995	20.52	0.53	6.87	0.84

Table 8-30
Raw Score Descriptive Statistics for Social Studies by Race/Ethnicity

Content	Race/Ethnicity	Grade	N Count	Mean	Test Difficulty	SD	Alpha
Social Studies	W	4	42953	28.70	0.80	5.20	0.83
		8	45374	29.71	0.74	6.58	0.87
		10	49171	33.90	0.68	8.75	0.89
	AA	4	6400	22.75	0.63	6.86	0.86
		8	6325	22.36	0.56	7.52	0.87
		10	5853	24.28	0.49	8.84	0.87
	H	4	6266	25.33	0.70	6.00	0.84
		8	5313	25.20	0.63	7.23	0.86
		10	4789	27.85	0.56	9.11	0.88
	A	4	2369	27.22	0.76	5.78	0.84
		8	2219	28.08	0.70	7.23	0.88
		10	2445	31.58	0.63	9.42	0.90
	AI	4	1006	25.88	0.72	6.13	0.85
		8	996	25.97	0.65	7.28	0.87
		10	987	29.41	0.59	8.97	0.88

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

Content	Race/Ethnicity	Grade	N Count	Mean	Test Difficulty	SD	Alpha
Science	W	4	42927	30.19	0.75	6.31	0.85
		8	45342	31.79	0.79	6.04	0.86
		10	49189	33.18	0.66	9.33	0.90
	AA	4	6413	22.67	0.57	7.39	0.86
		8	6322	24.04	0.60	7.48	0.87
		10	5877	21.35	0.43	8.84	0.87
	H	4	6278	25.62	0.64	6.89	0.85
		8	5316	26.93	0.67	7.07	0.86
		10	4797	25.54	0.51	9.59	0.89
	A	4	2382	28.22	0.71	6.97	0.87
		8	2231	29.44	0.74	7.00	0.88
		10	2450	30.35	0.61	10.28	0.91
	AI	4	1005	26.56	0.66	7.00	0.86
		8	993	28.36	0.71	7.05	0.87
		10	999	27.92	0.56	9.76	0.90

Table 8-32
Raw Score Descriptive Statistics by Socioeconomic Status

Content	Grade	Economically Disadvantaged					Not Economically Disadvantaged				
		N Count	Mean	Test Difficulty	SD	Alpha	N Count	Mean	Test Difficulty	SD	Alpha
Reading	3	26004	35.96	0.60	12.62	0.94	33000	44.25	0.74	10.38	0.92
	4	25905	33.23	0.56	11.55	0.92	32961	41.55	0.70	10.08	0.91
	5	25854	35.06	0.58	11.18	0.91	33957	43.21	0.72	9.36	0.90
	6	25455	37.20	0.62	10.97	0.91	35228	45.10	0.75	8.86	0.89
	7	24568	34.69	0.58	10.65	0.90	35729	42.32	0.71	9.19	0.89
	8	23962	36.70	0.61	10.26	0.89	36286	43.39	0.72	8.64	0.87
	10	21801	32.99	0.59	11.02	0.91	41745	40.68	0.73	9.35	0.90
Mathematics	3	26288	37.68	0.66	10.59	0.91	33052	44.83	0.79	8.51	0.89
	4	26089	37.10	0.65	10.66	0.91	32979	44.39	0.78	8.49	0.89
	5	26000	37.56	0.61	11.07	0.91	33998	45.64	0.74	9.80	0.90
	6	25523	38.17	0.62	11.15	0.91	35253	46.40	0.75	9.49	0.90
	7	24684	36.13	0.58	11.63	0.91	35756	44.82	0.72	10.46	0.91
	8	24053	32.41	0.52	11.72	0.91	36303	41.59	0.67	11.78	0.92
	10	21872	26.65	0.48	11.13	0.91	41774	36.13	0.65	11.50	0.92
Language Arts	4	25910	16.66	0.56	5.33	0.80	32937	20.38	0.68	4.99	0.80
	8	23907	19.10	0.66	6.04	0.87	36279	22.94	0.79	4.95	0.85
	10	21609	20.25	0.52	6.79	0.84	41676	25.55	0.66	6.67	0.85
Social Studies	4	26023	25.37	0.70	6.33	0.86	32982	29.34	0.82	4.84	0.81
	8	23940	25.24	0.63	7.41	0.87	36290	30.51	0.76	6.26	0.86
	10	21585	27.94	0.56	9.23	0.89	41666	34.70	0.69	8.53	0.89
Science	4	26050	25.96	0.65	7.27	0.87	32966	30.94	0.77	5.97	0.84
	8	23950	27.43	0.69	7.39	0.88	36257	32.37	0.81	5.71	0.85
	10	21628	26.32	0.53	9.90	0.90	41690	33.90	0.68	9.20	0.90

Table 8-33
Raw Score Descriptive Statistics by Disability

Content	Grade	Disabled					Not Disabled				
		N Count	Mean	Test Difficulty	SD	Alpha	N Count	Mean	Test Difficulty	SD	Alpha
Reading	RD	3	7353	29.68	0.49	13.41	3	51651	42.15	0.70	11.11
	RD	4	7676	27.29	0.46	12.12	4	51191	39.48	0.67	10.54
	RD	5	8034	28.39	0.47	11.83	5	51777	41.43	0.69	9.71
	RD	6	7812	29.64	0.49	11.35	6	52871	43.58	0.73	9.14
	RD	7	7699	27.56	0.46	10.66	7	52598	40.92	0.68	9.33
	RD	8	7690	29.18	0.49	10.40	8	52558	42.42	0.71	8.58
	RD	10	7518	25.80	0.46	10.35	10	56029	39.68	0.71	9.51
Mathematics	MA	3	7409	35.29	0.62	11.25	3	51931	42.57	0.75	9.63
	MA	4	7737	33.72	0.59	11.53	4	51332	42.30	0.74	9.46
	MA	5	8094	32.93	0.53	11.80	5	51904	43.57	0.70	10.29
	MA	6	7843	31.71	0.51	11.71	6	52933	44.61	0.72	9.85
	MA	7	7741	28.70	0.46	11.56	7	52699	43.12	0.70	10.60
	MA	8	7725	25.50	0.41	10.49	8	52631	39.76	0.64	11.81
	MA	10	7523	20.33	0.36	9.09	10	56124	34.56	0.62	11.61
Language Arts	LA	4	7684	15.09	0.50	5.31	4	51164	19.30	0.64	5.27
	LA	8	7662	15.36	0.53	6.03	8	52524	22.30	0.77	5.11
	LA	10	7406	16.18	0.41	5.82	10	55880	24.75	0.63	6.71
Social Studies	SS	4	7715	24.24	0.67	6.72	4	51291	28.10	0.78	5.58
	SS	8	7673	21.37	0.53	7.67	8	52557	29.44	0.74	6.54
	SS	10	7386	23.32	0.47	8.77	10	55866	33.59	0.67	8.74
Science	SC	4	7724	24.69	0.62	7.48	4	51293	29.35	0.73	6.75
	SC	8	7683	23.80	0.59	7.89	8	52524	31.37	0.78	6.14
	SC	10	7424	21.90	0.44	9.13	10	55895	32.56	0.65	9.56

Table 8-34
Raw Score Descriptive Statistics by English Language Proficiency

Content	Grade	Limited English Proficient					Fully English Proficient				
		N Count	Mean	Test Difficulty	SD	Alpha	N Count	Mean	Test Difficulty	SD	Alpha
Reading	3	54477	41.22	0.69	11.97	0.94	4527	33.06	0.55	11.65	0.92
	4	54662	38.55	0.65	11.34	0.92	4205	29.23	0.50	10.26	0.89
	5	56034	40.29	0.67	10.76	0.92	3777	30.60	0.51	9.84	0.88
	6	57436	42.35	0.71	10.32	0.91	3247	31.83	0.53	9.39	0.87
	7	57117	39.78	0.66	10.30	0.90	3180	29.10	0.48	8.90	0.85
	8	57297	41.17	0.69	9.70	0.89	2951	32.20	0.54	9.28	0.86
	10	61323	38.51	0.69	10.36	0.91	2224	25.13	0.45	8.92	0.85
Mathematics	3	54486	42.04	0.74	10.07	0.91	4854	37.41	0.66	9.90	0.90
	4	54653	41.61	0.73	10.05	0.91	4416	35.74	0.63	10.08	0.90
	5	56053	42.62	0.69	11.03	0.91	3945	35.34	0.57	10.13	0.88
	6	57445	43.40	0.70	10.88	0.92	3331	35.00	0.56	9.81	0.88
	7	57150	41.78	0.67	11.64	0.92	3290	32.45	0.52	10.14	0.88
	8	57281	38.42	0.62	12.53	0.92	3075	28.79	0.46	9.96	0.86
	10	61340	33.33	0.60	12.13	0.93	2307	20.67	0.37	8.11	0.83
Language Arts	4	54599	19.00	0.63	5.43	0.82	4249	15.46	0.52	4.88	0.74
	8	57215	21.66	0.75	5.65	0.87	2971	16.69	0.58	5.23	0.80
	10	61067	24.01	0.62	7.09	0.86	2219	16.40	0.42	5.08	0.71
Social Studies	4	54638	27.86	0.77	5.81	0.85	4368	24.24	0.67	5.85	0.82
	8	57235	28.73	0.72	7.11	0.88	2995	22.31	0.56	6.50	0.82
	10	61029	32.76	0.66	9.21	0.90	2223	22.42	0.45	7.11	0.79
Science	4	54627	29.10	0.73	6.94	0.87	4390	24.27	0.61	6.52	0.82
	8	57179	30.75	0.77	6.72	0.88	3028	23.86	0.60	6.46	0.81
	10	61084	31.72	0.63	9.96	0.91	2235	20.09	0.40	7.08	0.79

Table 8-35
2011 and 2009 Scale Score Mean and Standard Deviation

Content	Grade	2011 Mean	2009 Mean	Diff = 2011 – 2009 Mean	2011 Standard Deviation	2009 Standard Deviation	Diff = 2011 - 2009 Standard Deviation
Reading	3	457.00	456.59	0.41	39.29	39.43	-0.14
	4	476.93	476.87	0.06	48.54	49.41	-0.87
	5	480.98	480.18	0.80	48.17	48.54	-0.37
	6	503.97	503.53	0.44	49.89	49.16	0.73
	7	517.78	518.26	-0.48	48.18	48.09	0.09
	8	525.16	525.12	0.04	49.46	49.55	-0.09
	10	548.91	545.35	3.56	64.92	66.22	-1.30
Mathematics	3	436.91	437.15	-0.24	46.39	46.83	-0.44
	4	473.29	472.64	0.65	45.91	44.65	1.26
	5	499.52	497.25	2.27	50.64	49.90	0.74
	6	517.27	516.00	1.27	45.40	45.14	0.26
	7	539.79	539.29	0.50	45.35	43.98	1.37
	8	548.26	546.30	1.96	49.88	49.64	0.24
	10	562.60	560.00	2.60	50.59	48.97	1.62
Language Arts	4	294.35	294.86	-0.51	30.12	30.48	-0.36
	8	397.89	398.19	-0.30	42.03	42.83	-0.80
	10	449.73	446.13	3.60	41.29	40.55	0.74
Social Studies	4	297.94	298.23	-0.29	26.91	27.64	-0.73
	8	397.08	396.88	0.20	42.23	42.62	-0.39
	10	447.77	446.84	0.93	46.48	45.60	0.88
Science	4	299.67	299.54	0.13	31.95	32.35	-0.40
	8	405.04	404.65	0.39	43.27	43.58	-0.31
	10	451.03	449.35	1.68	49.80	49.15	0.65

Table 8-36
Scale Score Descriptive Statistics

Content	Grade	N Count	Mean	SD	Skewness	Kurtosis	Min	Max	LOSS	HOSS
Reading	3	59005	457.00	39.29	-1.15	4.92	270	640	270	640
	4	58867	476.93	48.54	-1.00	2.77	280	650	280	650
	5	59811	480.98	48.17	-0.57	1.69	290	690	290	690
	6	60683	503.97	49.89	-0.68	1.89	300	730	300	730
	7	60297	517.78	48.18	-0.59	1.37	310	780	310	780
	8	60248	525.16	49.46	-0.47	1.84	330	790	330	790
	10	63547	548.91	64.92	-0.39	0.94	350	820	350	820
Mathematics	3	59340	436.91	46.39	-0.02	2.14	220	630	220	630
	4	59069	473.29	45.91	-0.17	1.75	240	650	240	650
	5	59998	499.52	50.64	-0.37	1.74	270	680	270	680
	6	60776	517.27	45.40	-0.40	1.18	310	700	310	700
	7	60441	539.79	45.35	-0.44	1.56	330	710	330	710
	8	60356	548.26	49.88	-0.53	1.84	350	730	350	730
	10	63647	562.60	50.59	-0.53	1.50	410	750	410	750
Language Arts	4	58848	294.35	30.12	-1.16	6.70	140	420	140	420
	8	60186	397.89	42.03	0.25	2.23	250	520	250	520
	10	63286	449.73	41.29	-0.01	1.48	290	630	290	630
Social Studies	4	59006	297.94	26.91	0.38	3.88	170	400	170	400
	8	60230	397.08	42.23	-0.23	1.89	230	530	230	530
	10	63252	447.77	46.48	-0.63	2.53	240	620	240	620
Science	4	59017	299.67	31.95	0.15	2.78	170	440	170	440
	8	60207	405.04	43.27	0.35	3.49	230	560	230	560
	10	63319	451.03	49.80	-1.05	3.57	240	610	240	610

Table 8-37
Scale Score Descriptive Statistics by Gender

Content	Grade	Male					Female				
		N Count	Mean	SD	Min	Max	N Count	Mean	SD	Min	Max
Reading	3	30224	453.45	40.74	270	640	28774	460.73	37.35	270	640
	4	30014	473.96	50.61	280	650	28843	480.02	46.09	280	650
	5	30689	476.82	49.11	290	690	29118	485.37	46.76	290	690
	6	31022	498.35	51.15	300	730	29660	509.84	47.84	300	730
	7	30859	513.32	49.55	310	780	29437	522.46	46.24	310	780
	8	30732	517.78	50.08	330	790	29513	532.84	47.61	330	790
	10	32374	544.26	66.02	350	820	31170	553.74	63.40	350	820
Mathematics	3	30413	438.18	46.93	220	630	28920	435.58	45.78	220	630
	4	30116	475.06	46.78	240	650	28943	471.45	44.92	240	650
	5	30800	499.30	51.30	270	680	29194	499.75	49.94	270	680
	6	31068	517.26	46.61	310	700	29706	517.28	44.10	310	700
	7	30937	538.33	46.38	330	710	29503	541.32	44.19	330	710
	8	30796	548.59	51.17	350	730	29557	547.92	48.50	350	730
	10	32430	563.08	52.76	410	750	31214	562.10	48.21	410	750
Language Arts	4	30005	291.28	30.66	140	420	28833	297.53	29.21	140	420
	8	30696	392.03	42.41	250	520	29487	404.00	40.74	250	520
	10	32240	442.63	41.30	290	630	31043	457.10	39.96	290	630
Social Studies	4	30087	297.45	27.68	170	400	28909	298.46	26.08	170	400
	8	30721	397.53	45.10	230	530	29506	396.60	39.03	230	530
	10	32228	447.52	50.11	240	620	31021	448.04	42.37	240	620
Science	4	30101	300.02	32.71	170	440	28906	299.30	31.13	170	440
	8	30717	407.05	47.30	230	560	29487	402.93	38.53	230	560
	10	32253	453.49	53.14	240	610	31063	448.49	45.93	240	610

Table 8-38
Scale Score Descriptive Statistics for Reading by Race/Ethnicity

Content	Race/Ethnicity	Grade	N Count	Mean	SD	Min	Max
Reading	W	3	42963	463.45	35.86	270	640
		4	42944	485.50	44.57	280	650
		5	43953	489.14	44.93	290	690
		6	45043	513.12	45.64	300	730
		7	45089	526.23	44.57	310	780
		8	45395	532.82	46.18	330	790
		10	49268	559.45	59.62	350	820
	AA	3	6478	431.73	45.54	270	640
		4	6411	443.02	52.60	280	583
		5	6306	447.36	50.54	290	610
		6	6537	464.82	54.48	300	730
		7	6380	481.34	50.00	310	652
		8	6353	490.01	52.13	330	706
		10	5994	495.14	66.59	350	737
	H	3	6167	440.43	40.10	270	601
		4	6151	455.31	48.01	280	583
		5	6118	459.87	45.58	290	690
		6	5738	481.74	46.67	300	679
		7	5538	495.53	46.59	310	780
		8	5294	504.69	47.98	330	731
		10	4827	516.14	65.82	350	737
	A	3	2412	456.18	37.43	270	640
		4	2344	475.47	49.58	280	650
		5	2448	479.66	50.56	290	683
		6	2391	499.08	50.61	300	730
		7	2309	513.83	48.70	310	780
		8	2205	524.98	51.76	330	790
		10	2448	541.78	71.88	350	820
AI	3	980	447.49	37.95	270	557	
	4	1005	462.89	45.79	280	599	
	5	981	466.76	45.12	290	615	
	6	974	486.39	47.53	300	614	
	7	979	501.75	47.68	310	639	
	8	998	509.25	47.56	330	633	
	10	1003	527.43	62.56	350	691	

Table 8-39
Scale Score Descriptive Statistics for Mathematics by Race/Ethnicity

Content	Race/Ethnicity	Grade	N Count	Mean	SD	Min	Max
Mathematics	W	3	42984	445.42	43.34	220	630
		4	42946	481.81	42.56	240	650
		5	43977	508.12	47.02	270	680
		6	45050	525.26	41.87	310	700
		7	45115	547.61	42.09	330	710
		8	45382	556.37	46.08	350	730
		10	49286	571.78	45.43	410	750
	AA	3	6489	399.78	47.56	220	630
		4	6412	435.77	46.63	240	650
		5	6318	459.83	53.07	270	680
		6	6547	479.53	46.79	310	645
		7	6388	502.25	46.81	330	656
		8	6362	509.08	51.93	350	677
		10	6006	511.84	52.59	410	750
	H	3	6421	418.59	40.85	220	630
		4	6303	454.92	42.14	240	650
		5	6233	478.94	46.66	270	680
		6	5793	498.60	41.68	310	700
		7	5621	520.42	41.08	330	710
		8	5366	526.59	47.27	350	730
		10	4882	535.28	48.86	410	750
	A	3	2462	439.51	46.88	220	630
		4	2389	475.98	47.92	240	650
		5	2482	506.09	52.90	270	680
		6	2412	522.82	47.83	310	700
		7	2334	545.57	45.71	330	710
		8	2247	555.42	54.19	350	730
		10	2465	564.54	53.87	410	750
AI	3	979	422.98	42.14	220	630	
	4	1007	457.84	39.10	275	604	
	5	983	483.90	46.72	270	626	
	6	974	498.57	42.67	310	625	
	7	981	521.96	42.34	330	671	
	8	996	529.63	48.71	350	730	
	10	1001	543.45	45.59	410	664	

Table 8-40
Scale Score Descriptive Statistics for Language Arts by Race/Ethnicity

Content	Race/Ethnicity	Grade	N Count	Mean	SD	Min	Max
Language Arts	W	4	42922	299.28	27.40	140	420
		8	45373	403.82	40.76	250	520
		10	49169	455.98	39.29	290	630
	AA	4	6392	274.39	34.76	140	420
		8	6318	370.90	40.98	250	520
		10	5880	416.62	38.67	290	630
	H	4	6164	282.39	30.75	140	420
		8	5296	381.87	36.84	250	520
		10	4788	429.99	37.02	290	598
	A	4	2357	293.69	31.28	140	420
		8	2206	397.52	41.43	250	520
		10	2447	449.70	42.73	290	630
	AI	4	1001	285.59	27.40	140	367
		8	990	385.16	40.69	250	520
		10	995	431.67	39.94	290	630

Table 8-41
Scale Score Descriptive Statistics for Social Studies by Race/Ethnicity

Content	Race/Ethnicity	Grade	N Count	Mean	SD	Min	Max
Social Studies	W	4	42953	302.54	25.64	170	400
		8	45374	404.15	39.53	230	530
		10	49171	455.07	42.69	240	620
	AA	4	6400	278.50	27.46	170	400
		8	6325	363.95	42.86	230	530
		10	5853	407.72	50.23	240	620
	H	4	6266	287.97	22.93	170	400
		8	5313	378.96	39.07	230	530
		10	4789	426.04	46.99	240	620
	A	4	2369	296.60	27.44	170	400
		8	2219	396.45	42.73	230	530
		10	2445	445.44	46.46	240	620
	AI	4	1006	290.56	24.23	170	400
		8	996	383.38	40.18	230	530
		10	987	433.23	44.10	240	543

Table 8-42
Scale Score Descriptive Statistics for Science by Race/Ethnicity

Content	Race/Ethnicity	Grade	N Count	Mean	SD	Min	Max
Science	W	4	42927	305.75	30.08	170	440
		8	45342	412.61	40.82	230	560
		10	49189	459.93	43.79	240	610
	AA	4	6413	274.33	31.41	170	440
		8	6322	370.70	41.27	230	560
		10	5877	401.56	57.30	240	566
	H	4	6278	286.28	27.67	170	440
		8	5316	385.35	39.01	230	560
		10	4797	425.04	52.39	240	610
	A	4	2382	297.72	32.06	170	440
		8	2231	400.34	43.79	230	560
		10	2450	448.09	50.11	240	610
	AI	4	1005	289.61	29.39	170	440
		8	993	393.66	42.12	230	560
		10	999	436.27	48.54	240	570

Table 8-43
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
Reading	3	26005	442.81	41.32	270	640	33000	468.18	33.64	270	640
	4	25905	458.09	50.22	280	650	32961	491.74	41.57	280	650
	5	25854	461.00	47.79	290	690	33957	496.20	42.57	290	690
	6	25455	482.87	50.15	300	695	35228	519.22	43.77	300	730
	7	24568	497.53	47.95	310	720	35729	531.71	43.13	310	780
	8	23962	505.32	49.39	330	790	36286	538.26	44.94	330	790
	10	21801	518.75	65.50	350	820	41745	564.65	58.75	350	820
Mathematics	3	26288	418.92	44.39	220	630	33052	451.22	42.83	220	630
	4	26089	455.39	44.86	240	650	32979	487.45	41.59	240	650
	5	26000	479.14	49.40	270	680	33998	515.11	45.82	270	680
	6	25523	497.92	44.39	310	700	35253	531.28	40.74	310	700
	7	24685	520.43	43.76	330	710	35756	553.15	41.44	330	710
	8	24053	527.18	49.29	350	730	36303	562.23	45.14	350	730
	10	21872	537.51	50.72	410	750	41774	575.73	45.28	410	750
Language Arts	4	25910	283.62	31.18	140	420	32937	302.78	26.35	140	420
	8	23907	382.21	39.61	250	520	36279	408.22	40.35	250	520
	10	21609	430.44	38.05	290	630	41676	459.72	39.31	290	630
Social Studies	4	26023	288.38	25.35	170	400	32982	305.49	25.69	170	400
	8	23940	379.24	40.85	230	530	36290	408.84	38.88	230	530
	10	21585	426.58	46.89	240	620	41666	458.75	42.28	240	620
Science	4	26050	287.51	30.58	170	440	32966	309.27	29.66	170	440
	8	23950	388.07	41.44	230	560	36257	416.25	40.73	230	560
	10	21628	427.86	53.20	240	610	41690	463.05	43.31	240	610

Table 8-44
Scale Score Descriptive Statistics by Disability

Content	Grade	Disabled					Not Disabled				
		N Count	Mean	SD	Min	Max	N Count	Mean	SD	Min	Max
Reading	3	7353	422.35	50.70	270	640	51652	461.93	34.68	270	640
	4	7676	430.92	61.47	280	650	51191	483.83	42.17	280	650
	5	8034	432.15	56.24	290	636	51777	488.56	41.98	290	690
	6	7812	448.67	57.18	300	695	52871	512.14	43.07	300	730
	7	7699	465.46	52.44	310	780	52598	525.44	42.41	310	780
	8	7690	469.71	53.75	330	686	52558	533.27	43.20	330	790
	10	7518	476.51	65.92	350	737	56029	558.62	58.30	350	820
Mathematics	3	7409	409.08	48.73	220	630	51931	440.89	44.66	220	630
	4	7737	440.79	50.73	240	650	51332	478.19	43.07	240	650
	5	8094	456.82	58.46	270	680	51904	506.18	45.86	270	680
	6	7843	471.74	50.24	310	645	52933	524.02	40.50	310	700
	7	7741	492.36	48.71	330	710	52700	546.76	40.39	330	710
	8	7725	495.77	54.54	350	730	52631	555.97	44.19	350	730
	10	7523	506.41	52.35	410	750	56124	570.13	45.33	410	750
Language Arts	4	7684	274.44	34.94	140	420	51164	297.34	28.13	140	420
	8	7662	359.50	40.10	250	520	52524	403.49	39.29	250	520
	10	7406	408.24	35.32	290	619	55880	455.23	38.82	290	630
Social Studies	4	7715	284.24	27.67	170	400	51291	300.00	26.18	170	400
	8	7673	357.60	45.37	230	530	52557	402.84	38.51	230	530
	10	7386	402.98	49.65	240	620	55866	453.69	42.66	240	620
Science	4	7724	282.52	32.23	170	440	51293	302.25	31.10	170	440
	8	7683	368.52	46.37	230	560	52524	410.38	40.10	230	560
	10	7424	404.78	58.26	240	610	55895	457.17	45.14	240	610

Table 8-45
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
Reading	3	4528	434.72	37.70	270	601	54477	458.85	38.85	270	640
	4	4205	443.43	46.53	280	596	54662	479.51	47.73	280	650
	5	3777	443.24	41.53	290	584	56034	483.53	47.52	290	690
	6	3247	460.64	42.21	300	588	57436	506.42	49.16	300	730
	7	3180	473.69	40.35	310	611	57117	520.24	47.39	310	780
	8	2951	483.79	43.64	330	658	57297	527.29	48.80	330	790
	10	2224	472.55	56.61	350	641	61323	551.68	63.50	350	820
Mathematics	3	4854	417.19	40.19	220	630	54486	438.67	46.50	220	630
	4	4416	449.30	40.90	240	650	54653	475.23	45.75	240	650
	5	3945	469.86	44.96	270	637	56053	501.61	50.36	270	680
	6	3331	486.00	38.16	310	645	57445	519.09	45.13	310	700
	7	3291	507.70	37.82	330	671	57150	541.64	45.05	330	710
	8	3075	513.57	46.54	350	676	57281	550.12	49.37	350	730
	10	2307	512.11	47.39	410	655	61340	564.50	49.71	410	750
Language Arts	4	4249	276.99	31.02	140	420	54599	295.70	29.62	140	420
	8	2971	368.01	31.05	250	520	57215	399.44	41.95	250	520
	10	2219	410.55	30.30	290	586	61067	451.15	40.93	290	630
Social Studies	4	4368	283.46	20.95	170	400	54638	299.10	27.00	170	400
	8	2995	364.41	34.12	230	530	57235	398.78	41.92	230	530
	10	2223	400.91	40.85	240	545	61029	449.48	45.77	240	620
Science	4	4390	280.66	25.76	170	440	54627	301.19	31.91	170	440
	8	3028	369.52	33.49	230	560	57179	406.92	42.92	230	560
	10	2235	398.20	51.03	240	551	61084	452.96	48.68	240	610

Table 8-46
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
Reading	394	430	466	396	440	489	401	444	497	418	457	514	434	467	523	445	480	539	456	503	555
Mathematics	392	407	452	421	438	484	445	463	505	464	485	532	480	504	555	483	513	573	516	541	595
Language Arts				252	277	308										358	385	418	393	428	484
Social Studies				242	263	288										334	364	403	408	420	455
Science				249	279	320										349	375	419	411	429	466

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

Table 8-47
 Percentage of Students in Each Performance Level by Sub-Group (Reading)

Grade	Proficiency Level	Examinees		Gender		Race/Ethnicity					ELP		Disability		SES	
		N	%	Female	Male	White	African American	Hispanic	Asian	American Indian	Fully English Proficient	Limited English Proficient	Disabled	Not Disabled	Economically Disadvantaged	Not Economically Disadvantaged
3	M	2715	4.60	3.40	5.74	2.84	13.26	8.11	3.52	5.10	4.27	8.59	20.41	2.35	8.12	1.83
	B	8791	14.90	13.15	16.56	11.10	28.73	24.91	17.16	21.22	13.70	29.35	31.71	12.50	22.75	8.71
	P	21416	36.30	35.87	36.70	34.87	38.33	42.08	39.01	42.24	35.59	44.79	31.28	37.01	40.65	32.86
	A	26083	44.21	47.58	40.99	51.18	19.68	24.91	40.30	31.43	46.44	17.27	16.59	48.14	28.48	56.59
Total		59005	100	28774	30224	42963	6478	6167	2412	980	54477	4528	7353	51652	26005	33000
4	M	2786	4.73	3.62	5.80	2.89	13.41	8.36	4.44	6.47	4.26	10.84	21.68	2.19	8.57	1.72
	B	7668	13.03	11.76	14.24	9.37	27.69	22.06	14.16	17.61	11.85	28.30	29.66	10.53	20.17	7.41
	P	22891	38.89	39.79	38.02	36.96	41.90	46.64	41.89	47.46	38.12	48.85	33.42	39.71	44.65	34.36
	A	25522	43.36	44.84	41.94	50.78	17.00	22.94	39.51	28.46	45.77	12.01	15.24	47.57	26.61	56.51
Total		58867	100	28843	30014	42944	6411	6151	2344	1005	54662	4205	7676	51191	25905	32961
5	M	3012	5.04	3.79	6.22	3.13	14.78	8.42	5.02	6.83	4.55	12.31	24.56	2.01	9.21	1.86
	B	8344	13.95	12.75	15.09	10.22	28.18	24.04	16.71	19.98	12.61	33.81	31.94	11.16	22.14	7.72
	P	25587	42.78	42.16	43.37	41.99	42.71	47.68	42.52	48.62	42.49	47.07	32.45	44.38	47.08	39.51
	A	22868	38.23	41.30	35.32	44.66	14.34	19.86	35.74	24.57	40.35	6.80	11.05	42.45	21.57	50.92
Total		59811	100	29118	30689	43953	6306	6118	2448	981	56034	3777	8034	51777	25854	33957

Table 8-47 Cont'd
 Percentage of Students in Each Performance Level by Sub-Group (Reading)

Grade	Proficiency Level	Examinees		Gender		Race/Ethnicity					ELP		Disability		SES	
		N	%	Female	Male	White	African-American	Hispanic	Asian	American Indian	Fully English Proficient	Limited English Proficient	Disabled	Not Disabled	Economically Disadvantaged	Not Economically Disadvantaged
6	M	2922	4.82	3.35	6.21	2.77	15.71	8.19	4.56	6.88	4.36	12.78	24.71	1.88	8.89	1.87
	B	6262	10.32	8.65	11.92	7.30	23.54	17.03	12.84	15.71	9.37	27.10	28.15	7.68	16.72	5.69
	P	23800	39.22	38.30	40.10	36.55	43.89	50.96	44.83	48.25	38.42	53.31	36.42	39.63	47.19	33.46
	A	27699	45.65	49.69	41.77	53.38	16.86	23.82	37.77	29.16	47.84	6.81	10.73	50.80	27.20	58.97
Total		60683	100	29660	31022	45043	6537	5738	2391	974	57436	3247	7812	52871	25455	35228
7	M	2969	4.92	3.53	6.25	2.99	14.80	8.85	5.07	7.05	4.40	14.28	24.31	2.09	8.96	2.15
	B	5047	8.37	7.22	9.47	5.78	19.03	15.57	9.40	14.91	7.57	22.77	25.34	5.89	13.97	4.52
	P	22539	37.38	36.80	37.93	34.46	46.49	47.07	43.92	42.49	36.41	54.84	37.04	37.43	46.02	31.44
	A	29742	49.33	52.45	46.34	56.77	19.69	28.51	41.62	35.55	51.62	8.11	13.30	54.60	31.05	61.89
Total		60297	100	29437	30859	45089	6380	5538	2309	979	57117	3180	7699	52598	24568	35729
8	M	3216	5.34	3.51	7.09	3.30	15.84	9.77	5.31	8.02	4.81	15.55	27.27	2.13	9.83	2.37
	B	6076	10.09	8.18	11.91	7.58	21.33	16.72	10.98	15.23	9.34	24.50	28.05	7.46	16.00	6.18
	P	26581	44.12	42.22	45.94	42.78	47.03	50.66	44.85	50.30	43.63	53.64	36.62	45.22	49.99	40.24
	A	24375	40.46	46.08	35.05	46.35	15.80	22.86	38.87	26.45	42.22	6.30	8.06	45.20	24.18	51.21
Total		60248	100	29513	30732	45395	6353	5294	2205	998	57297	2951	7690	52558	23962	36286

Table 8-47 Cont'd
 Percentage of Students in Each Performance Level by Sub-Group (Reading)

Grade	Proficiency Level	Examinees		Gender		Race/Ethnicity					ELP		Disability		SES	
		N	%	Female	Male	White	African-American	Hispanic	Asian	American Indian	Fully English Proficient	Limited English Proficient	Disabled	Not Disabled	Economically Disadvantaged	Not Economically Disadvantaged
10	M	4948	7.79	6.34	9.17	4.60	24.86	16.66	10.42	12.86	6.84	33.81	35.14	4.12	15.43	3.80
	B	8218	12.93	11.72	14.10	10.20	26.31	21.55	16.34	17.75	12.16	34.13	29.66	10.69	21.05	8.69
	P	18984	29.87	30.29	29.48	29.24	30.73	34.00	30.23	34.90	29.99	26.71	23.84	30.68	33.27	28.10
	A	31397	49.41	51.65	47.25	55.96	18.10	27.80	43.01	34.50	51.01	5.35	11.36	54.51	30.26	59.41
Total		63547	100	31170	32374	49268	5994	4827	2448	1003	61323	2224	7518	56029	21801	41745

Table 8-48
 Percentage of Students in Each Performance Level by Sub-Group (Mathematics)

Grade	Proficiency Level	Examinees		Gender		Race/Ethnicity					ELP		Disability		SES	
		N	%	Female	Male	White	African-American	Hispanic	Asian	American Indian	Fully English Proficient	Limited English Proficient	Disabled	Not Disabled	Economically Disadvantaged	Not Economically Disadvantaged
3	M	8654	14.58	14.85	14.33	9.29	40.48	23.25	13.20	21.96	13.78	23.65	33.86	11.83	24.19	6.95
	B	5069	8.54	8.68	8.41	7.03	13.55	13.24	8.37	11.54	8.10	13.54	12.92	7.92	12.08	5.73
	P	23306	39.28	40.62	38.00	39.29	33.70	44.11	39.64	43.11	38.72	45.55	35.58	39.80	41.75	37.30
	A	22311	37.60	35.84	39.27	44.39	12.27	19.41	38.79	23.39	39.41	17.26	17.64	40.45	21.98	50.02
Total		59340	100	28920	30413	42984	6489	6421	2462	979	54486	4854	7409	51931	26288	33052
4	M	6693	11.33	11.72	10.96	6.74	34.79	18.78	9.38	15.69	10.53	21.20	32.18	8.19	19.57	4.82
	B	4846	8.20	8.75	7.68	6.35	14.89	12.93	8.96	13.21	7.70	14.38	14.02	7.33	12.20	5.04
	P	23016	38.97	40.02	37.94	38.26	36.26	44.98	39.81	46.28	38.35	46.58	35.13	39.54	42.54	36.14
	A	24514	41.50	39.51	43.42	48.65	14.05	23.31	41.86	24.83	43.41	17.84	18.66	44.94	25.69	54.01
Total		59069	100	28943	30116	42946	6412	6303	2389	1007	54653	4416	7737	51332	26089	32979
5	M	7324	12.21	11.73	12.66	7.93	34.19	20.21	9.95	16.89	11.36	24.31	38.03	8.18	20.92	5.55
	B	4972	8.29	8.52	8.07	6.48	15.34	13.41	8.14	11.90	7.75	15.89	13.78	7.43	12.30	5.22
	P	18935	31.56	32.24	30.92	30.48	31.94	37.99	31.06	38.05	30.94	40.38	28.87	31.98	36.43	27.84
	A	28767	47.95	47.52	48.35	55.11	18.53	28.38	50.85	33.16	49.95	19.42	19.32	52.41	30.36	61.40
Total		59998	100	29194	30800	43977	6318	6233	2482	983	56053	3945	8094	51904	26000	33998

Table 8-48 Cont'd
 Percentage of Students in Each Performance Level by Sub-Group (Mathematics)

Grade	Proficiency Level	Examinees		Gender		Race/Ethnicity					ELP		Disability		SES	
		N	%	Female	Male	White	African-American	Hispanic	Asian	American Indian	Fully English Proficient	Limited English Proficient	Disabled	Not Disabled	Economically Disadvantaged	Not Economically Disadvantaged
6	M	6712	11.04	10.30	11.75	6.82	33.66	18.02	9.04	17.97	10.28	24.29	41.48	6.53	19.63	4.83
	B	6049	9.95	10.36	9.56	7.79	18.47	16.02	10.24	15.81	9.30	21.13	17.56	8.83	15.25	6.12
	P	24682	40.61	41.75	39.52	40.49	36.28	46.00	39.47	46.10	40.35	45.12	30.19	42.16	43.48	38.53
	A	23333	38.39	37.58	39.17	44.90	11.59	19.96	41.25	20.12	40.07	9.46	10.77	42.48	21.64	50.52
Total		60776	100	29706	31068	45050	6547	5793	2412	974	57445	3331	7843	52933	25523	35253
7	M	4998	8.27	7.27	9.23	4.99	26.06	14.32	6.34	13.05	7.65	19.08	36.18	4.17	14.87	3.71
	B	6275	10.38	9.91	10.83	7.85	21.67	17.58	8.44	16.72	9.69	22.39	22.56	8.59	16.27	6.32
	P	26643	44.08	44.73	43.47	43.64	41.47	48.92	45.80	49.54	43.73	50.20	33.02	45.71	48.69	40.90
	A	22525	37.27	38.09	36.48	43.52	10.80	19.18	39.42	20.69	38.93	8.33	8.24	41.53	20.17	49.07
Total		60441	100	29503	30937	45115	6388	5621	2334	981	57150	3291	7741	52700	24685	35756
8	M	4972	8.24	7.97	8.50	5.09	24.80	14.82	6.54	14.06	7.62	19.77	34.91	4.32	14.88	3.83
	B	7199	11.93	11.98	11.88	9.46	23.58	18.75	10.24	16.87	11.31	23.41	25.85	9.88	18.31	7.70
	P	29312	48.57	50.04	47.14	48.90	43.16	52.09	47.17	51.81	48.48	50.18	33.42	50.79	51.21	46.81
	A	18873	31.27	30.01	32.48	36.54	8.46	14.35	36.05	17.27	32.59	6.63	5.81	35.01	15.60	41.65
Total		60356	100	29557	30796	45382	6362	5366	2247	996	57281	3075	7725	52631	24053	36303

Table 8-48 Cont'd
 Percentage of Students in Each Performance Level by Sub-Group (Mathematics)

Grade	Proficiency Level	Examinees		Gender		Race/Ethnicity					ELP		Disability		SES	
		N	%	Female	Male	White	African-American	Hispanic	Asian	American Indian	Fully English Proficient	Limited English Proficient	Disabled	Not Disabled	Economically Disadvantaged	Not Economically Disadvantaged
10	M	9233	14.51	13.63	15.34	9.11	46.20	28.64	13.83	22.48	13.32	45.99	52.25	9.45	27.39	7.76
	B	8561	13.45	13.57	13.34	11.40	22.31	21.51	13.59	21.68	13.01	25.14	22.33	12.26	20.07	9.99
	P	29690	46.65	49.08	44.31	49.47	28.32	41.09	46.37	45.55	47.36	27.61	22.28	49.91	42.22	48.96
	A	16163	25.40	23.72	27.01	30.02	3.16	8.77	26.21	10.29	26.30	1.26	3.14	28.38	10.32	33.29
Total		63647	100	31214	32430	49286	6006	4882	2465	1001	61340	2307	7523	56124	21872	41774

Table 8-49
 Percentage of Students in Each Performance Level by Sub-Group (Language Arts)

Grade	Proficiency Level	Examinees		Gender		Race/Ethnicity					ELP		Disability		SES	
		N	%	Female	Male	White	African-American	Hispanic	Asian	American Indian	Fully English Proficient	Limited English Proficient	Disabled	Not Disabled	Economically Disadvantaged	Not Economically Disadvantaged
4	M	2909	4.94	4.00	5.85	2.80	14.69	9.25	5.56	6.49	4.41	11.74	15.72	3.32	8.87	1.86
	B	10260	17.44	15.12	19.66	13.48	33.56	26.51	18.84	24.58	16.38	31.04	31.62	15.30	25.67	10.96
	P	26891	45.70	44.65	46.70	46.24	39.91	47.39	44.72	51.45	45.54	47.75	41.02	46.40	47.04	44.64
	A	18788	31.93	36.23	27.79	37.48	11.84	16.86	30.89	17.48	33.67	9.46	11.63	34.97	18.42	42.55
Total		58848	100	28833	30005	42922	6392	6164	2357	1001	54599	4249	7684	51164	25910	32937
8	M	7630	12.68	8.94	16.27	9.05	31.77	20.07	12.19	18.59	11.82	29.25	42.76	8.29	21.37	6.95
	B	13577	22.56	20.24	24.78	19.81	31.85	32.23	25.39	31.31	21.53	42.28	34.18	20.86	30.27	17.48
	P	22887	38.03	39.50	36.61	39.87	27.73	35.50	37.04	35.05	38.65	26.09	18.59	40.86	34.32	40.47
	A	16092	26.74	31.32	22.34	31.27	8.66	12.20	25.39	15.05	28.00	2.39	4.48	29.98	14.04	35.10
Total		60186	100	29487	30696	45373	6318	5296	2206	990	57215	2971	7662	52524	23907	36279
10	M	4418	6.98	4.06	9.80	4.56	21.89	12.91	6.21	11.96	6.43	22.26	29.64	3.98	13.39	3.66
	B	13662	21.59	17.92	25.12	17.61	40.51	34.96	24.89	33.47	20.49	51.69	44.44	18.56	33.71	15.30
	P	33227	52.50	54.81	50.29	55.58	34.22	45.66	50.18	47.04	53.48	25.60	24.03	56.28	45.80	55.98
	A	11979	18.93	23.22	14.80	22.24	3.38	6.47	18.72	7.54	19.60	0.45	1.89	21.19	7.10	25.06
Total		63286	100	31043	32240	49169	5880	4788	2447	995	61067	2219	7406	55880	21609	41676

Table 8-50
 Percentage of Students in Each Performance Level by Sub-Group (Social Studies)

Grade	Proficiency Level	Examinees		Gender		Race/Ethnicity					ELP		Disability		SES	
		N	%	Female	Male	White	African-American	Hispanic	Asian	American Indian	Fully English Proficient	Limited English Proficient	Disabled	Not Disabled	Economically Disadvantaged	Not Economically Disadvantaged
4	M	881	1.49	1.27	1.71	0.68	6.33	2.04	1.48	2.19	1.40	2.66	4.55	1.03	2.79	0.47
	B	2914	4.94	4.49	5.36	2.85	15.23	8.31	4.60	8.25	4.54	9.87	11.54	3.95	8.68	1.98
	P	15208	25.77	25.44	26.09	21.02	41.83	38.96	30.60	32.90	24.20	45.44	38.31	23.89	35.92	17.77
	A	40003	67.80	68.80	66.84	75.45	36.61	50.69	63.32	56.66	69.85	42.03	45.60	71.13	52.61	79.77
Total		59006	100	28909	30087	42953	6400	6266	2369	1006	54638	4368	7715	51291	26023	32982
8	M	3289	5.46	4.38	6.50	3.23	17.38	9.84	5.23	8.63	4.99	14.52	24.15	2.73	10.20	2.33
	B	7690	12.77	12.60	12.93	9.43	29.19	20.16	13.79	18.78	11.92	28.95	30.72	10.15	20.85	7.44
	P	22279	36.99	39.58	34.50	35.77	37.61	44.68	39.25	42.47	36.48	46.68	31.80	37.75	42.00	33.68
	A	26972	44.78	43.44	46.07	51.57	15.83	25.32	41.73	30.12	46.61	9.85	13.33	49.37	26.95	56.55
Total		60230	100	29506	30721	45374	6325	5313	2219	996	57235	2995	7673	52557	23940	36290
10	M	10269	16.24	14.34	18.05	11.29	45.17	29.23	17.96	23.40	14.93	52.14	52.03	11.50	29.55	9.34
	B	4340	6.86	7.04	6.69	5.82	11.72	10.17	8.67	9.22	6.58	14.71	11.51	6.25	10.20	5.13
	P	19275	30.47	33.24	27.81	30.17	28.04	34.87	31.29	36.88	30.58	27.58	24.26	31.29	33.33	28.99
	A	29368	46.43	45.38	47.45	52.72	15.07	25.73	42.09	30.50	47.92	5.58	12.20	50.96	26.93	56.53
Total		63252	100	31021	32228	49171	5853	4789	2445	987	61029	2223	7386	55866	21585	41666

Table 8-51
 Percentage of Students in Each Performance Level by Sub-Group (Science)

Grade	Proficiency Level	Examinees		Gender		Race/Ethnicity					ELP		Disability		SES	
		N	%	Female	Male	White	African-American	Hispanic	Asian	American Indian	Fully English Proficient	Limited English Proficient	Disabled	Not Disabled	Economically Disadvantaged	Not Economically Disadvantaged
4	M	2835	4.80	4.65	4.95	2.46	17.48	7.49	4.70	7.66	4.47	9.00	11.79	3.75	8.74	1.69
	B	10432	17.68	17.82	17.53	12.83	36.60	29.88	19.23	23.88	16.34	34.31	31.43	15.60	26.88	10.40
	P	31715	53.74	54.58	52.93	55.82	40.06	52.96	54.53	55.32	53.84	52.44	46.74	54.79	52.21	54.94
	A	14035	23.78	22.95	24.58	28.90	5.86	9.67	21.54	13.13	25.35	4.26	10.03	25.85	12.16	32.96
Total		59017	100	28906	30101	42927	6413	6278	2382	1005	54627	4390	7724	51293	26050	32966
8	M	4181	6.94	5.93	7.92	3.82	23.38	12.94	8.70	8.96	6.17	21.53	27.72	3.90	13.08	2.89
	B	7299	12.12	12.67	11.60	8.65	26.43	21.86	15.78	19.13	11.15	30.42	25.81	10.12	19.49	7.26
	P	28337	47.07	50.92	43.36	47.34	41.95	50.24	46.88	50.55	47.22	44.19	36.03	48.68	48.58	46.07
	A	20390	33.87	30.48	37.12	40.19	8.24	14.95	28.64	21.35	35.46	3.86	10.44	37.29	18.85	43.78
Total		60207	100	29487	30717	45342	6322	5316	2231	993	57179	3028	7683	52524	23950	36257
10	M	10107	15.96	15.83	16.09	9.96	51.64	31.52	17.67	22.72	14.59	53.42	49.43	11.52	30.08	8.64
	B	6567	10.37	11.28	9.50	8.82	15.55	16.68	14.49	15.62	9.98	20.94	16.35	9.58	14.92	8.01
	P	21613	34.13	37.06	31.31	35.32	24.11	34.13	33.51	36.24	34.56	22.51	24.03	35.48	33.81	34.30
	A	25032	39.53	35.83	43.10	45.90	8.69	17.68	34.33	25.43	40.87	3.13	10.18	43.43	21.19	49.05
Total		63319	100	31063	32253	49189	5877	4797	2450	999	61084	2235	7424	55895	21628	41690

Table 8-52
Cut Scores and Associated Impact Data for WKCE-CRT Reading

Grade	Score Range				Impact Data				
	Minimal	Basic	Proficient	Advanced	Minimal	Basic	Proficient	Advanced	Proficient +Advanced
3	270-393	394-429	430-465	466-640	4.60	14.90	36.30	44.20	80.50
4	280-395	396-439	440-488	489-650	4.73	13.03	38.89	43.36	82.24
5	290-400	401-443	444-496	497-690	5.04	13.95	42.78	38.23	81.01
6	300-417	418-456	457-513	514-730	4.82	10.32	39.22	45.65	84.87
7	310-433	434-466	467-522	523-780	4.92	8.37	37.38	49.33	86.71
8	330-444	445-479	480-538	539-790	5.34	10.09	44.12	40.46	84.58
10	350-455	456-502	503-554	555-820	7.79	12.93	29.87	49.41	79.28

Table 8-53
Cut Scores and Associated Impact Data for WKCE-CRT Mathematics

Grade	Score Range				Impact Data				
	Minimal	Basic	Proficient	Advanced	Minimal	Basic	Proficient	Advanced	Proficient +Advanced
3	220-391	392-406	407-451	452-630	14.58	8.54	39.28	37.60	76.87
4	240-420	421-437	438-483	484-650	11.33	8.20	38.96	41.50	80.47
5	270-444	445-462	463-504	505-680	12.21	8.29	31.56	47.95	79.51
6	310-463	464-484	485-531	532-700	11.04	9.95	40.61	38.39	79.00
7	330-479	480-503	504-554	555-710	8.27	10.38	44.08	37.27	81.35
8	350-482	483-512	513-572	573-730	8.24	11.93	48.57	31.27	79.83
10	410-515	516-540	541-594	595-750	14.51	13.45	46.65	25.39	72.04

Table 8-54
Cut Scores and Associated Impact Data for WKCE-CRT Language Arts

Grade	Score Range				Impact Data				
	Minimal	Basic	Proficient	Advanced	Minimal	Basic	Proficient	Advanced	Proficient +Advanced
4	140-251	252-276	277-307	308-420	4.94	17.43	45.70	31.93	77.62
8	250-357	358-384	385-417	418-520	12.68	22.56	38.03	26.74	64.76
10	290-392	393-427	428-483	484-630	6.98	21.59	52.50	18.93	71.43

Table 8-55
Cut Scores and Associated Impact Data for WKCE-CRT Social Studies

Grade	Score Range				Impact Data				
	Minimal	Basic	Proficient	Advanced	Minimal	Basic	Proficient	Advanced	Proficient +Advanced
4	170-241	242-262	263-287	288-400	1.49	4.94	25.77	67.79	93.57
8	230-333	334-363	364-402	403-530	5.46	12.77	36.99	44.78	81.77
10	240-407	408-419	420-454	455-620	16.24	6.86	30.47	46.43	76.90

Table 8-56
 Cut Scores and Associated Impact Data for WKCE-CRT Science

Grade	Score Range				Impact Data				
	Minimal	Basic	Proficient	Advanced	Minimal	Basic	Proficient	Advanced	Proficient +Advanced
4	170-248	249-278	279-319	320-440	4.80	17.68	53.74	23.78	77.52
8	230-348	349-374	375-418	419-560	6.94	12.12	47.07	33.87	80.93
10	240-410	411-428	429-465	466-610	15.96	10.37	34.13	39.53	73.67

Table 8-57
Summary Statistics for Reading Content Standards Raw and SPI Scores

Grade	N	Content Standard	Standard	No. of Items		Total Score Points	Mean	Mean <i>p</i> -value	SD	SPI	
				MC	CR					Mean	SD
3	59004	1	Determines Meaning	12	0	12	8.52	0.71	2.75	70.21	20.62
	59004	2	Understands Text	20	0	20	14.99	0.75	4.53	74.87	21.95
	59004	3	Analyzes Text	18	1	21	13.92	0.66	4.48	66.87	20.23
	59004	4	Evaluates/Extends Text	4	1	7	3.16	0.45	1.64	46.15	17.94
4	58867	1	Determines Meaning	11	0	11	7.28	0.66	2.48	66.04	20.14
	58867	2	Understands Text	19	0	19	12.99	0.68	4.16	68.35	20.43
	58867	3	Analyzes Text	18	1	21	12.97	0.62	4.31	61.80	19.34
	58867	4	Evaluates/Extends Text	5	1	8	4.64	0.58	1.87	58.73	19.29
5	59811	1	Determines Meaning	11	0	11	8.21	0.75	2.43	73.93	19.77
	59811	2	Understands Text	17	0	17	12.01	0.71	3.44	70.69	18.69
	59811	3	Analyzes Text	18	2	24	14.58	0.61	4.33	61.35	16.97
	59811	4	Evaluates/Extends Text	8	0	8	4.88	0.61	2.02	61.53	21.03
6	60683	1	Determines Meaning	10	0	10	6.73	0.67	2.16	67.01	18.26
	60683	2	Understands Text	14	0	14	10.56	0.75	2.80	75.55	18.15
	60683	3	Analyzes Text	19	1	22	15.40	0.70	3.90	70.09	16.74
	60683	4	Evaluates/Extends Text	11	1	14	9.09	0.65	2.95	65.08	18.52
7	60297	1	Determines Meaning	10	0	10	6.61	0.66	2.32	65.08	20.23
	60297	2	Understands Text	14	0	14	10.59	0.76	2.77	75.23	17.94
	60297	3	Analyzes Text	19	1	22	14.12	0.64	4.20	64.64	17.80
	60297	4	Evaluates/Extends Text	11	1	14	7.89	0.56	2.59	57.22	16.20
8	60248	1	Determines Meaning	10	0	10	7.47	0.75	2.03	74.24	17.56
	60248	2	Understands Text	14	0	14	10.24	0.73	2.78	73.26	17.66
	60248	3	Analyzes Text	19	1	22	14.01	0.64	3.75	64.18	15.40
	59004	4	Evaluates/Extends Text	11	1	14	9.01	0.64	2.67	64.47	16.86

Table 8-57 Cont'd
 Summary Statistics for Reading Content Standards Raw and SPI Scores

Grade	N	Content Standard	Standard	No. of Items		Total Score Points	Mean	Mean <i>p</i> -value	SD	SPI	
				MC	CR					Mean	SD
10	63547	1	Determines Meaning	7	0	7	5.91	0.84	1.37	83.22	16.84
	63547	2	Understands Text	7	0	7	4.62	0.66	1.71	66.04	19.49
	63547	3	Analyzes Text	22	1	25	16.60	0.66	5.15	66.43	19.60
	63547	4	Evaluates/Extends Text	14	1	17	10.92	0.64	3.55	64.76	19.48

Table 8-58
Summary Statistics for Mathematics Content Standards Raw and SPI Scores

Grade	N	Content Standard	Standard	No. of Items		Total Score Points	Mean	Mean <i>p</i> -value	SD	SPI	
				MC	CR					Mean	SD
3	59340	A	Mathematical Processes	3	3	9	5.18	0.58	2.28	57.88	21.26
	59340	B	Number Operations	11	1	12	9.50	0.79	2.49	78.93	19.28
	59340	C	Geometry	9	2	11	8.98	0.82	1.85	81.85	14.40
	59340	D	Measurement	8	0	8	6.01	0.75	1.73	75.10	17.74
	59340	E	Statistics/Probability	8	0	8	5.31	0.66	2.05	66.37	21.96
	59340	F	Algebraic Relationships	7	1	9	6.68	0.74	1.92	74.13	17.75
4	59069	A	Mathematical Processes	3	3	9	4.72	0.52	2.24	52.68	21.45
	59069	B	Number Operations	11	0	11	8.59	0.78	2.22	78.04	17.98
	59069	C	Geometry	9	1	10	8.02	0.80	1.81	80.19	14.98
	59069	D	Measurement	8	1	9	6.78	0.75	1.91	75.95	17.79
	59069	E	Statistics/Probability	7	1	8	5.82	0.73	1.85	71.16	19.81
	59069	F	Algebraic Relationships	8	1	10	7.26	0.73	2.24	73.10	19.36
5	59998	A	Mathematical Processes	3	3	9	5.87	0.65	2.00	64.96	18.78
	59998	B	Number Operations	11	0	11	8.27	0.75	2.31	74.64	19.04
	59998	C	Geometry	9	1	11	7.51	0.68	2.10	68.95	14.20
	59998	D	Measurement	9	1	10	7.33	0.73	2.08	73.27	17.83
	59998	E	Statistics/Probability	9	1	10	6.00	0.60	2.39	60.17	20.44
	59998	F	Algebraic Relationships	10	1	11	7.15	0.65	2.57	65.51	20.71

Table 8-58 Cont'd
 Summary Statistics for Mathematics Content Standards Raw and SPI Scores

Grade	N	Content Standard	Standard	No. of Items		Total Score Points	Mean	Mean <i>p</i> -value	SD	SPI	
				MC	CR					Mean	SD
6	60776	A	Mathematical Processes	3	3	9	4.95	0.55	2.04	55.79	20.01
	60776	B	Number Operations	12	0	12	9.43	0.79	2.32	78.59	17.32
	60776	C	Geometry	9	1	10	7.95	0.80	1.75	77.30	13.38
	60776	D	Measurement	9	1	10	6.41	0.64	2.37	64.76	20.40
	60776	E	Statistics/Probability	8	1	10	6.37	0.64	2.38	63.89	18.92
	60776	F	Algebraic Relationships	10	1	11	7.84	0.71	2.41	71.65	19.51
7	60440	A	Mathematical Processes	3	3	9	5.12	0.57	2.29	58.15	22.22
	60440	B	Number Operations	12	0	12	8.77	0.73	2.67	73.11	20.25
	60440	C	Geometry	10	2	12	8.03	0.67	2.57	66.69	18.25
	60440	D	Measurement	9	0	9	5.81	0.65	1.99	64.21	19.05
	60440	E	Statistics/Probability	8	1	10	5.96	0.60	2.26	59.45	18.73
	60440	F	Algebraic Relationships	9	1	10	7.59	0.76	2.22	75.46	19.66
8	60356	A	Mathematical Processes	3	3	9	5.56	0.62	2.46	62.87	24.36
	60356	B	Number Operations	7	0	7	4.26	0.61	1.83	61.21	21.71
	60356	C	Geometry	8	1	9	5.15	0.57	2.08	56.75	19.11
	60356	D	Measurement	11	1	12	7.82	0.65	2.81	65.15	21.22
	60356	E	Statistics/Probability	8	1	10	4.81	0.48	2.59	48.05	22.71
	60356	F	Algebraic Relationships	14	1	15	10.33	0.69	3.11	68.39	18.70
10	63647	A	Mathematical Processes	6	1	8	5.07	0.63	2.01	63.75	21.92
	63647	B	Number Operations	7	0	7	4.21	0.60	1.89	59.36	22.40
	63647	C	Geometry	8	1	10	5.46	0.55	2.58	55.05	22.37
	63647	D	Measurement	9	1	11	6.39	0.58	2.82	58.43	23.47
	63647	E	Statistics/Probability	8	0	8	4.39	0.55	1.95	54.75	19.99
	63647	F	Algebraic Relationships	10	1	12	7.35	0.61	3.08	61.21	23.35

Table 8-59
 Summary Statistics for Language Arts Content Standards Raw and SPI Scores

Grade	N	Content Standard	Standard	No. of Items		Total Score Points	Mean	Mean <i>p</i> -value	SD	SPI	
				MC	CR					Mean	SD
4	58848	B	Writing	20	0	20	13.02	0.65	3.71	65.32	17.66
	58848	D	Language	4	0	4	2.36	0.59	1.09	58.89	18.37
	58848	F	Research and Inquiry	6	0	6	3.36	0.56	1.61	56.51	20.69
8	60186	B	Writing	17	0	17	13.26	0.78	3.36	77.99	19.17
	60186	D	Language	6	0	6	4.20	0.70	1.55	69.91	22.27
	60186	F	Research and Inquiry	6	0	6	3.95	0.66	1.61	65.84	20.47
10	63286	B	Writing	15	2	24	14.88	0.62	4.12	61.15	16.35
	63286	D	Language	9	0	9	5.35	0.59	2.33	61.69	22.89
	63286	F	Research and Inquiry	6	0	6	3.51	0.59	1.61	58.96	20.96

Table 8-60
Summary Statistics for Social Studies Content Standards Raw and SPI Scores

Grade	N	Content Standard	Standard	No. of Items		Total Score Points	Mean	Mean <i>p</i> -value	SD	SPI	
				MC	CR					Mean	SD
4	59006	A	Geography	8	0	8	6.01	0.75	1.63	75.05	16.68
	59006	B	History	8	0	8	6.24	0.78	1.53	78.40	15.48
	59006	C	Political Science	7	0	7	5.07	0.72	1.48	72.33	16.71
	59006	D	Economics	6	0	6	4.78	0.80	1.20	80.26	16.39
	59006	E	Behavioral Science	7	0	7	5.49	0.78	1.56	78.44	18.44
8	60230	A	Geography	10	0	10	7.73	0.77	1.90	77.30	16.47
	60230	B	History	13	0	13	8.75	0.67	2.71	67.65	18.60
	60230	C	Political Science	6	0	6	4.18	0.70	1.49	69.18	19.31
	60230	D	Economics	6	0	6	4.62	0.77	1.36	76.59	18.63
	60230	E	Behavioral Science	5	0	5	3.13	0.63	1.32	63.70	19.83
10	63252	A	Geography	10	0	10	6.74	0.67	2.26	66.56	18.87
	63252	B	History	12	0	12	6.74	0.56	2.59	56.92	18.46
	63252	C	Political Science	12	0	12	7.87	0.66	2.61	65.71	19.56
	63252	D	Economics	8	0	8	6.25	0.78	1.78	77.56	19.24
	63252	E	Behavioral Science	8	0	8	4.79	0.60	1.89	60.72	19.50

Table 8-61
Summary Statistics for Science Content Standards Raw and SPI Scores

Grade	N	Content Standard	Standard	No. of Items		Total Score Points	Mean	Mean <i>p</i> -value	SD	SPI	
				MC	CR					Mean	SD
4	59017	A/B	Connections & Nature of Sci	8	0	8	5.67	0.71	1.81	71.10	19.04
	59017	C	Science Inquiry	7	0	7	4.90	0.70	1.83	70.96	21.76
	59017	D	Physical Science	6	0	6	4.16	0.69	1.17	70.34	13.68
	59017	E	Earth and Space	6	0	6	4.23	0.71	1.32	69.34	16.10
	59017	F	Life and Environment	6	0	6	4.49	0.75	1.33	72.96	16.84
	59017	G/H	Appl & Social Perspectives	7	0	7	5.29	0.76	1.64	75.57	19.77
8	60207	A/B	Connections & Nature of Sci	7	0	7	5.46	0.78	1.52	77.61	18.42
	60207	C	Science Inquiry	8	0	8	6.64	0.83	1.49	83.10	15.78
	60207	D	Physical Science	6	0	6	4.03	0.67	1.40	67.01	18.30
	60207	E	Earth and Space	6	0	6	3.95	0.66	1.48	66.01	18.56
	60207	F	Life and Environment	6	0	6	4.60	0.77	1.36	77.83	17.99
	60207	G/H	Appl & Social Perspectives	7	0	7	5.72	0.82	1.48	81.40	18.37
10	63319	A/B	Connections & Nature of Sci	10	0	10	6.72	0.67	2.38	67.08	21.30
	63319	C	Science Inquiry	10	0	10	6.73	0.67	2.30	66.53	20.12
	63319	D	Physical Science	7	0	7	3.87	0.55	1.76	55.42	19.26
	63319	E	Earth and Space	6	0	6	3.85	0.64	1.49	64.10	19.62
	63319	F	Life and Environment	7	0	7	3.76	0.54	1.83	55.46	20.67
	63319	G/H	Appl & Social Perspectives	10	0	10	6.37	0.64	2.54	64.03	22.78

Table 8-62
SPI Cut Scores

Content	Content Standard	Performance Level	SPI Cut Score Ranges						
			Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 10
RD	Standard 1 Determines Meaning	1	0-30	0-26	0-32	0-31	0-28	0-39	0-53
		2	31-49	27-44	33-55	32-47	29-37	40-55	54-74
		3	50-77	45-73	56-84	48-70	38-67	56-81	75-88
		4	78-100	74-100	85-100	71-100	68-100	82-100	89-100
	Standard 2 Understands Text	1	0-27	0-28	0-33	0-38	0-38	0-37	0-35
		2	28-54	29-46	34-53	39-54	39-53	38-54	36-48
		3	55-84	47-75	54-79	55-82	54-79	55-81	49-67
		4	85-100	76-100	80-100	83-100	80-100	82-100	68-100
	Standard 3 Analyzes Text	1	0-26	0-26	0-28	0-34	0-30	0-36	0-33
		2	27-48	27-41	29-45	35-51	31-42	37-47	34-49
		3	49-74	42-67	46-68	52-75	43-67	48-68	50-69
		4	75-100	68-100	69-100	76-100	68-100	69-100	70-100
	Standard 4 Evaluates/Extends Text	1	0-18	0-21	0-24	0-29	0-26	0-31	0-32
		2	19-29	22-39	25-40	30-43	27-38	32-46	33-47
		3	30-48	40-65	41-70	44-70	39-59	47-71	48-68
		4	49-100	66-100	71-100	71-100	60-100	72-100	69-100
MA	Standard A Mathematical Processes	1	0-33	0-23	0-42	0-29	0-22	0-26	0-37
		2	34-40	24-31	43-49	30-38	23-36	27-40	38-50
		3	41-65	32-58	50-67	39-63	37-69	41-80	51-81
		4	66-100	59-100	68-100	64-100	70-100	81-100	82-100
	Standard B Number Operations	1	0-57	0-52	0-48	0-54	0-41	0-31	0-32
		2	58-67	53-64	49-57	55-65	42-53	32-39	33-42
		3	68-89	65-86	58-80	66-87	54-84	40-73	43-77
		4	90-100	87-100	81-100	88-100	85-100	74-100	78-100
	Standard C Geometry	1	0-67	0-61	0-53	0-60	0-39	0-31	0-29
		2	68-74	62-69	54-58	61-67	40-50	32-38	30-38
		3	75-88	70-86	59-70	68-82	51-74	39-66	39-71
		4	89-100	87-100	71-100	83-100	75-100	67-100	72-100
	Standard D Measurement	1	0-55	0-51	0-50	0-36	0-36	0-32	0-30
		2	56-63	52-60	51-57	37-47	37-45	33-43	31-41
		3	64-84	61-83	58-77	48-72	46-71	44-79	42-77
		4	85-100	84-100	78-100	73-100	72-100	80-100	78-100
	Standard E Statistics/Probability	1	0-39	0-43	0-34	0-38	0-32	0-21	0-32
		2	40-47	44-52	35-41	39-48	33-41	22-26	33-40
		3	48-76	53-78	42-61	49-71	42-65	27-57	41-68
		4	77-100	79-100	62-100	72-100	66-100	58-100	69-100
	Standard F Algebraic Relationships	1	0-55	0-46	0-38	0-44	0-43	0-39	0-32
		2	56-63	47-55	39-46	45-56	44-58	40-51	33-44
		3	64-82	56-81	47-68	57-80	59-86	52-79	45-80
		4	83-100	82-100	69-100	81-100	87-100	80-100	81-100

Table 8-62 Cont'd
SPI Cut Scores

Content	Content Standard	Performance Level	SPI Cut Score Ranges						
			Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 10
LA	Standard B Writing	1		0-34				0-53	0-35
		2		35-51				54-75	36-50
		3		52-74				76-91	51-75
		4		75-100				92-100	76-100
	Standard D Language	1		0-29				0-40	0-26
		2		30-45				41-61	27-48
		3		46-68				62-87	49-85
		4		69-100				88-100	86-100
	Standard F Research and Inquiry	1		0-23				0-42	0-29
		2		24-38				43-60	30-44
		3		39-67				61-79	45-79
		4		68-100				80-100	80-100
SS	Standard A Geography	1		0-31				0-44	0-45
		2		32-46				45-63	46-50
		3		47-69				64-83	51-70
		4		70-100				84-100	71-100
	Standard B History	1		0-36				0-34	0-37
		2		37-52				35-49	38-41
		3		53-74				50-71	42-57
		4		75-100				72-100	58-100
	Standard C Political Science	1		0-29				0-36	0-43
		2		30-43				37-49	44-49
		3		44-67				50-74	50-69
		4		68-100				75-100	70-100
Standard D Economics	1		0-29				0-39	0-57	
	2		30-51				40-59	58-65	
	3		52-77				60-83	66-85	
	4		78-100				84-100	86-100	
Standard E Behavioral Science	1		0-31				0-30	0-38	
	2		32-44				31-44	39-44	
	3		45-73				45-67	45-64	
	4		74-100				68-100	65-100	

Table 8-62 Cont'd
SPI Cut Scores

Content	Content Standard	Performance Level	SPI Cut Score Ranges						
			Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 10
SC	Standard A/B Connections & Nature of Science	1		0-35				0-43	0-42
		2		36-56				44-61	43-52
		3		57-86				62-88	53-76
		4		87-100				89-100	77-100
	Standard C Science Inquiry	1		0-30				0-56	0-43
		2		31-53				57-72	44-54
		3		54-89				73-92	55-74
		4		90-100				93-100	75-100
	Standard D Physical Science	1		0-46				0-38	0-34
		2		47-60				39-50	35-40
		3		61-80				51-74	41-59
		4		81-100				75-100	60-100
	Standard E Earth and Space	1		0-40				0-38	0-42
		2		41-57				39-48	43-51
		3		58-80				49-74	52-71
		4		81-100				75-100	72-100
	Standard F Life and Environment	1		0-41				0-46	0-32
		2		42-60				47-62	33-40
		3		61-85				63-88	41-61
		4		86-100				89-100	62-100
	Standard G/H Science Applications & Social Perspectives	1		0-35				0-46	0-36
		2		36-61				47-66	37-46
		3		62-91				67-92	47-74
		4		92-100				93-100	75-100

Table 9-1
Reliability for Total Group and Subgroups Using Cronbach's Alpha

Content	Grade	Total	Gender		Race/Ethnicity					ELP		Disability		SES		
			Female	Male	White	African American	Hispanic	Asian	American Indian	Fully English Proficient	Limited English Proficient	Disabled	Not Disabled	Economically Disadvantaged	Not Economically Disadvantaged	
Reading	3	0.94	0.93	0.94	0.93	0.93	0.93	0.93	0.93	0.93	0.94	0.92	0.94	0.93	0.94	0.92
	4	0.92	0.92	0.93	0.92	0.91	0.91	0.92	0.91	0.91	0.92	0.89	0.92	0.91	0.92	0.91
	5	0.92	0.92	0.92	0.91	0.91	0.91	0.92	0.91	0.91	0.92	0.88	0.92	0.90	0.91	0.90
	6	0.92	0.91	0.92	0.90	0.91	0.90	0.92	0.91	0.91	0.91	0.87	0.91	0.89	0.91	0.89
	7	0.91	0.90	0.91	0.90	0.90	0.90	0.91	0.90	0.90	0.90	0.85	0.90	0.89	0.90	0.89
	8	0.90	0.89	0.90	0.88	0.90	0.89	0.90	0.89	0.89	0.89	0.86	0.89	0.87	0.89	0.87
	10	0.92	0.91	0.92	0.90	0.91	0.91	0.92	0.91	0.91	0.91	0.85	0.90	0.90	0.91	0.90
Mathematics	3	0.91	0.91	0.91	0.90	0.92	0.90	0.91	0.90	0.90	0.91	0.90	0.92	0.91	0.91	0.89
	4	0.91	0.91	0.91	0.90	0.91	0.90	0.91	0.90	0.90	0.91	0.90	0.92	0.90	0.91	0.89
	5	0.92	0.91	0.92	0.90	0.91	0.90	0.92	0.90	0.90	0.91	0.88	0.92	0.90	0.91	0.90
	6	0.92	0.92	0.92	0.91	0.91	0.90	0.92	0.90	0.90	0.92	0.88	0.92	0.90	0.91	0.90
	7	0.92	0.92	0.92	0.91	0.91	0.91	0.92	0.91	0.91	0.92	0.88	0.91	0.91	0.91	0.91
	8	0.92	0.92	0.93	0.92	0.89	0.90	0.93	0.90	0.90	0.92	0.86	0.88	0.91	0.91	0.92
	10	0.93	0.92	0.93	0.92	0.87	0.90	0.93	0.90	0.90	0.93	0.83	0.86	0.92	0.91	0.92
Language Arts	4	0.82	0.82	0.81	0.80	0.79	0.78	0.83	0.77	0.77	0.82	0.74	0.78	0.81	0.80	0.80
	8	0.87	0.86	0.88	0.86	0.87	0.86	0.87	0.87	0.87	0.87	0.80	0.85	0.85	0.87	0.85
	10	0.87	0.86	0.86	0.86	0.81	0.83	0.87	0.84	0.84	0.86	0.71	0.78	0.85	0.84	0.85
Social Studies	4	0.86	0.85	0.86	0.83	0.86	0.84	0.84	0.85	0.85	0.82	0.87	0.84	0.86	0.81	
	8	0.88	0.87	0.89	0.87	0.87	0.86	0.88	0.87	0.88	0.82	0.87	0.86	0.87	0.86	
	10	0.90	0.89	0.91	0.89	0.87	0.88	0.90	0.88	0.90	0.79	0.87	0.89	0.89	0.89	
Science	4	0.87	0.87	0.88	0.85	0.86	0.85	0.87	0.86	0.86	0.87	0.82	0.87	0.87	0.87	0.84
	8	0.88	0.87	0.89	0.86	0.87	0.86	0.88	0.87	0.87	0.81	0.88	0.86	0.88	0.85	
	10	0.91	0.90	0.92	0.90	0.87	0.89	0.91	0.90	0.90	0.79	0.88	0.90	0.90	0.90	

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

Content	Grade	Total	Gender		Race/Ethnicity					ELP		Disability		SES	
			Female	Male	White	African American	Hispanic	Asian	American Indian	Fully English Proficient	Limited English Proficient	Disabled	Not Disabled	Economically Disadvantaged	Not Economically Disadvantaged
Reading	3	3.03	2.99	3.06	2.95	3.28	3.24	3.09	3.17	3.00	3.33	3.32	2.98	3.21	2.88
	4	3.18	3.17	3.19	3.11	3.39	3.37	3.22	3.33	3.16	3.45	3.40	3.14	3.34	3.04
	5	3.10	3.07	3.12	3.02	3.35	3.32	3.14	3.25	3.07	3.45	3.39	3.05	3.29	2.94
	6	3.07	3.02	3.10	2.98	3.37	3.27	3.12	3.24	3.04	3.42	3.41	3.00	3.25	2.91
	7	3.19	3.17	3.21	3.13	3.38	3.33	3.23	3.31	3.17	3.45	3.42	3.15	3.33	3.08
	8	3.18	3.12	3.22	3.12	3.42	3.35	3.21	3.31	3.16	3.50	3.49	3.12	3.33	3.07
	10	3.07	3.05	3.07	3.00	3.32	3.27	3.14	3.20	3.05	3.42	3.33	3.02	3.24	2.96
Mathematics	3	2.98	2.97	2.98	2.89	3.26	3.16	2.95	3.13	2.96	3.17	3.23	2.94	3.15	2.82
	4	3.00	3.02	2.98	2.92	3.26	3.19	2.97	3.15	2.98	3.22	3.23	2.96	3.16	2.85
	5	3.23	3.24	3.22	3.17	3.43	3.39	3.18	3.36	3.22	3.44	3.43	3.19	3.37	3.11
	6	3.13	3.12	3.13	3.06	3.37	3.30	3.07	3.31	3.11	3.39	3.39	3.08	3.29	2.99
	7	3.28	3.28	3.28	3.21	3.50	3.46	3.22	3.45	3.26	3.54	3.51	3.23	3.45	3.14
	8	3.48	3.49	3.46	3.43	3.60	3.61	3.43	3.58	3.46	3.66	3.57	3.45	3.59	3.38
	10	3.29	3.30	3.28	3.26	3.35	3.38	3.27	3.40	3.29	3.36	3.35	3.27	3.38	3.23
Language Arts	4	2.32	2.29	2.36	2.28	2.45	2.43	2.33	2.41	2.31	2.47	2.46	2.30	2.41	2.25
	8	2.03	1.95	2.09	1.96	2.27	2.19	2.03	2.18	2.01	2.33	2.36	1.97	2.19	1.90
	10	2.63	2.56	2.66	2.59	2.76	2.71	2.63	2.74	2.62	2.74	2.73	2.59	2.72	2.56
Social Studies	4	2.24	2.23	2.25	2.15	2.53	2.43	2.29	2.38	2.21	2.51	2.46	2.20	2.41	2.09
	8	2.48	2.47	2.47	2.40	2.76	2.66	2.49	2.63	2.46	2.79	2.81	2.42	2.66	2.34
	10	2.96	2.97	2.94	2.90	3.20	3.13	2.98	3.10	2.95	3.25	3.21	2.92	3.13	2.86
Science	4	2.50	2.51	2.49	2.41	2.77	2.69	2.53	2.64	2.48	2.76	2.72	2.46	2.66	2.36
	8	2.37	2.38	2.35	2.28	2.70	2.61	2.45	2.53	2.34	2.78	2.75	2.31	2.57	2.22
	10	3.02	3.06	2.98	2.98	3.22	3.19	3.06	3.14	3.01	3.27	3.21	3.00	3.17	2.94

Table 9-3
Cronbach's Alpha Reliability Coefficients for Content Standards

Content Area	Grade	Alpha Per Content Standard								
		A/1	A/B	B/2	C/3	D/4	E	F	G/H	Total
Reading	3	0.76		0.87	0.83	0.52				0.94
	4	0.69		0.82	0.81	0.56				0.92
	5	0.73		0.78	0.80	0.63				0.92
	6	0.66		0.74	0.78	0.72				0.92
	7	0.69		0.75	0.76	0.64				0.91
	8	0.66		0.73	0.71	0.65				0.90
	10	0.62		0.58	0.83	0.76				0.92
Mathematics	3	0.57		0.77	0.65	0.61	0.69	0.63		0.91
	4	0.61		0.71	0.60	0.64	0.66	0.66		0.91
	5	0.55		0.71	0.55	0.68	0.68	0.73		0.92
	6	0.57		0.74	0.59	0.70	0.67	0.72		0.92
	7	0.62		0.75	0.69	0.61	0.63	0.71		0.92
	8	0.62		0.62	0.60	0.73	0.68	0.75		0.92
	10	0.62		0.63	0.68	0.74	0.60	0.76		0.93
Language Arts	4			0.75		0.35		0.56		0.82
	8			0.82		0.61		0.58		0.87
	10			0.76		0.70		0.55		0.87
Social Studies	4	0.56		0.55	0.48	0.50	0.63			0.86
	8	0.66		0.70	0.54	0.59	0.50			0.88
	10	0.67		0.64	0.69	0.68	0.60			0.90
Science	4		0.60		0.67	0.33	0.50	0.50	0.61	0.87
	8		0.59		0.60	0.50	0.49	0.55	0.63	0.88
	10		0.69		0.68	0.54	0.55	0.57	0.73	0.91

Table 9-4
Standard Error of Measurement per Content Standard

Content Area	Grade	SEM Per Content Standard								
		A/1	A/B	B/2	C/3	D/4	E	F	G/H	Total
Reading	3	1.35		1.63	1.85	1.14				3.03
	4	1.38		1.76	1.88	1.24				3.18
	5	1.26		1.61	1.94	1.23				3.10
	6	1.26		1.43	1.83	1.56				3.07
	7	1.29		1.39	2.06	1.55				3.19
	8	1.18		1.44	2.02	1.58				3.18
	10	0.84		1.11	2.12	1.74				3.07
Mathematics	3	1.50		1.19	1.09	1.08	1.14	1.17		2.98
	4	1.40		1.20	1.14	1.15	1.08	1.31		3.00
	5	1.34		1.24	1.41	1.18	1.35	1.34		3.23
	6	1.34		1.18	1.12	1.30	1.37	1.28		3.13
	7	1.41		1.34	1.43	1.24	1.37	1.20		3.28
	8	1.52		1.13	1.32	1.46	1.47	1.56		3.48
	10	1.24		1.15	1.46	1.44	1.23	1.51		3.29
Language Arts	4			1.86		0.88		1.07		2.32
	8			1.43		0.97		1.04		2.03
	10			2.02		1.28		1.08		2.63
Social Studies	4	1.08		1.03	1.07	0.85	0.95			2.24
	8	1.11		1.48	1.01	0.87	0.93			2.48
	10	1.30		1.55	1.45	1.01	1.20			2.96
Science	4		1.14		1.05	0.96	0.93	0.94	1.02	2.50
	8		0.97		0.94	0.99	1.06	0.91	0.90	2.37
	10		1.33		1.30	1.19	1.00	1.20	1.32	3.02

Table 9-5
 Classification Consistency and Classification Accuracy for Reading Grade 3

Contingency Table with All Cut Scores

	Minimal Performance	Basic Proficient	Proficient	Advanced	Sum
Minimal Performance	0.05	0.02	0.00	0.00	0.07
Basic Proficient	0.02	0.12	0.02	0.00	0.16
Proficient	0.00	0.03	0.27	0.05	0.34
Advanced	0.00	0.00	0.05	0.38	0.42
Sum	0.07	0.17	0.34	0.42	.

Indexes for Classification Consistency and Classification Accuracy

	Cut 1	Cut 2	Cut 3	All cuts
Classification Consistency (P)	0.96	0.95	0.91	0.82
Probability of Chance	0.87	0.64	0.51	0.33
Kappa (k)	0.70	0.85	0.81	0.73
Classification Accuracy	0.97	0.97	0.93	0.87

Table 9-6
 Classification Consistency and Classification Accuracy for Reading Grade 4

Contingency Table with All Cut Scores

	Minimal Performance	Basic Proficient	Proficient	Advanced	Sum
Minimal Performance	0.04	0.02	0.00	0.00	0.06
Basic Proficient	0.02	0.09	0.03	0.00	0.15
Proficient	0.00	0.03	0.30	0.04	0.37
Advanced	0.00	0.00	0.05	0.37	0.42
Sum	0.06	0.14	0.38	0.41	.

Indexes for Classification Consistency and Classification Accuracy

	Cut 1	Cut 2	Cut 3	All cuts
Classification Consistency (P)	0.96	0.94	0.91	0.80
Probability of Chance	0.88	0.67	0.51	0.34
Kappa (k)	0.65	0.81	0.81	0.70
Classification Accuracy	0.97	0.95	0.93	0.86

Table 9-7
 Classification Consistency and Classification Accuracy for Reading Grade 5

Contingency Table with All Cut Scores

	Minimal Performance	Basic Proficient	Proficient	Advanced	Sum
Minimal Performance	0.04	0.02	0.00	0.00	0.06
Basic Proficient	0.02	0.10	0.03	0.00	0.15
Proficient	0.00	0.03	0.31	0.05	0.39
Advanced	0.00	0.00	0.06	0.35	0.40
Sum	0.06	0.14	0.40	0.40	.

Indexes for Classification Consistency and Classification Accuracy

	Cut 1	Cut 2	Cut 3	All cuts
Classification Consistency (P)	0.97	0.94	0.89	0.80
Probability of Chance	0.89	0.67	0.52	0.34
Kappa (k)	0.69	0.82	0.78	0.69
Classification Accuracy	0.98	0.96	0.92	0.86

Table 9-8
 Classification Consistency and Classification Accuracy for Reading Grade 6

Contingency Table with All Cut Scores

	Minimal Performance	Basic Proficient	Proficient	Advanced	Sum
Minimal Performance	0.04	0.01	0.00	0.00	0.06
Basic Proficient	0.01	0.07	0.03	0.00	0.11
Proficient	0.00	0.02	0.31	0.05	0.39
Advanced	0.00	0.00	0.07	0.38	0.45
Sum	0.06	0.11	0.40	0.43	

Indexes for Classification Consistency and Classification Accuracy

	Cut 1	Cut 2	Cut 3	All cuts
Classification Consistency (P)	0.97	0.95	0.88	0.80
Probability of Chance	0.90	0.72	0.51	0.36
Kappa (k)	0.75	0.81	0.76	0.69
Classification Accuracy	0.98	0.96	0.92	0.86

Table 9-9
 Classification Consistency and Classification Accuracy for Reading Grade 7

Contingency Table with All Cut Scores

	Minimal Performance	Basic Proficient	Proficient	Advanced	Sum
Minimal Performance	0.03	0.01	0.00	0.00	0.05
Basic Proficient	0.01	0.04	0.03	0.00	0.08
Proficient	0.00	0.02	0.26	0.05	0.33
Advanced	0.00	0.00	0.07	0.47	0.54
Sum	0.05	0.08	0.35	0.52	

Indexes for Classification Consistency and Classification Accuracy

	Cut 1	Cut 2	Cut 3	All cuts
Classification Consistency (P)	0.97	0.95	0.88	0.81
Probability of Chance	0.91	0.78	0.50	0.41
Kappa (k)	0.70	0.76	0.77	0.67
Classification Accuracy	0.98	0.96	0.92	0.87

Table 9-10
 Classification Consistency and Classification Accuracy for Reading Grade 8

Contingency Table with All Cut Scores

	Minimal Performance	Basic Proficient	Proficient	Advanced	Sum
Minimal Performance	0.04	0.02	0.00	0.00	0.06
Basic Proficient	0.01	0.06	0.03	0.00	0.10
Proficient	0.00	0.03	0.31	0.06	0.41
Advanced	0.00	0.00	0.07	0.37	0.44
Sum	0.06	0.10	0.41	0.43	

Indexes for Classification Consistency and Classification Accuracy

	Cut 1	Cut 2	Cut 3	All cuts
Classification Consistency (P)	0.97	0.94	0.87	0.78
Probability of Chance	0.89	0.73	0.51	0.37
Kappa (k)	0.71	0.78	0.74	0.65
Classification Accuracy	0.98	0.96	0.91	0.84

Table 9-11
 Classification Consistency and Classification Accuracy for Reading Grade 10

Contingency Table with All Cut Scores

	Minimal Performance	Basic Proficient	Proficient	Advanced	Sum
Minimal Performance	0.07	0.02	0.00	0.00	0.10
Basic Proficient	0.02	0.08	0.04	0.00	0.14
Proficient	0.00	0.03	0.20	0.06	0.29
Advanced	0.00	0.00	0.05	0.42	0.47
Sum	0.09	0.14	0.28	0.48	

Indexes for Classification Consistency and Classification Accuracy

	Cut 1	Cut 2	Cut 3	All cuts
Classification Consistency (P)	0.96	0.93	0.89	0.77
Probability of Chance	0.82	0.64	0.50	0.34
Kappa (k)	0.74	0.80	0.78	0.66
Classification Accuracy	0.97	0.95	0.92	0.84

Table 9-12
 Classification Consistency and Classification Accuracy for Mathematics Grade 3

Contingency Table with All Cut Scores

	Minimal Performance	Basic Proficient	Proficient	Advanced	Sum
Minimal Performance	0.12	0.02	0.00	0.00	0.14
Basic Proficient	0.02	0.03	0.03	0.00	0.08
Proficient	0.01	0.03	0.27	0.06	0.36
Advanced	0.00	0.00	0.06	0.36	0.41
Sum	0.15	0.08	0.36	0.42	.

Indexes for Classification Consistency and Classification Accuracy

	Cut 1	Cut 2	Cut 3	All cuts
Classification Consistency (P)	0.94	0.94	0.88	0.77
Probability of Chance	0.75	0.65	0.51	0.33
Kappa (k)	0.77	0.82	0.76	0.66
Classification Accuracy	0.96	0.95	0.92	0.84

Table 9-13
 Classification Consistency and Classification Accuracy for Mathematics Grade 4

Contingency Table with All Cut Scores

	Minimal Performance	Basic Proficient	Proficient	Advanced	Sum
Minimal Performance	0.11	0.02	0.01	0.00	0.14
Basic Proficient	0.02	0.03	0.03	0.00	0.08
Proficient	0.01	0.03	0.28	0.06	0.37
Advanced	0.00	0.00	0.06	0.36	0.42
Sum	0.14	0.08	0.37	0.42	

Indexes for Classification Consistency and Classification Accuracy

	Cut 1	Cut 2	Cut 3	All cuts
Classification Consistency (P)	0.94	0.93	0.89	0.77
Probability of Chance	0.76	0.66	0.51	0.33
Kappa (k)	0.77	0.79	0.77	0.66
Classification Accuracy	0.96	0.95	0.92	0.84

Table 9-14
 Classification Consistency and Classification Accuracy for Mathematics Grade 5

Contingency Table with All Cut Scores

	Minimal Performance	Basic Proficient	Proficient	Advanced	Sum
Minimal Performance	0.10	0.02	0.00	0.00	0.13
Basic Proficient	0.02	0.03	0.03	0.00	0.08
Proficient	0.01	0.03	0.21	0.05	0.31
Advanced	0.00	0.00	0.05	0.43	0.48
Sum	0.14	0.08	0.30	0.48	

Indexes for Classification Consistency and Classification Accuracy

	Cut 1	Cut 2	Cut 3	All cuts
Classification Consistency (P)	0.94	0.92	0.89	0.77
Probability of Chance	0.77	0.66	0.50	0.35
Kappa (k)	0.74	0.77	0.78	0.65
Classification Accuracy	0.96	0.95	0.93	0.84

Table 9-15
 Classification Consistency and Classification Accuracy for Mathematics Grade 6

Contingency Table with All Cut Scores

	Minimal Performance	Basic Proficient	Proficient	Advanced	Sum
Minimal Performance	0.09	0.02	0.00	0.00	0.11
Basic Proficient	0.02	0.04	0.03	0.00	0.10
Proficient	0.00	0.03	0.30	0.06	0.38
Advanced	0.00	0.00	0.06	0.35	0.41
Sum	0.11	0.09	0.39	0.41	

Indexes for Classification Consistency and Classification Accuracy

	Cut 1	Cut 2	Cut 3	All cuts
Classification Consistency (P)	0.95	0.94	0.88	0.78
Probability of Chance	0.80	0.67	0.52	0.34
Kappa (k)	0.77	0.80	0.76	0.67
Classification Accuracy	0.97	0.95	0.92	0.84

Table 9-16
 Classification Consistency and Classification Accuracy for Mathematics Grade 7

Contingency Table with All Cut Scores

	Minimal Performance	Basic Proficient	Proficient	Advanced	Sum
Minimal Performance	0.06	0.02	0.00	0.00	0.08
Basic Proficient	0.02	0.05	0.03	0.00	0.10
Proficient	0.00	0.03	0.33	0.06	0.42
Advanced	0.00	0.00	0.05	0.35	0.40
Sum	0.08	0.10	0.42	0.41	.

Indexes for Classification Consistency and Classification Accuracy

	Cut 1	Cut 2	Cut 3	All cuts
Classification Consistency (P)	0.96	0.94	0.89	0.79
Probability of Chance	0.85	0.71	0.52	0.35
Kappa (k)	0.75	0.79	0.77	0.68
Classification Accuracy	0.97	0.95	0.93	0.86

Table 9-17
 Classification Consistency and Classification Accuracy for Mathematics Grade 8

Contingency Table with All Cut Scores

	Minimal Performance	Basic Proficient	Proficient	Advanced	Sum
Minimal Performance	0.07	0.03	0.00	0.00	0.10
Basic Proficient	0.02	0.05	0.04	0.00	0.11
Proficient	0.00	0.03	0.38	0.04	0.45
Advanced	0.00	0.00	0.04	0.29	0.34
Sum	0.09	0.11	0.47	0.33	

Indexes for Classification Consistency and Classification Accuracy

	Cut 1	Cut 2	Cut 3	All cuts
Classification Consistency (P)	0.95	0.93	0.92	0.79
Probability of Chance	0.83	0.67	0.55	0.34
Kappa (k)	0.68	0.77	0.81	0.68
Classification Accuracy	0.96	0.95	0.95	0.86

Table 9-18
 Classification Consistency and Classification Accuracy for Mathematics Grade 10

Contingency Table with All Cut Scores

	Minimal Performance	Basic Proficient	Proficient	Advanced	Sum
Minimal Performance	0.14	0.04	0.00	0.00	0.18
Basic Proficient	0.03	0.07	0.05	0.00	0.15
Proficient	0.00	0.05	0.36	0.04	0.45
Advanced	0.00	0.00	0.03	0.18	0.22
Sum	0.18	0.15	0.45	0.22	.

Indexes for Classification Consistency and Classification Accuracy

	Cut 1	Cut 2	Cut 3	All cuts
Classification Consistency (P)	0.92	0.90	0.93	0.76
Probability of Chance	0.70	0.56	0.66	0.30
Kappa (k)	0.73	0.78	0.79	0.65
Classification Accuracy	0.94	0.94	0.95	0.83

Table 9-19
 Classification Consistency and Classification Accuracy for Language Arts Grade 4

Contingency Table with All Cut Scores

	Minimal Performance	Basic Proficient	Proficient	Advanced	Sum
Minimal Performance	0.04	0.02	0.00	0.00	0.06
Basic Proficient	0.02	0.08	0.05	0.00	0.15
Proficient	0.00	0.06	0.27	0.07	0.41
Advanced	0.00	0.00	0.08	0.30	0.37
Sum	0.06	0.16	0.40	0.37	

Indexes for Classification Consistency and Classification Accuracy

	Cut 1	Cut 2	Cut 3	All cuts
Classification Consistency (P)	0.95	0.88	0.85	0.69
Probability of Chance	0.88	0.66	0.53	0.33
Kappa (k)	0.59	0.66	0.68	0.54
Classification Accuracy	0.96	0.92	0.89	0.77

Table 9-20
 Classification Consistency and Classification Accuracy for Language Arts Grade 8

Contingency Table with All Cut Scores

	Minimal Performance	Basic Proficient	Proficient	Advanced	Sum
Minimal Performance	0.10	0.03	0.00	0.00	0.13
Basic Proficient	0.03	0.12	0.05	0.00	0.21
Proficient	0.00	0.06	0.21	0.07	0.34
Advanced	0.00	0.00	0.08	0.25	0.32
Sum	0.13	0.22	0.34	0.32	

Indexes for Classification Consistency and Classification Accuracy

	Cut 1	Cut 2	Cut 3	All cuts
Classification Consistency (P)	0.94	0.88	0.85	0.68
Probability of Chance	0.78	0.55	0.56	0.28
Kappa (k)	0.72	0.74	0.66	0.55
Classification Accuracy	0.96	0.92	0.89	0.77

Table 9-21
 Classification Consistency and Classification Accuracy for Language Arts Grade 10

Contingency Table with All Cut Scores

	Minimal Performance	Basic Proficient	Proficient	Advanced	Sum
Minimal Performance	0.09	0.03	0.00	0.00	0.12
Basic Proficient	0.04	0.12	0.06	0.00	0.22
Proficient	0.00	0.06	0.38	0.05	0.49
Advanced	0.00	0.00	0.05	0.12	0.17
Sum	0.13	0.21	0.49	0.17	

Indexes for Classification Consistency and Classification Accuracy

	Cut 1	Cut 2	Cut 3	All cuts
Classification Consistency (P)	0.93	0.88	0.90	0.72
Probability of Chance	0.78	0.55	0.72	0.33
Kappa (k)	0.68	0.73	0.66	0.58
Classification Accuracy	0.95	0.91	0.93	0.79

Table 9-22
 Classification Consistency and Classification Accuracy for Social Studies Grade 4

Contingency Table with All Cut Scores

	Minimal Performance	Basic Proficient	Proficient	Advanced	Sum
Minimal Performance	0.03	0.01	0.00	0.00	0.04
Basic Proficient	0.01	0.03	0.01	0.00	0.04
Proficient	0.00	0.01	0.15	0.02	0.18
Advanced	0.00	0.00	0.02	0.71	0.74
Sum	0.04	0.04	0.18	0.74	

Indexes for Classification Consistency and Classification Accuracy

	Cut 1	Cut 2	Cut 3	All cuts
Classification Consistency (P)	0.99	0.98	0.95	0.92
Probability of Chance	0.93	0.85	0.61	0.58
Kappa (k)	0.84	0.84	0.88	0.80
Classification Accuracy	0.98	0.98	0.93	0.90

Table 9-23
 Classification Consistency and Classification Accuracy for Social Studies Grade 8

Contingency Table with All Cut Scores

	Minimal Performance	Basic Proficient	Proficient	Advanced	Sum
Minimal Performance	0.05	0.02	0.00	0.00	0.08
Basic Proficient	0.02	0.07	0.04	0.00	0.14
Proficient	0.00	0.03	0.24	0.07	0.34
Advanced	0.00	0.00	0.05	0.39	0.44
Sum	0.08	0.13	0.34	0.45	

Indexes for Classification Consistency and Classification Accuracy

	Cut 1	Cut 2	Cut 3	All cuts
Classification Consistency (P)	0.95	0.92	0.88	0.76
Probability of Chance	0.86	0.67	0.51	0.34
Kappa (k)	0.68	0.77	0.76	0.63
Classification Accuracy	0.97	0.95	0.92	0.83

Table 9-24
 Classification Consistency and Classification Accuracy for Social Studies Grade 10

Contingency Table with All Cut Scores

	Minimal Performance	Basic Proficient	Proficient	Advanced	Sum
Minimal Performance	0.16	0.02	0.02	0.00	0.20
Basic Proficient	0.02	0.02	0.03	0.00	0.07
Proficient	0.02	0.03	0.19	0.06	0.29
Advanced	0.00	0.00	0.06	0.38	0.44
Sum	0.20	0.08	0.28	0.44	.

Indexes for Classification Consistency and Classification Accuracy

	Cut 1	Cut 2	Cut 3	All cuts
Classification Consistency (P)	0.92	0.90	0.89	0.74
Probability of Chance	0.68	0.60	0.51	0.32
Kappa (k)	0.75	0.76	0.77	0.62
Classification Accuracy	0.94	0.94	0.92	0.81

Table 9-25
 Classification Consistency and Classification Accuracy for Science Grade 4

Contingency Table with All Cut Scores

	Minimal Performance	Basic Proficient	Proficient	Advanced	Sum
Minimal Performance	0.04	0.02	0.00	0.00	0.07
Basic Proficient	0.02	0.11	0.05	0.00	0.19
Proficient	0.00	0.05	0.36	0.06	0.48
Advanced	0.00	0.00	0.07	0.20	0.27
Sum	0.07	0.19	0.48	0.26	.

Indexes for Classification Consistency and Classification Accuracy

	Cut 1	Cut 2	Cut 3	All cuts
Classification Consistency (P)	0.95	0.90	0.87	0.72
Probability of Chance	0.87	0.62	0.61	0.34
Kappa (k)	0.60	0.73	0.67	0.58
Classification Accuracy	0.96	0.93	0.91	0.80

Table 9-26
 Classification Consistency and Classification Accuracy for Science Grade 8

Contingency Table with All Cut Scores

	Minimal Performance	Basic Proficient	Proficient	Advanced	Sum
Minimal Performance	0.07	0.02	0.00	0.00	0.09
Basic Proficient	0.02	0.07	0.04	0.00	0.13
Proficient	0.00	0.04	0.31	0.07	0.42
Advanced	0.00	0.00	0.07	0.29	0.36
Sum	0.09	0.13	0.42	0.36	

Indexes for Classification Consistency and Classification Accuracy

	Cut 1	Cut 2	Cut 3	All cuts
Classification Consistency (P)	0.95	0.91	0.86	0.73
Probability of Chance	0.83	0.66	0.54	0.33
Kappa (k)	0.70	0.75	0.69	0.59
Classification Accuracy	0.97	0.94	0.91	0.81

Table 9-27
 Classification Consistency and Classification Accuracy for Science Grade 10

Contingency Table with All Cut Scores

	Minimal Performance	Basic Proficient	Proficient	Advanced	Sum
Minimal Performance	0.17	0.03	0.01	0.00	0.21
Basic Proficient	0.04	0.05	0.04	0.00	0.12
Proficient	0.01	0.04	0.23	0.05	0.33
Advanced	0.00	0.00	0.04	0.30	0.34
Sum	0.22	0.12	0.32	0.35	.

Indexes for Classification Consistency and Classification Accuracy

	Cut 1	Cut 2	Cut 3	All cuts
Classification Consistency (P)	0.91	0.9	0.91	0.74
Probability of Chance	0.66	0.55	0.55	0.28
Kappa (k)	0.74	0.77	0.80	0.64
Classification Accuracy	0.94	0.93	0.93	0.81

Table 9-28
Inter-Rater Reliability, Reading*

Grade	Item No.	Max	Percentage Absolute Difference				Codes	Intra. Corr.	Weighted Kappa	Mean	No. of Reads	Frequency			
			Perfect	Adjacent	Discrepant							0	1	2	3
3	31	3	0.69	0.27	0.01	0.03	0.90	0.80	1.18	6078	1529	2157	2147	245	
3	56	3	0.74	0.22	0.02	0.03	0.89	0.79	0.98	6078	1825	2736	1363	154	
4	13	3	0.71	0.25	0.02	0.02	0.86	0.71	1.07	6340	1400	3289	1460	191	
4	56	3	0.79	0.15	0.02	0.04	0.95	0.89	1.01	6340	2711	1126	2252	251	
5	33	3	0.72	0.25	0.01	0.02	0.83	0.67	0.84	6124	1782	3592	691	59	
5	50	3	0.68	0.27	0.03	0.02	0.86	0.72	1.09	6124	1679	2306	2037	102	
6	38	3	0.65	0.31	0.02	0.02	0.89	0.77	1.39	6392	1269	1964	2531	628	
6	56	3	0.57	0.39	0.03	0.01	0.81	0.62	1.42	6392	889	2400	2666	437	
7	36	3	0.75	0.20	0.03	0.02	0.92	0.84	1.51	6258	1560	842	2953	903	
7	56	3	0.75	0.22	0.02	0.02	0.88	0.75	0.62	6258	3328	2094	743	93	
8	19	3	0.65	0.31	0.02	0.03	0.88	0.76	1.47	6324	979	2179	2412	754	
8	40	3	0.69	0.26	0.02	0.03	0.89	0.78	1.22	6324	1775	1644	2650	255	
10	12	3	0.67	0.28	0.01	0.04	0.91	0.81	1.11	6618	1965	2338	1934	381	
10	43	3	0.62	0.28	0.02	0.08	0.91	0.83	1.41	6618	1550	1833	2229	1006	

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

Table 9-29
Inter-Rater Reliability, Mathematics*

Grade	Item No.	Max	Percentage Absolute Difference				Codes	Intra. Corr.	Weighted Kappa	Mean	No. of Reads	Frequency		
			Perfect	Adjacent	Discrepant							0	1	2
3	10	2	0.84	0.15	0.00	0.01	0.92	0.84	1.28	6078	960	2439	2679	
3	25A	1	0.96	0.02	0.00	0.02	0.98	0.96	0.49	6078	3084	2994		
3	25B	2	0.78	0.17	0.02	0.03	0.92	0.84	0.84	6078	2677	1729	1672	
3	28A	1	0.98	0.00	0.00	0.01	0.98	0.97	0.96	6078	258	5820		
3	28B	2	0.91	0.05	0.01	0.02	0.97	0.94	1.44	6078	1483	459	4136	
3	44A	1	0.96	0.02	0.00	0.02	0.98	0.95	0.71	6078	1765	4313		
3	44B	2	0.83	0.09	0.06	0.03	0.91	0.82	0.65	6078	3851	490	1737	
4	13	2	0.94	0.05	0.00	0.01	0.98	0.97	0.92	6340	2885	1060	2395	
4	20A	1	0.99	0.01	0.00	0.01	1.00	0.99	0.50	6340	3165	3175		
4	20B	2	0.85	0.09	0.03	0.03	0.92	0.84	0.73	6340	2923	2235	1182	
4	29A	1	0.98	0.01	0.00	0.01	1.00	0.99	0.72	6340	1773	4567		
4	29B	2	0.87	0.08	0.02	0.03	0.94	0.88	0.55	6340	3928	1312	1100	
4	41A	1	0.99	0.01	0.00	0.01	1.00	0.99	0.69	6340	1996	4344		
4	41B	2	0.83	0.14	0.01	0.02	0.93	0.86	1.29	6340	1510	1514	3316	
5	12A	1	0.97	0.02	0.00	0.01	0.98	0.97	0.59	6124	2520	3604		
5	12B	2	0.87	0.11	0.01	0.02	0.93	0.86	1.12	6124	1095	3201	1828	
5	19	2	0.92	0.06	0.00	0.01	0.95	0.90	1.56	6124	463	1767	3894	
5	23A	1	0.98	0.01	0.00	0.01	1.00	0.99	0.39	6124	3720	2404		
5	23B	2	0.91	0.07	0.01	0.01	0.97	0.94	0.72	6124	3715	382	2027	
5	46A	1	0.99	0.01	0.00	0.01	0.99	0.98	0.87	6124	811	5313		
5	46B	2	0.93	0.05	0.02	0.01	0.90	0.81	1.83	6124	393	256	5475	
6	10A	1	0.99	0.00	0.00	0.01	1.00	0.99	0.62	6392	2416	3976		
6	10B	2	0.89	0.09	0.01	0.01	0.96	0.92	1.29	6392	1925	698	3769	
6	22A	1	0.99	0.00	0.00	0.01	0.99	0.97	0.92	6392	522	5870		
6	22B	2	0.93	0.06	0.00	0.01	0.95	0.89	0.52	6392	3212	3066	114	

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

Table 9-29 Cont'd
Inter-Rater Reliability, Mathematics*

Grade	Item No.	Max	Percentage Absolute Difference				Codes	Intra. Corr.	Weighted Kappa	Mean	No. of Reads	Frequency		
			Perfect	Adjacent	Discrepant							0	1	2
6	35A	1	0.98	0.01	0.00	0.01	0.99	0.98	0.32	6392	4374	2018		
6	35B	2	0.85	0.12	0.01	0.02	0.95	0.89	0.79	6392	2897	1941	1554	
6	53	2	0.97	0.01	0.00	0.01	0.99	0.99	1.24	6392	1325	2236	2831	
7	4A	1	0.99	0.00	0.00	0.01	1.00	0.99	0.69	6258	1955	4303		
7	4B	2	0.89	0.09	0.01	0.02	0.96	0.93	1.23	6258	2008	791	3459	
7	29A	1	0.96	0.02	0.00	0.02	0.98	0.96	0.49	6258	3174	3084		
7	29B	2	0.86	0.09	0.02	0.02	0.95	0.90	1.27	6258	1968	657	3633	
7	32A	1	0.98	0.01	0.00	0.01	0.99	0.98	0.47	6258	3311	2947		
7	32B	2	0.93	0.05	0.00	0.02	0.96	0.91	0.66	6258	2338	3695	225	
7	51	2	0.97	0.02	0.00	0.01	0.99	0.98	0.81	6258	2571	2327	1360	
8	9A	1	0.98	0.01	0.00	0.01	0.98	0.95	0.90	6324	642	5682		
8	9B	2	0.89	0.08	0.02	0.01	0.94	0.88	1.52	6324	1301	447	4576	
8	20A	1	0.97	0.01	0.00	0.02	0.99	0.98	0.42	6324	3652	2672		
8	20B	2	0.86	0.11	0.01	0.02	0.95	0.91	1.06	6324	2125	1712	2487	
8	40A	1	0.98	0.01	0.00	0.02	1.00	0.99	0.46	6324	3429	2895		
8	40B	2	0.90	0.07	0.01	0.02	0.98	0.95	0.89	6324	3347	302	2675	
8	53	2	0.91	0.05	0.00	0.04	0.98	0.96	0.56	6324	4234	639	1451	
10	27	2	0.82	0.11	0.00	0.07	0.96	0.92	0.80	6618	3151	1650	1817	
10	33	2	0.85	0.11	0.00	0.04	0.95	0.91	1.31	6618	1363	1821	3434	
10	38	2	0.87	0.02	0.01	0.10	0.99	0.98	0.49	6618	4879	247	1492	
10	52	2	0.85	0.08	0.01	0.06	0.97	0.93	0.86	6618	3005	1520	2093	

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

Table 9-30
Inter-Rater Reliability, Writing Prompts*

Grade	Item No.	Max Score	Percentage Absolute Difference				Codes	Intra. Corr.	Weighted Kappa	Mean	No. of Reads	Frequency					
			P	A	D							0	1	2	3	4	5
4	1A	6	0.64	0.33	0.02	2	0.90	0.80	3.33	6340	117	113	920	2377	2069	685	59
4	1B	3	0.88	0.11	0.00	1	0.78	0.56	1.97	6340	59	340	5705	236			
8	1A	6	0.58	0.36	0.03	3	0.88	0.76	3.16	6324	199	77	1097	2524	2064	325	38
8	1B	3	0.92	0.06	0.00	1	0.85	0.70	1.97	6324	94	147	5960	123			
10	1A	6	0.58	0.37	0.02	3	0.92	0.83	3.47	6618	205	110	784	2064	2336	969	150
10	1B	3	0.74	0.23	0.00	3	0.80	0.61	2.12	6618	182	93	5114	1229			

* Note that P is percent perfect agreement, A is percent adjacent agreement, and D is percent discrepant. Also, note that the sum of the modes of agreement and codes may not equal exactly 100% due to rounding.

Table 10-1
Items Flagged for DIF, By Gender

Content	Grade	Test Book Number	Item Type	Female			Male			SMD	Delta	LH Flag Female	LH Flag Male	Flag MH
				D+	D-	Z	D+	D-	Z					
RD	5	5	MC	0.01	-0.06	-7.95	0.05	0.00	8.00	-0.25	-1.93			-C
RD	6	56	CR	0.10	0.00	6.74	0.00	-0.11	-7.68	0.26			-CC	CC
RD	8	19	CR	0.12	0.00	8.69	0.01	-0.16	-10.06	0.30		CC	-CC	CC
RD	10	4	MC	0.02	-0.06	-6.94	0.05	-0.01	5.39	-0.21	-1.50			-C
RD	10	43	CR	0.14	0.00	9.35	0.03	-0.19	-11.23	0.32		CC	-CC	CC
LA	4	1A	CR	0.14	0.00	9.40	0.02	-0.16	-9.40	0.31		CC	-CC	CC
LA	8	1A	CR	0.07	-0.07	4.12	0.16	-0.15	-7.74	0.21			-CC	BB
SS	8	14	MC	0.02	0.00	4.32	0.00	-0.01	-2.90	0.14	1.71			C

Table 10-2
 Items Flagged for DIF, By Race/Ethnicity, African American

Content	Grade	Test Book Number	Item Type	D+	D-	Z	SMD	Delta	LH Flag	MH Flag
RD	4	56	CR	0.10	-0.12	-4.20	-0.06		-CC	
RD	5	5	MC	0.01	-0.04	-2.42	-0.29	-1.54		-C
RD	7	36	CR	0.01	-0.23	-6.84	-0.27		-CC	-CC
RD	8	40	CR	0.03	-0.13	-3.60	-0.03		-CC	
RD	10	8	MC	0.01	-0.06	-3.12	-0.41	-1.98		-C
RD	10	43	CR	0.00	-0.13	-4.48	-0.09		-CC	
SS	10	13	MC	0.01	-0.03	-2.09	-0.49	-2.11		-C

Table 10-3
 Items Flagged for DIF, By Race/Ethnicity, Hispanic

Content	Grade	Test Book Number	Item Type	D+	D-	Z	SMD	Delta	LH Flag	MH Flag
RD	6	7	MC	0.00	-0.14	-8.67	-0.28	-1.34	-C	-B
RD	7	1	MC	0.00	-0.11	-6.66	-0.27	-1.70	-C	-C
RD	7	13	MC	0.11	0.00	6.17	0.23	1.21	C	B
SC	10	14	MC	0.11	0.00	6.49	0.16	0.92	C	

Table 10-4
Items Flagged for DIF, By Race/Ethnicity, Asian

Content	Grade	Test Book Number	Item Type	D+	D-	Z	SMD	Delta	LH Flag	MH Flag
RD	3	31	CR	0.25	-0.08	4.25	0.14		CC	
RD	4	10	MC	0.10	0.00	5.05	0.25	1.31	C	B
RD	4	56	CR	0.15	-0.02	3.52	0.10		CC	
RD	5	33	CR	0.13	-0.03	3.55	0.16		CC	
RD	7	36	CR	0.15	-0.07	3.44	0.07		CC	
RD	8	10	MC	0.00	-0.14	-6.95	-0.30	-1.45	-C	-B
RD	8	19	CR	0.21	0.00	6.15	0.23		CC	BB
RD	8	35	MC	0.04	0.00	2.89	0.18	1.63		C
RD	8	40	CR	0.19	0.00	5.69	0.20		CC	BB
RD	10	3	MC	0.10	0.00	5.73	0.20	1.45	C	B
RD	10	8	MC	0.01	-0.15	-6.87	-0.41	-2.62	-C	-C
RD	10	12	CR	0.14	-0.08	3.62	0.22		CC	BB
RD	10	15	MC	0.00	-0.10	-5.80	-0.26	-1.51	-C	-C
RD	10	20	MC	0.04	-0.12	-7.08	-0.22	-1.42	-C	-B
RD	10	43	CR	0.23	0.00	6.91	0.21		CC	BB
MA	5	20	MC	0.01	-0.04	-2.38	-0.24	-2.16		-C
MA	6	47	MC	0.03	-0.11	-5.48	-0.18	-1.07	-C	-B
MA	6	53	CR	0.15	-0.06	3.60	0.07		CC	
LA	4	1A	CR	0.18	0.00	4.85	0.18		CC	BB
LA	4	1B	CR	0.11	0.00	5.99	0.14		CC	
LA	8	14	MC	0.01	-0.12	-4.44	-0.24	-1.61		-C
LA	10	1	MC	0.06	-0.12	-5.10	-0.22	-1.28	-C	-B
LA	10	22	MC	0.03	-0.15	-6.41	-0.23	-1.46	-C	-B
LA	10	1A	CR	0.15	-0.06	3.37	0.23		CC	BB
SS	10	9	MC	0.00	-0.10	-5.65	-0.18	-0.98	-C	
SS	10	34	MC	0.14	-0.04	5.09	0.17	0.98	C	

Table 10-4 Cont'd
 Items Flagged for DIF, By Race/Ethnicity, Asian

Content	Grade	Test Book Number	Item Type	D+	D-	Z	SMD	Delta	LH Flag	MH Flag
SS	10	37	MC	0.07	0.00	4.28	0.24	1.58		C
SS	10	39	MC	0.02	-0.01	2.29	0.09	1.77		C
SC	10	1	MC	0.04	-0.13	-6.22	-0.17	-1.07	-C	-B

Table 10-5
 Items Flagged for DIF, By Race/Ethnicity, American Indian*

Content	Grade	Test Book Number	Item Type	D+	D-	Z	SMD	Delta	LH Flag	MH Flag
RD	3	7	MC	0.09	-0.24	-2.83	0.01	0.07	-C	
RD	5	19	MC	0.08	-0.26	-2.88	0.07	0.42	-C	
RD	10	19	MC	0.09	-0.17	-3.39	0.07	0.39	-C	
MA	5	35	MC	0.00	-0.15	-2.99	-0.15	-0.80	-C	

* Note: DIF statistics can only be calculated for items with sufficient student N counts. In some cases here, the size of the tested population was too small to include valid DIF statistics.

Table 10-6
Items Flagged for DIF, By English Language Proficiency

Content	Grade	Test Book Number	Item Type	Limited English Proficient			Fully English Proficient			SMD	Delta	LH Flag Limited English Proficient	LH Flag Fully English Proficient	MH Flag
				D+	D-	Z	D+	D-	Z					
RD	3	31	CR	0.19	-0.11	2.85	0.03	-0.05	-2.30	0.12		CC		
RD	4	10	MC	0.13	-0.02	5.96	0.03	-0.03	-2.06	0.23	1.14	C		B
RD	4	56	CR	0.16	-0.04	3.15	0.02	-0.05	-2.10	0.14		CC		
RD	5	2	MC	0.00	-0.10	-5.52	0.02	-0.01	2.33	-0.23	-1.14	-C		-B
RD	5	33	CR	0.10	0.00	3.86	0.01	-0.03	-2.31	0.11		CC		
RD	6	7	MC	0.00	-0.12	-6.14	0.02	-0.01	2.13	-0.33	-1.49	-C		-B
RD	7	1	MC	0.03	-0.10	-3.37	0.01	-0.03	1.50	-0.26	-1.72			-C
RD	8	10	MC	0.00	-0.11	-4.51	0.04	-0.03	0.34	-0.21	-1.09	-C		-B
RD	8	12	MC	0.01	-0.08	-3.57	0.01	-0.01	1.93	-0.36	-1.60			-C
RD	8	19	CR	0.21	-0.02	4.91	0.00	-0.02	-2.17	0.25		CC		CC
RD	8	40	CR	0.24	-0.08	5.48	0.02	-0.05	-2.66	0.17		CC		
RD	8	56	MC	0.07	-0.15	-5.18	0.02	-0.01	1.62	-0.19	-0.97	-C		
RD	10	8	MC	0.11	-0.12	-4.50	0.02	-0.01	1.51	-0.34	-1.66			-C
RD	10	12	CR	0.13	-0.06	3.52	0.02	-0.03	-2.28	0.20		CC		BB
RD	10	20	MC	0.04	-0.13	-5.83	0.02	-0.01	2.32	-0.28	-1.24	-C		-B
RD	10	43	CR	0.16	-0.01	3.75	0.00	-0.03	-2.16	0.13		CC		
MA	3	28B	CR	0.16	0.00	4.44	0.02	-0.03	-1.90	0.15		CC		
MA	3	35	MC	0.10	0.00	5.50	0.00	-0.02	-2.85	0.16	0.88	C		
MA	5	26	MC	0.15	-0.05	5.13	0.01	-0.03	-2.81	0.09	0.47	C		
MA	6	24	MC	0.12	0.00	5.67	0.01	-0.04	-2.10	0.09	0.51	C		
MA	6	34	MC	0.12	0.00	5.75	0.01	-0.04	-1.86	0.13	0.73	C		
MA	6	47	MC	0.00	-0.10	-5.48	0.01	0.00	1.75	-0.15	-0.74	-C		
MA	6	53	CR	0.19	0.00	6.29	0.02	-0.04	-1.88	0.15		CC		
MA	8	20B	CR	0.14	-0.03	3.00	0.03	-0.03	-2.41	0.06		CC		
MA	8	26	MC	0.07	-0.14	-4.40	0.02	-0.01	1.45	-0.16	-0.88	-C		

Table 10-6 Cont'd
 Items Flagged for DIF, By English Language Proficiency

Content	Grade	Test Book Number	Item Type	Limited English Proficient			Fully English Proficient			SMD	Delta	LH Flag Limited English Proficient	LH Flag Fully English Proficient	MH Flag
				D+	D-	Z	D+	D-	Z					
MA	8	43	MC	0.00	-0.15	-6.49	0.02	-0.01	1.50	-0.03	-0.18	-C		
MA	10	54	MC	0.14	-0.02	5.16	0.01	-0.02	-0.78	0.12	0.61	C		
LA	4	2	MC	0.02	-0.04	-0.64	0.02	0.00	4.13	-0.28	-1.66			-C
LA	4	1A	CR	0.22	-0.13	5.31	0.01	-0.03	-1.90	0.15		CC		
LA	8	8	MC	0.02	-0.14	-5.33	0.02	-0.01	0.63	-0.21	-1.00	-C		-B
LA	8	12	MC	0.00	-0.11	-5.03	0.02	-0.02	0.31	-0.13	-0.81	-C		
LA	8	15	MC	0.00	-0.12	-5.09	0.03	-0.01	1.47	-0.17	-0.87	-C		
LA	8	1A	CR	0.18	-0.10	3.66	0.03	-0.06	-3.75	0.17		CC		BB
LA	10	22	MC	0.00	-0.12	-5.14	0.01	-0.01	1.17	-0.15	-0.88	-C		
LA	10	1A	CR	0.19	-0.05	4.10	0.04	-0.04	-0.56	0.18		CC		BB
SS	8	2	MC	0.01	-0.07	-1.88	0.01	0.00	0.75	-0.30	-1.55			-C
SC	10	27	MC	0.12	-0.02	5.10	0.01	-0.02	-0.64	0.10	0.51	C		

Table 10-7
Items Flagged for DIF, By Disability Status

Content	Grade	Test Book Number	Item Type	Not Disabled			Disabled			SMD	Delta	LH Flag Not Disabled	LH Flag Disabled	MH Flag
				D+	D-	Z	D+	D-	Z					
RD	8	19	CR	0.02	-0.02	0.54	0.16	-0.13	-3.90	-0.08			-CC	
RD	10	43	CR	0.02	-0.02	-0.19	0.07	-0.14	-3.11	-0.10			-CC	
LA	4	2	MC	0.02	0.00	5.08	0.00	-0.06	-3.13	-0.40	-2.24			-C
LA	4	25	MC	0.03	-0.01	3.57	0.04	-0.11	-5.55	-0.33	-1.75			-C
LA	4	1A	CR	0.05	-0.05	2.51	0.05	-0.28	-7.62	-0.32			-CC	-CC
LA	4	1B	CR	0.02	-0.02	1.87	0.01	-0.11	-6.34	-0.32			-CC	-CC
LA	8	1A	CR	0.04	-0.06	-0.72	0.06	-0.20	-4.99	-0.20			-CC	-BB
LA	10	1A	CR	0.05	-0.05	2.08	0.10	-0.20	-4.08	-0.27				-CC

Table 10-8
Correlations among Reading Objectives

Grade	CS	1	2	3
3	2	0.78	-	-
	3	0.77	0.84	-
	4	0.59	0.63	0.65
4	2	0.73	-	-
	3	0.74	0.79	-
	4	0.64	0.67	0.69
5	2	0.73	-	-
	3	0.74	0.78	-
	4	0.64	0.68	0.71
6	2	0.66	-	-
	3	0.68	0.77	-
	4	0.68	0.74	0.76
7	2	0.67	-	-
	3	0.72	0.74	-
	4	0.65	0.67	0.72
8	2	0.68	-	-
	3	0.68	0.72	-
	4	0.66	0.69	0.70
10	2	0.55	-	-
	3	0.69	0.69	-
	4	0.66	0.67	0.81

Table 10-9
Correlations among Mathematics Objectives

Grade	CS	A	B	C	D	E
3	B	0.63	-	-	-	-
	C	0.59	0.64	-	-	-
	D	0.57	0.64	0.61	-	-
	E	0.61	0.65	0.60	0.61	-
	F	0.54	0.69	0.57	0.56	0.58
4	B	0.61	-	-	-	-
	C	0.58	0.60	-	-	-
	D	0.62	0.67	0.60	-	-
	E	0.61	0.63	0.57	0.62	-
	F	0.62	0.69	0.60	0.65	0.64
5	B	0.62	-	-	-	-
	C	0.52	0.55	-	-	-
	D	0.61	0.68	0.56	-	-
	E	0.67	0.65	0.58	0.64	-
	F	0.62	0.69	0.57	0.66	0.65
6	B	0.64	-	-	-	-
	C	0.54	0.56	-	-	-
	D	0.67	0.69	0.55	-	-
	E	0.58	0.61	0.52	0.62	-
	F	0.69	0.73	0.56	0.69	0.61
7	B	0.69	-	-	-	-
	C	0.66	0.64	-	-	-
	D	0.62	0.68	0.60	-	-
	E	0.63	0.67	0.62	0.62	-
	F	0.68	0.70	0.60	0.62	0.63
8	B	0.61	-	-	-	-
	C	0.65	0.58	-	-	-
	D	0.73	0.68	0.64	-	-
	E	0.61	0.64	0.61	0.67	-
	F	0.68	0.66	0.63	0.71	0.66
10	B	0.65	-	-	-	-
	C	0.64	0.63	-	-	-
	D	0.68	0.67	0.72	-	-
	E	0.61	0.61	0.62	0.65	-
	F	0.69	0.67	0.72	0.75	0.65

Table 10-10
Correlations among Language Arts Objectives

Grade	CS	B	D
4	D	0.50	-
	F	0.58	0.39
8	D	0.70	-
	F	0.61	0.52
10	D	0.73	-
	F	0.60	0.57

Table 10-11
Correlations among Social Studies Objectives

Grade	CS	A	B	C	D
4	B	0.55	-	-	-
	C	0.52	0.53	-	-
	D	0.52	0.52	0.49	-
	E	0.56	0.58	0.54	0.54
8	B	0.65	-	-	-
	C	0.57	0.62	-	-
	D	0.61	0.62	0.54	-
	E	0.52	0.59	0.51	0.50
10	B	0.62	-	-	-
	C	0.66	0.64	-	-
	D	0.63	0.57	0.66	-
	E	0.61	0.60	0.65	0.60

Table 10-12
Correlations among Science Objectives

Grade	CS	A/B	C	D	E	F
4	C	0.63	-	-	-	-
	D	0.43	0.43	-	-	-
	E	0.50	0.49	0.40	-	-
	F	0.52	0.52	0.42	0.48	-
	G/H	0.59	0.60	0.45	0.52	0.55
8	C	0.59	-	-	-	-
	D	0.53	0.52	-	-	-
	E	0.51	0.47	0.49	-	-
	F	0.58	0.56	0.52	0.51	-
	G/H	0.62	0.62	0.53	0.51	0.59
10	C	0.69	-	-	-	-
	D	0.57	0.57	-	-	-
	E	0.60	0.58	0.51	-	-
	F	0.62	0.59	0.52	0.51	-
	G/H	0.70	0.69	0.57	0.58	0.62

Table 10-13
Principal Components Analysis

Content Area	Grade	First Eigenvalue	Second Eigenvalue	Ratio of First Two Eigenvalues
Reading	3	13.61	1.77	7.71
	4	11.60	1.50	7.73
	5	11.09	1.55	7.16
	6	11.04	1.85	5.96
	7	10.14	1.63	6.24
	8	9.61	1.59	6.04
	10	10.54	1.46	7.22
Mathematics	3	10.71	1.64	6.53
	4	10.33	1.49	6.94
	5	10.55	1.70	6.20
	6	10.87	1.78	6.12
	7	11.12	1.51	7.36
	8	11.47	1.67	6.86
	10	11.49	1.56	7.38
Language Arts	4	5.05	1.33	3.80
	8	6.73	1.25	5.40
	10	6.56	1.16	5.68
Social Studies	4	6.37	1.36	4.69
	8	7.77	1.58	4.93
	10	9.15	1.66	5.52
Science	4	7.09	1.24	5.70
	8	7.52	1.51	4.98
	10	9.63	1.19	8.08

Figure 7-1
SEM Curves, Reading Grades 3-6

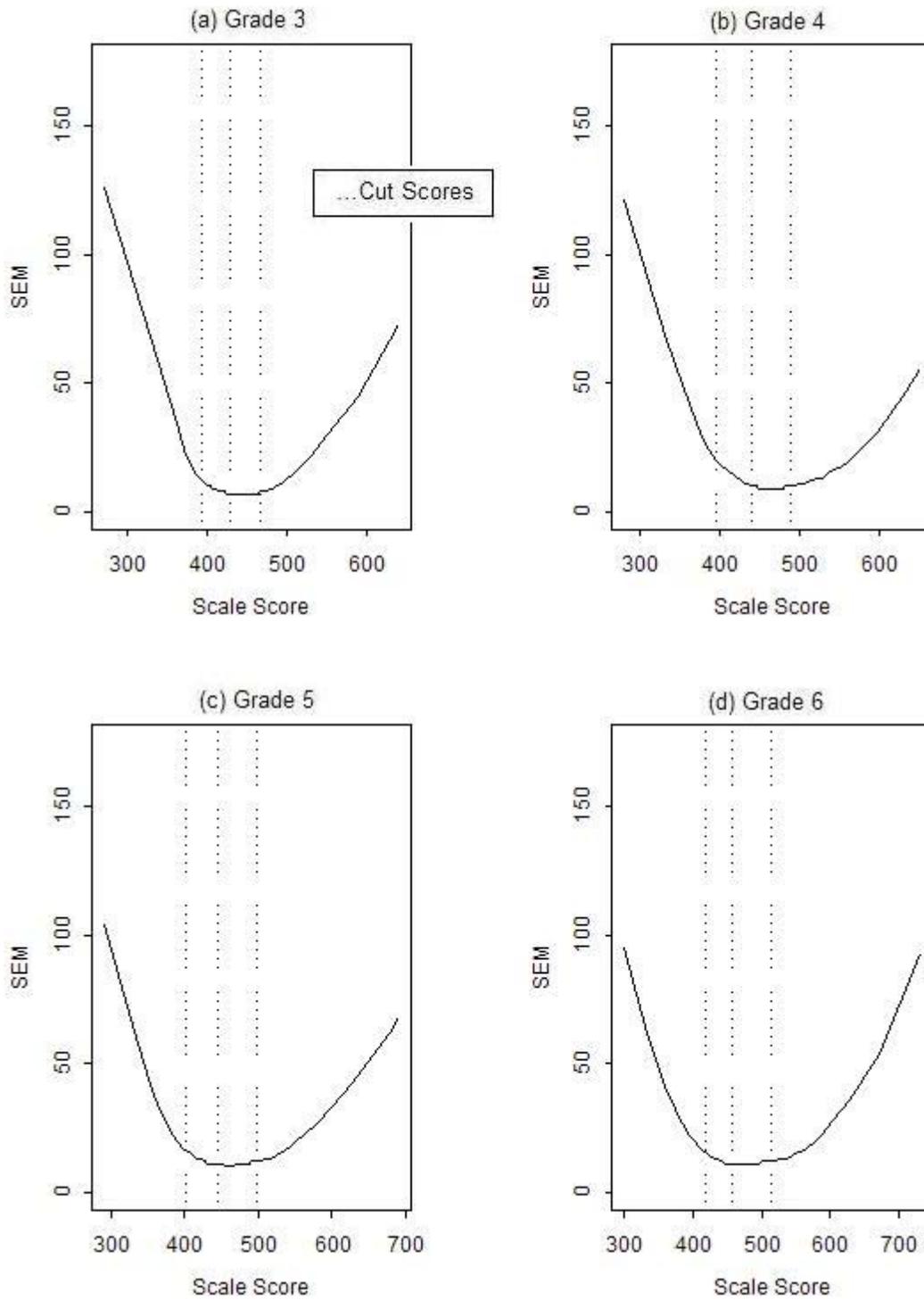


Figure 7-1 Cont'd
SEM Curves, Reading Grades 7, 8, 10

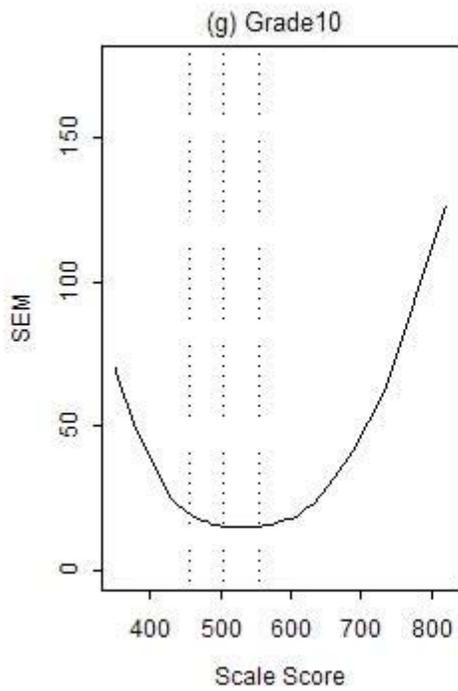
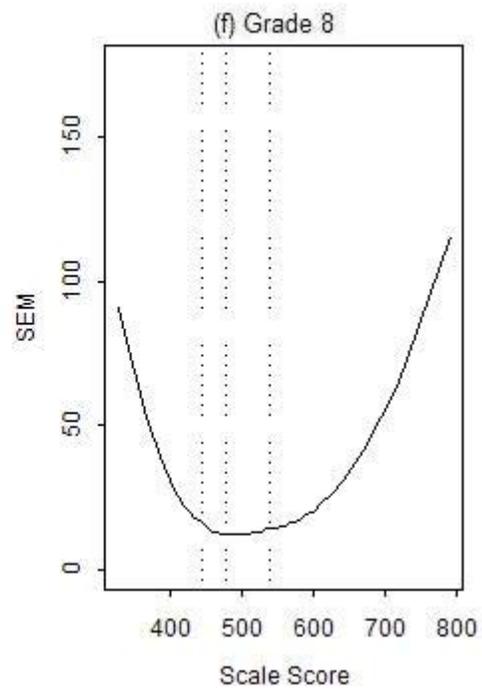
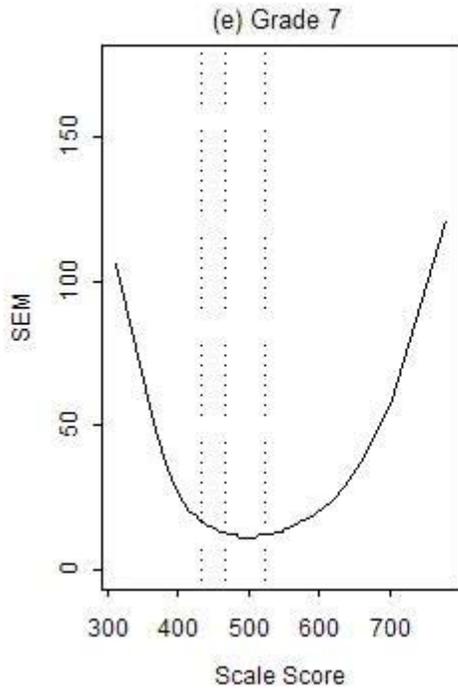


Figure 7-2
SEM Curves, Mathematics Grades 3-6

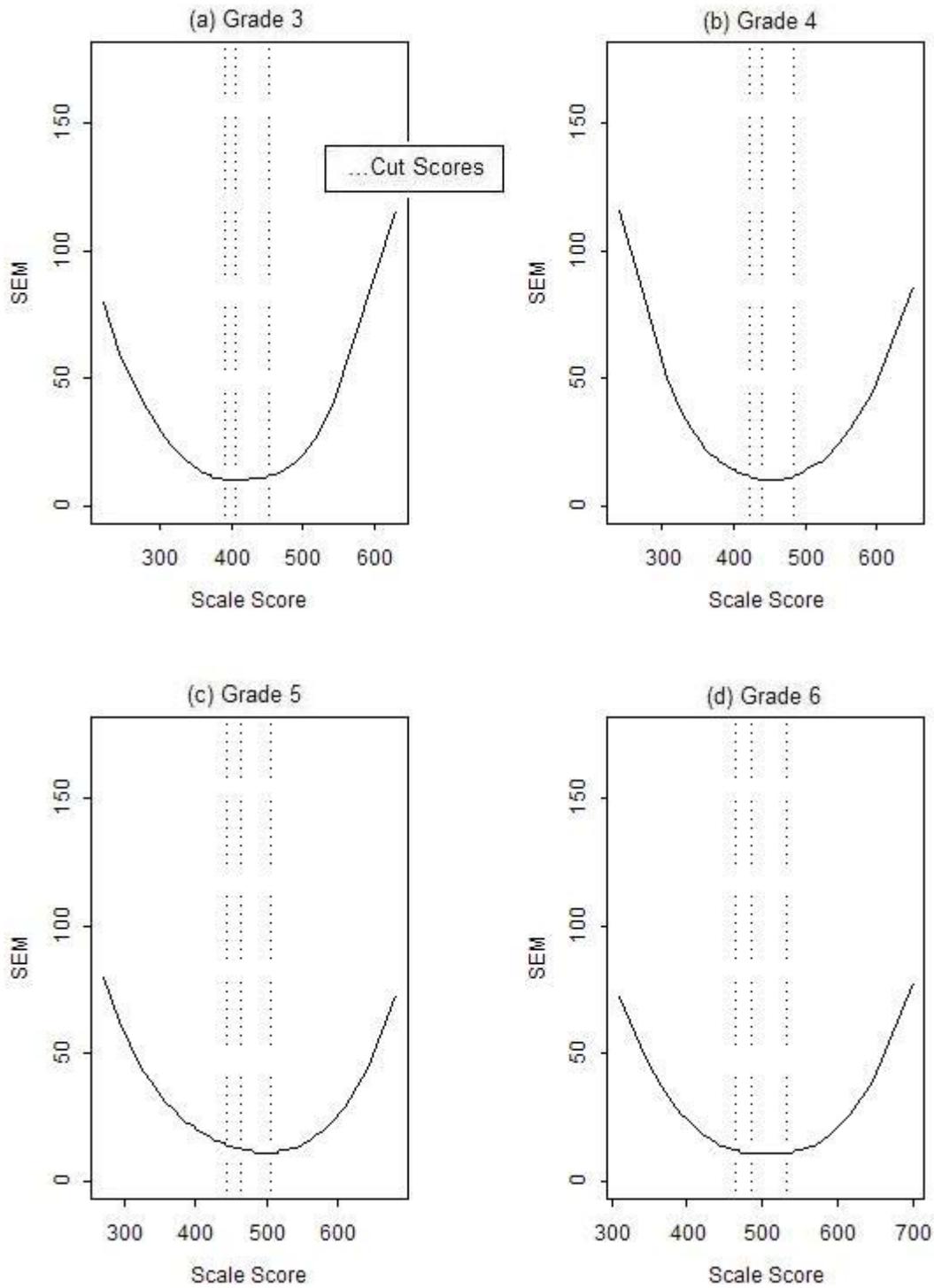


Figure 7-2 Cont'd
SEM Curves, Mathematics Grades 7, 8, 10

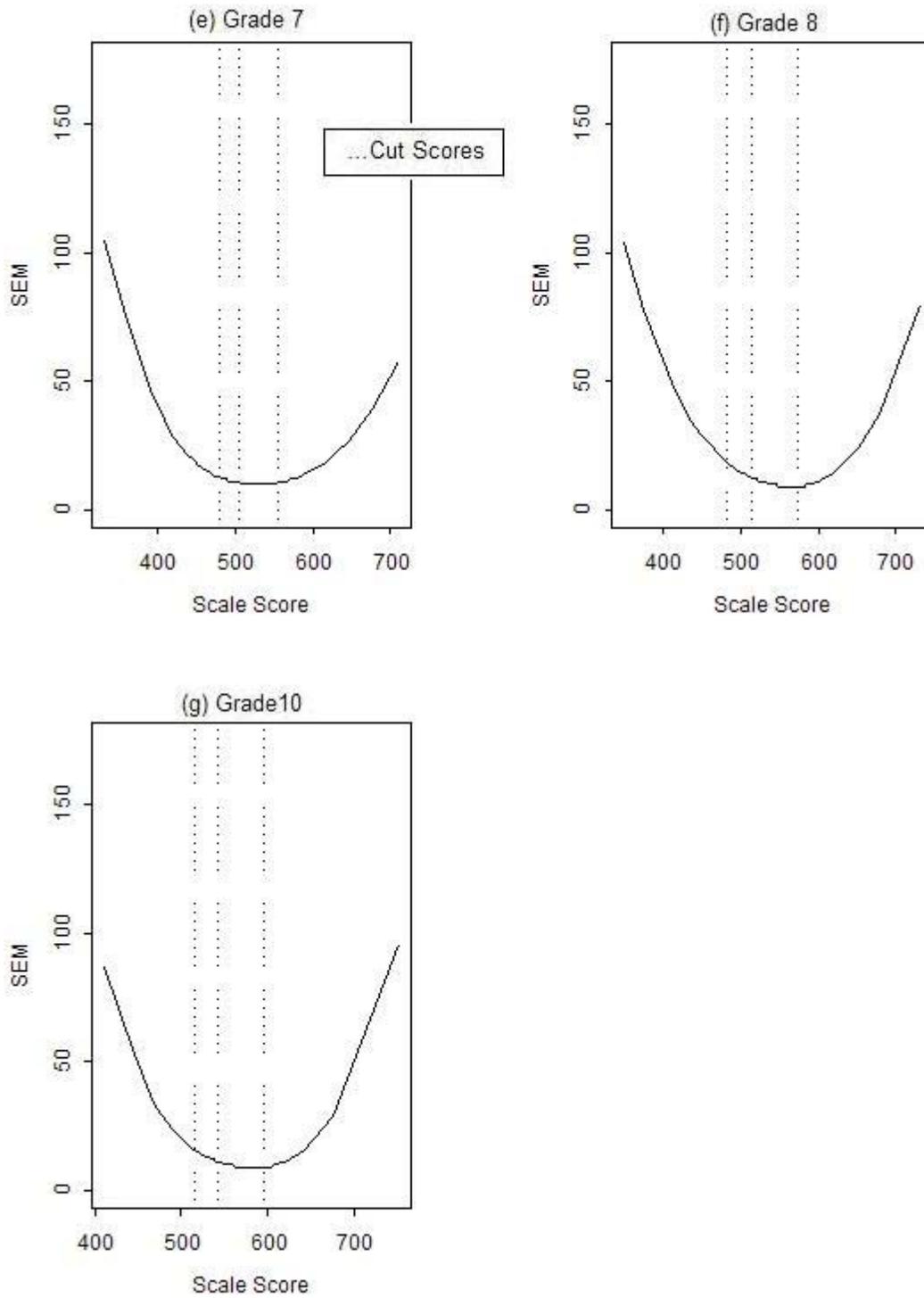


Figure 7-3
SEM Curves, Language Arts Grades 4, 8, 10

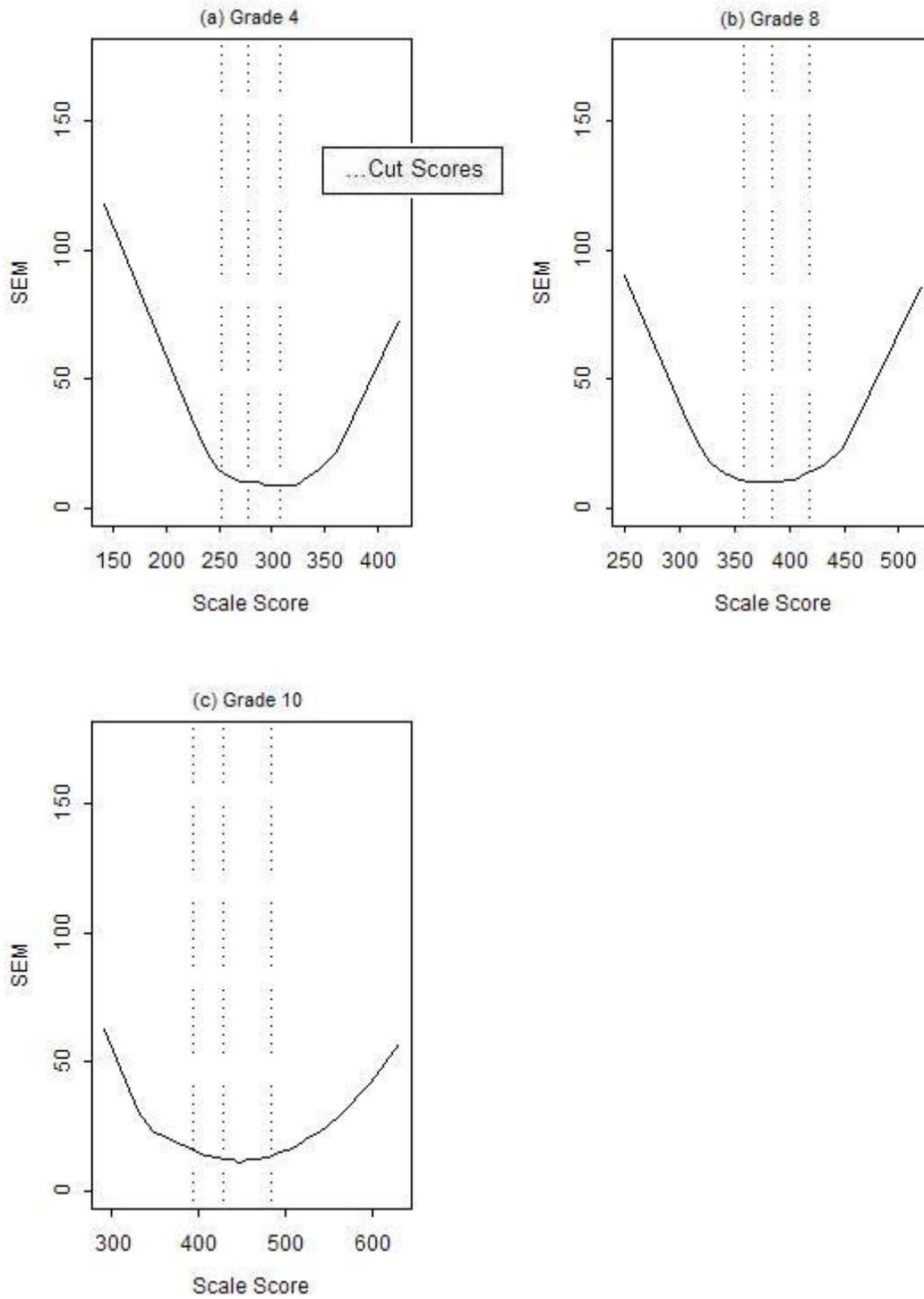


Figure 7-4
SEM Curves, Social Studies Grades 4, 8, 10

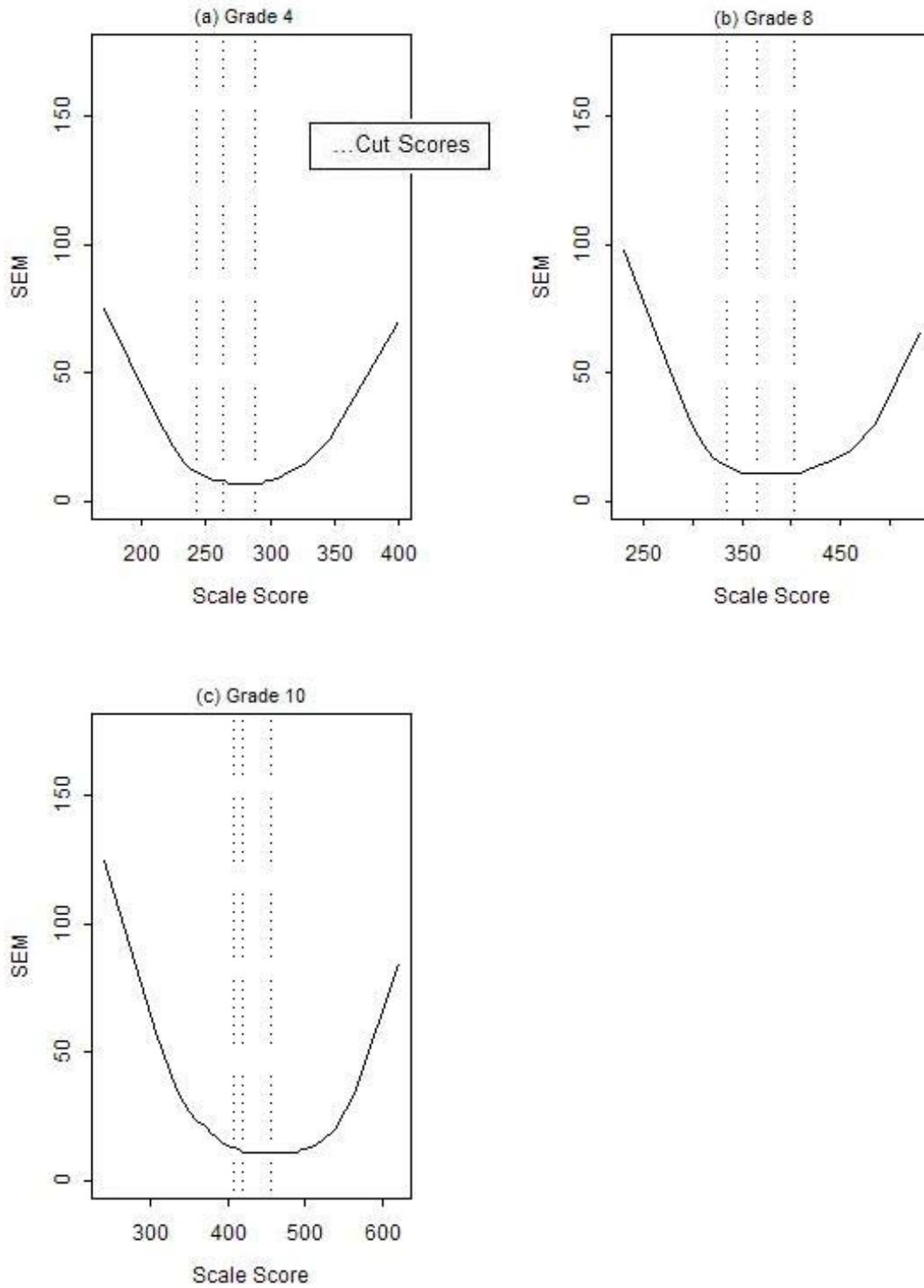


Figure 7-5
SEM Curves, Science Grades 4, 8, 10

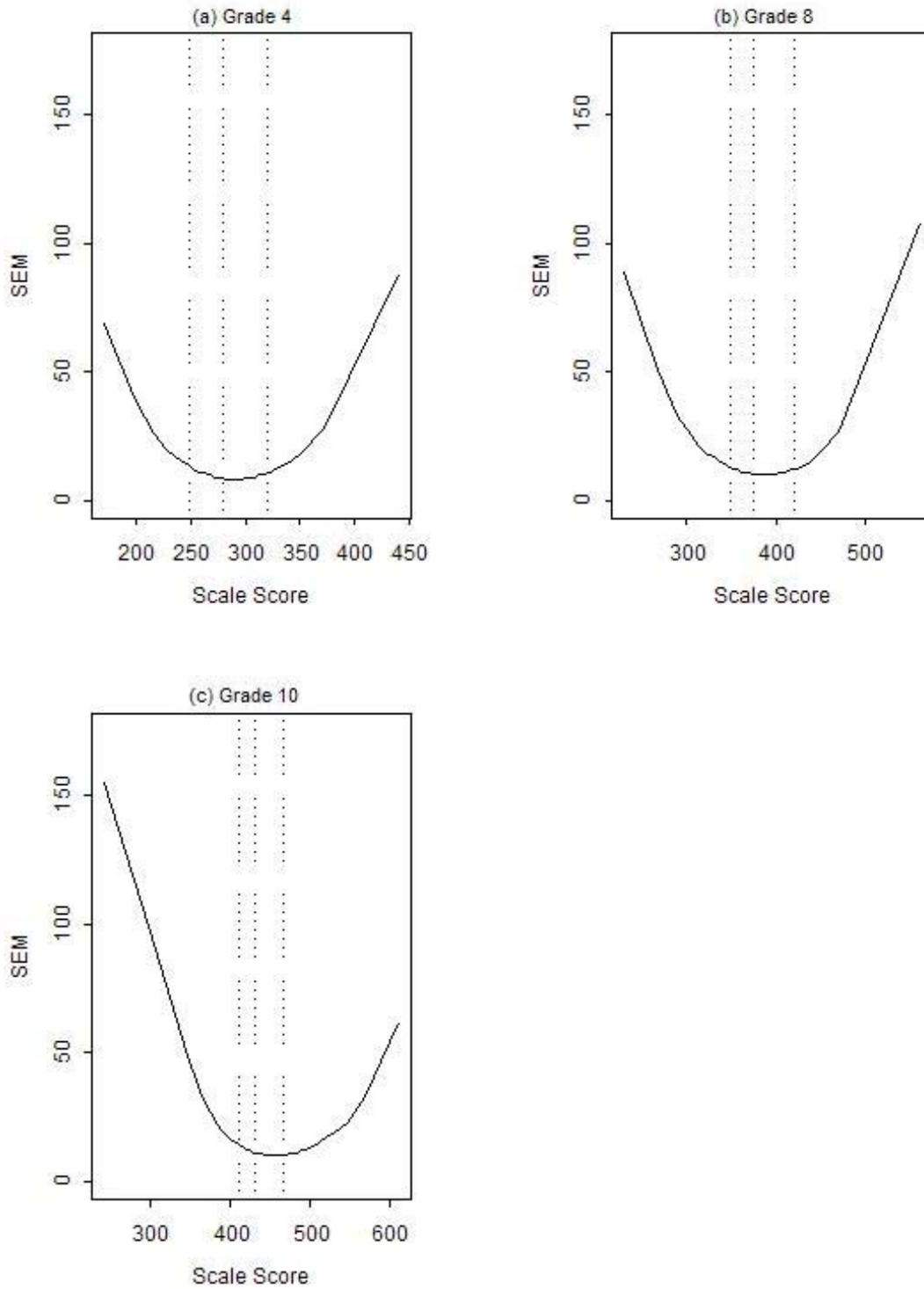


Figure 7-6
TCC Curve for Reading Grades 3-8, 10

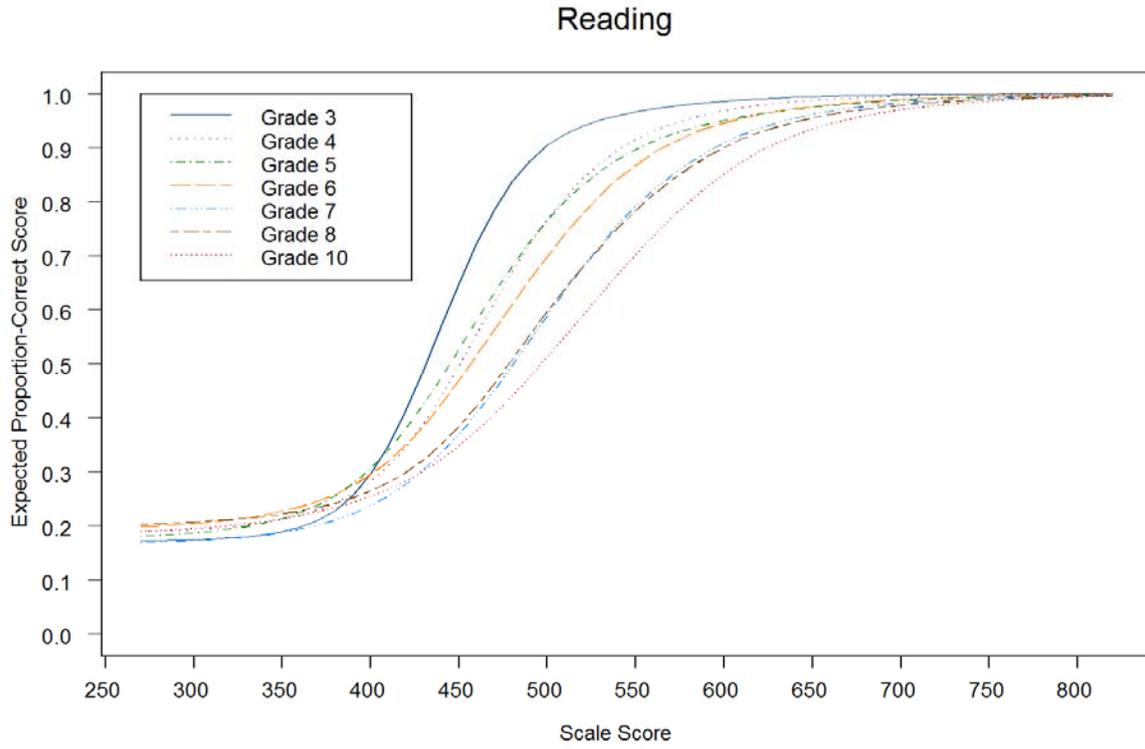


Figure 7-7
TCC Curve for Mathematics Grades 3-8, 10

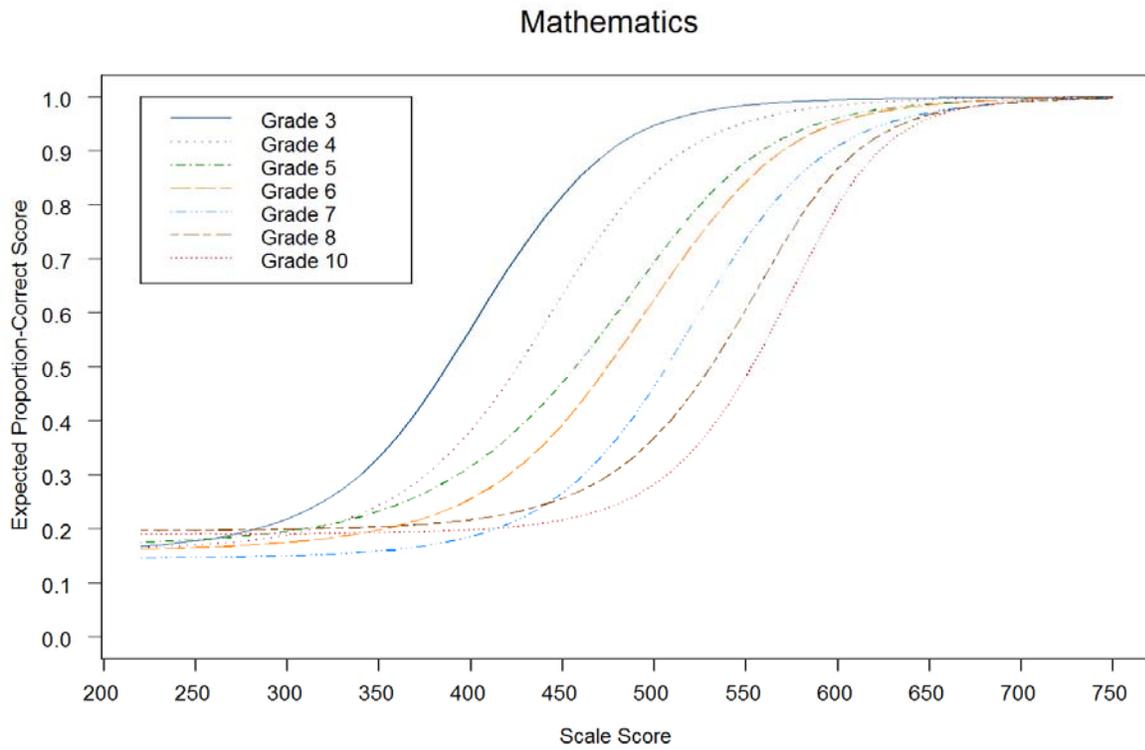


Figure 7-8
TCC Curve for Language Arts Grades 4, 8, 10

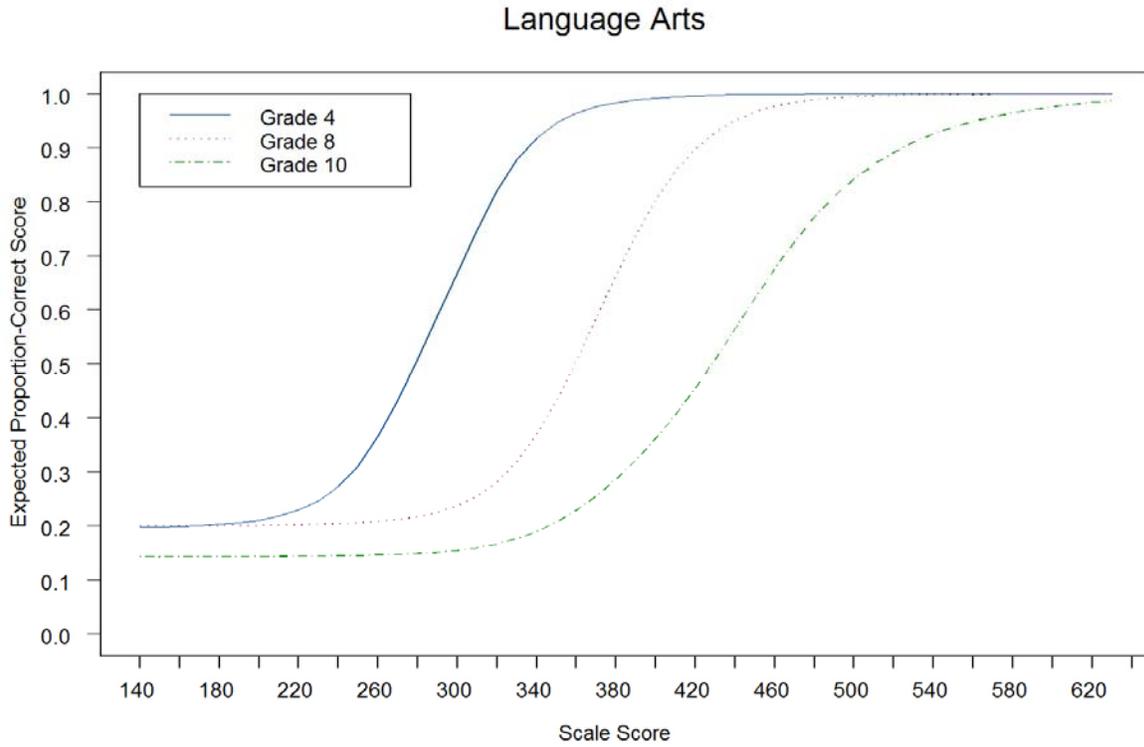


Figure 7-9
TCC Curve for Social Studies Grades 4, 8, 10

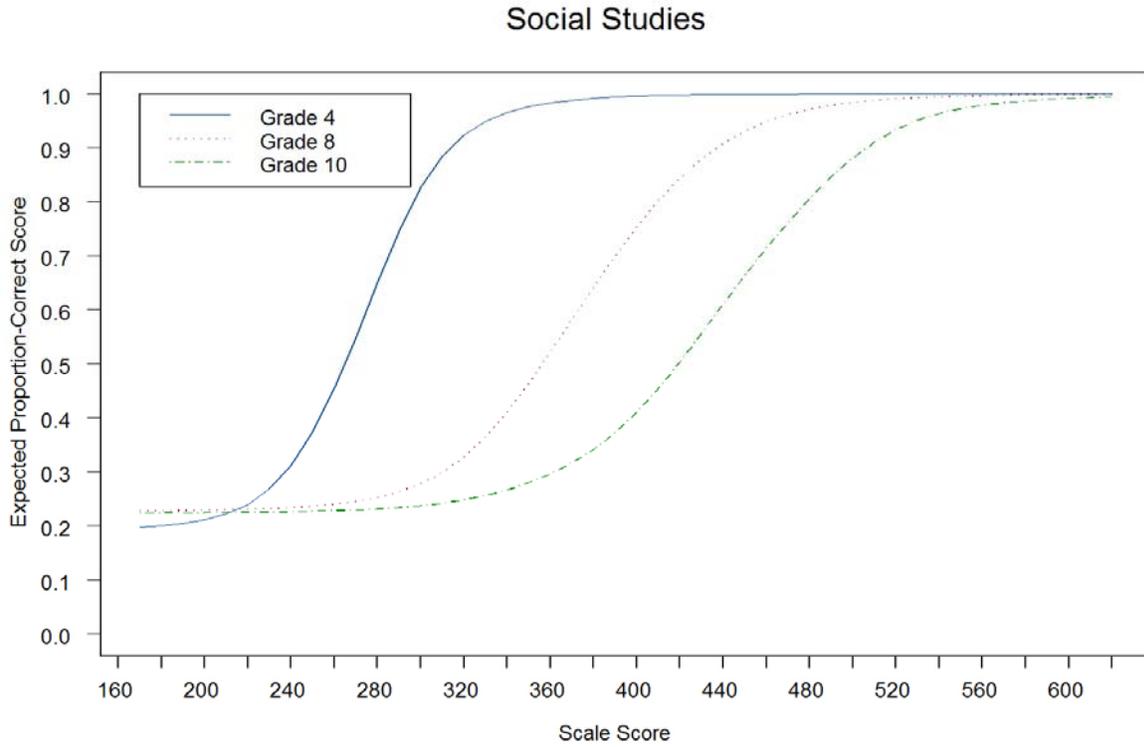


Figure 7-10
TCC Curve for Science Grades 4, 8, 10

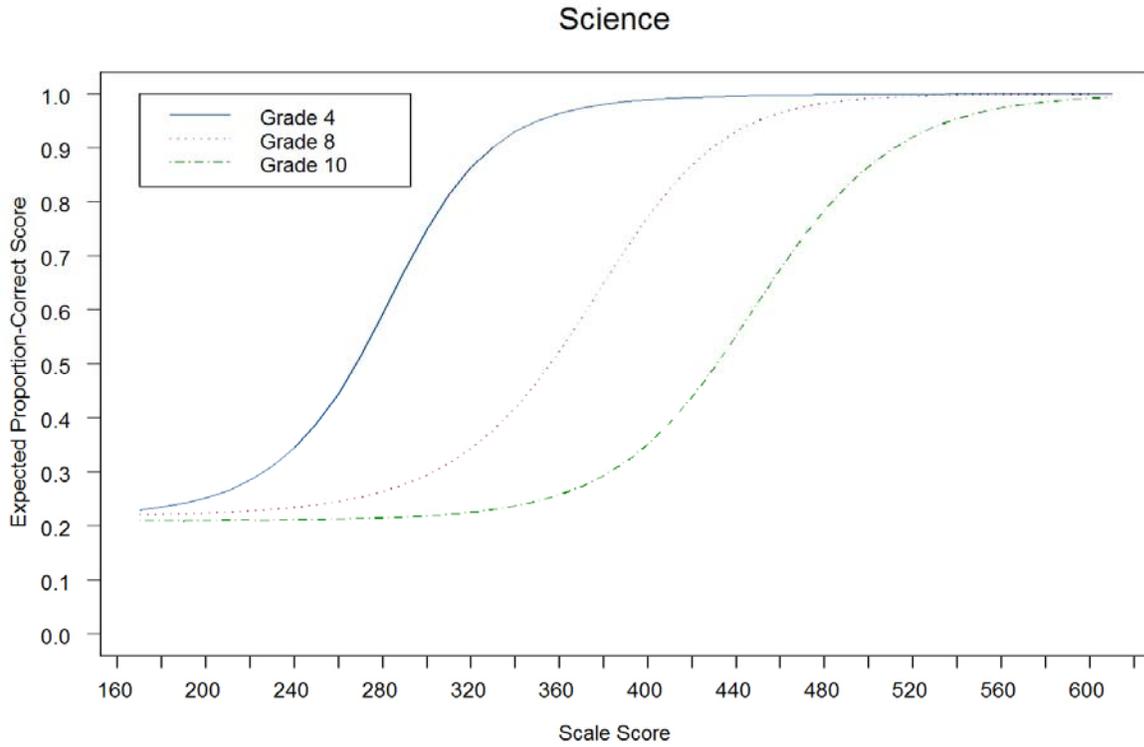


Figure 9-1
 Reading Indices for Classification Consistency and Classification Accuracy

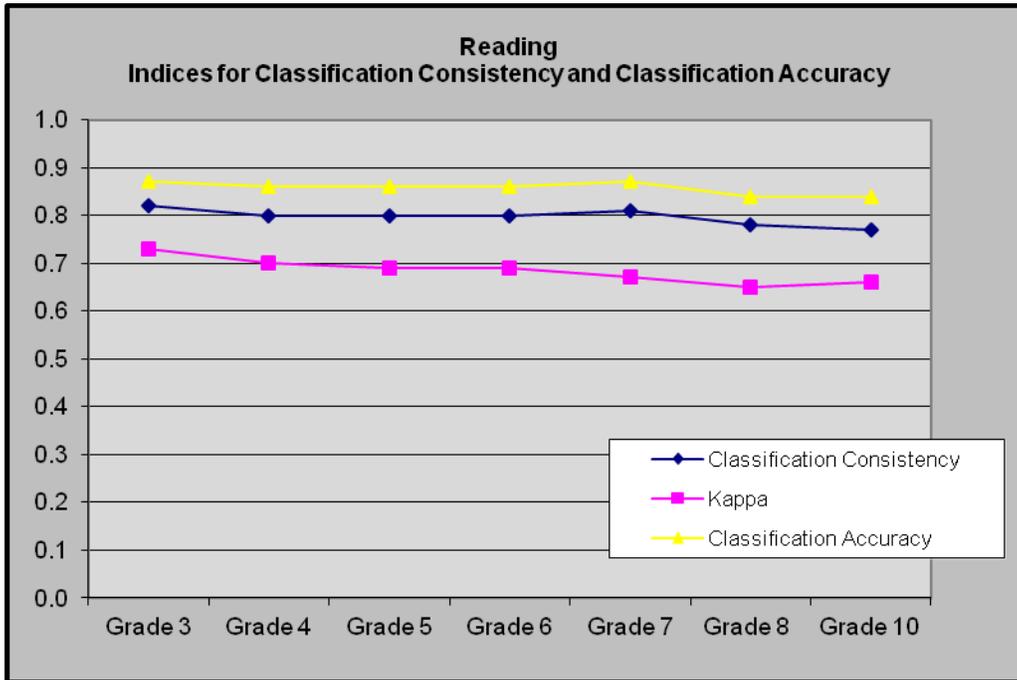


Figure 9-2
 Mathematics Indices for Classification Consistency and Classification Accuracy

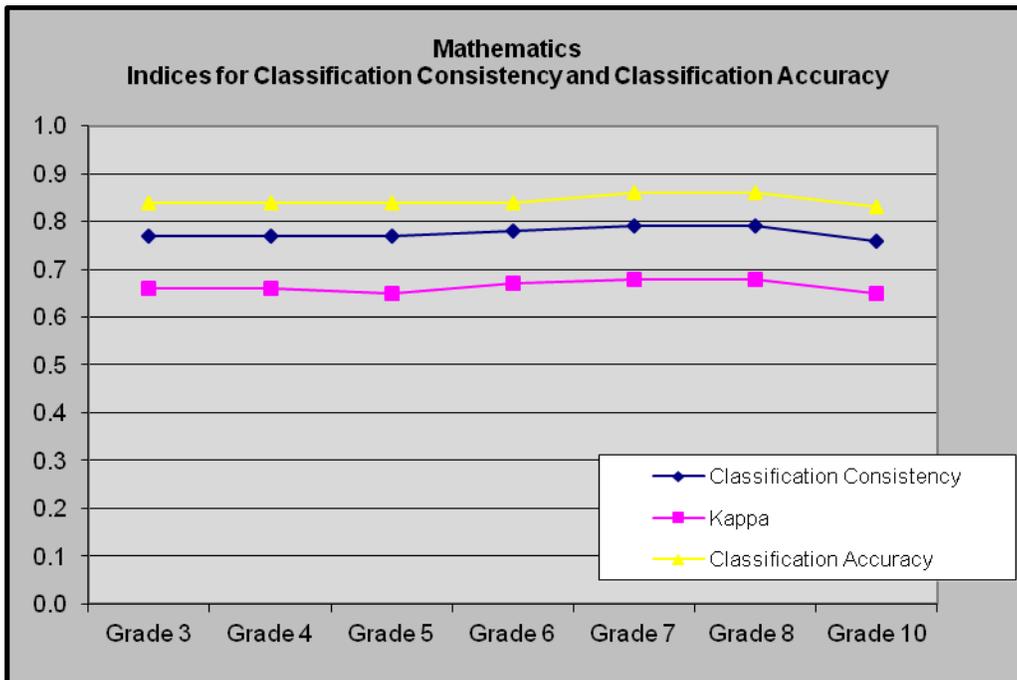


Figure 9-3
 Language Arts Indices for Classification Consistency and Classification Accuracy

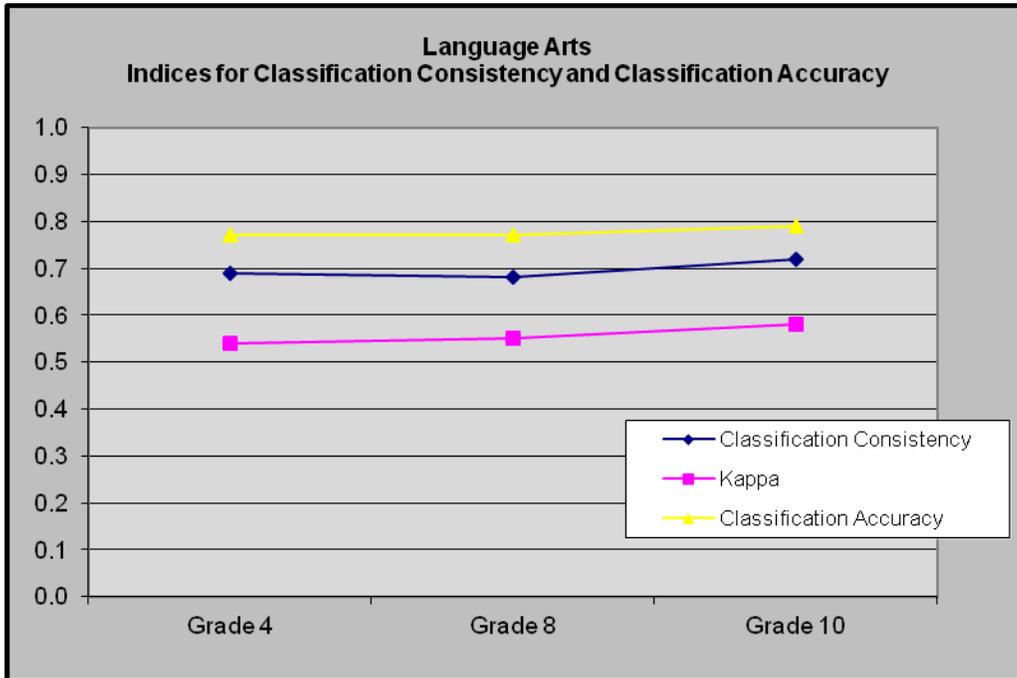


Figure 9-4
 Social Studies Indices for Classification Consistency and Classification Accuracy

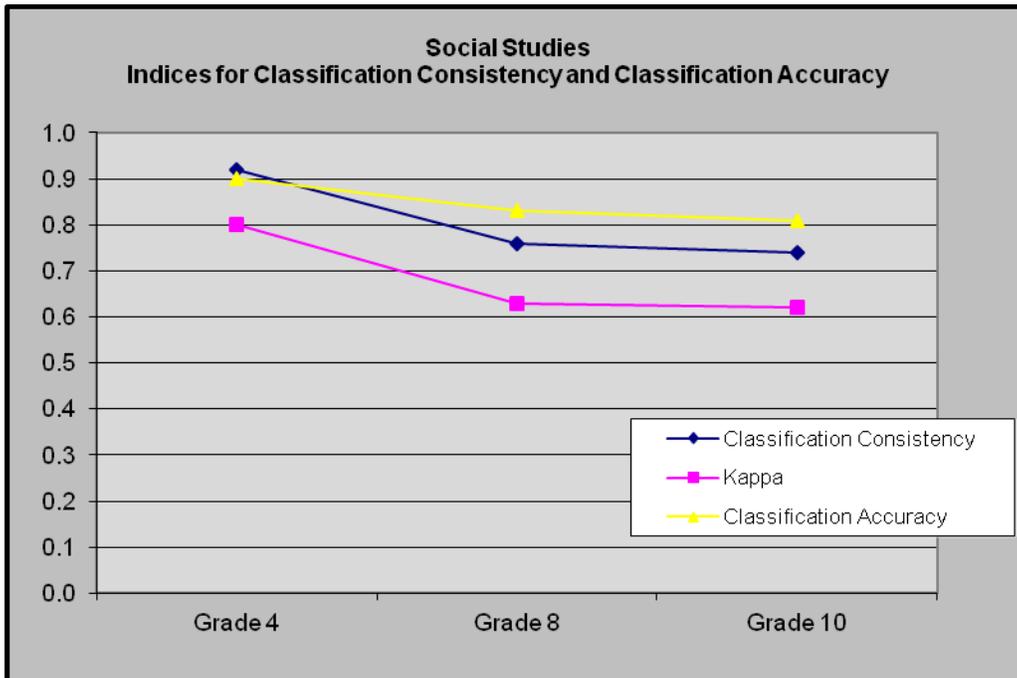
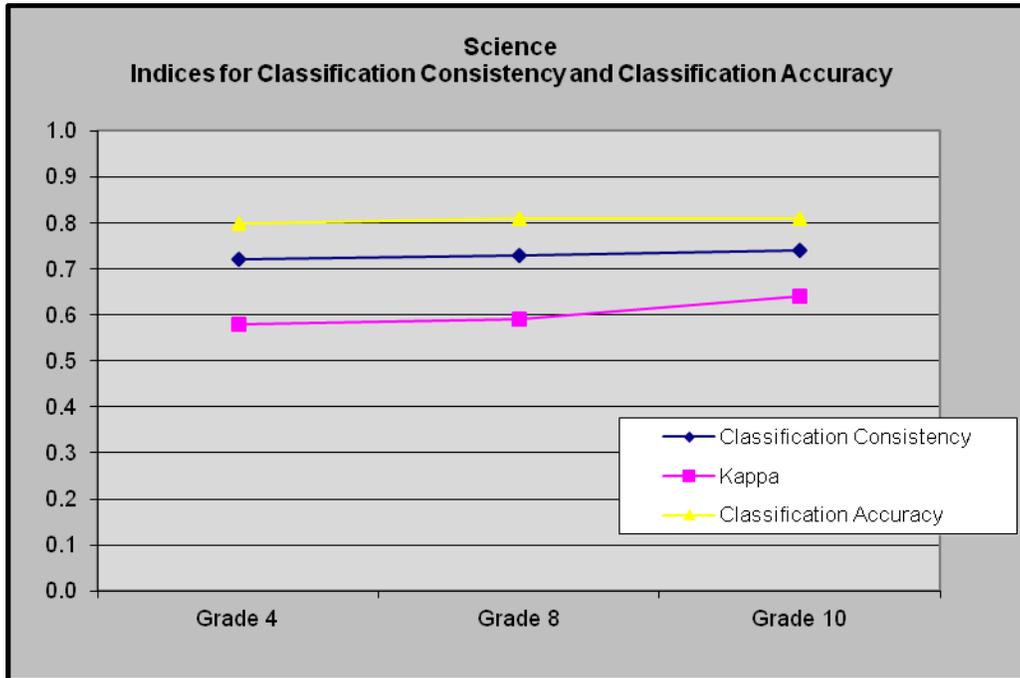


Figure 9-5
Science Indices for Classification Consistency and Classification Accuracy



Appendix 1: Fall 2011 Item Selection Check-Off Form

Fall 2011 Item Selection Check-Off Form

Program Name:	Wisconsin Knowledge and Concepts Examinations (WKCE)
Administration Year:	Fall 2011
Content Area:	
Grade Level:	

	Fall 2010				Anchor Items: Fall 2011				Total Form: Fall 2011			
	No. Items	% No. Items	No. Points	% No. Points	No. Items	% No. Items	No. Points	% No. Points	No. Items	% No. Items	No. Points	% No. Points
SR												
CR												
Prompt												
Total												

Blueprint Comparison (Number of items)

Reporting Category	Fall 2010 Blueprint Requirement			Fall 2011 Blueprint Requirement			Fall 2010 Actual Content Distribution			Fall 2010 Anchors			Fall 2011 Anchors			Fall 2011 Complete Form		
	SR	CR	Prompt	SR	CR	Prompt	SR	CR	Prompt	SR	CR	Prompt	SR	CR	Prompt	SR	CR	Prompt
A																		
B																		
C																		
D																		
E																		
F																		
G																		
Total	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Blueprint Comparison (% of items)

Reporting Category	Fall 2010 Blueprint Requirement			Fall 2011 Blueprint Requirement			Fall 2010 Actual Content Distribution			Fall 2010 Anchors			Fall 2011 Anchors			Fall 2011 Complete Form		
	SR	CR	Prompt	SR	CR	Prompt	SR	CR	Prompt	SR	CR	Prompt	SR	CR	Prompt	SR	CR	Prompt
A																		
B																		
C																		
D																		
E																		
F																		
Total																		

Fall 2011 Form Distribution of Items by DOK & Objective (number of items)

Objective	Obj DOK	DOK Level 1	DOK Level 2	DOK Level 3	DOK Level 4	50% ≥ Obj DOK?	Comments
A							
B							
C							
D							
E							
F							
G							

*Combine SR & CR items

Answer Key Distribution

	A	B	C	D
Selected Items				
Session 1				
Session 2				
Session 3				
Session 4				
Session 5				
Total Test				

- The "Selected Items" entry should be the same as the sum of the 5 sessions on the total test.

Number of Items on DPI Watch List

	Anchor Items	Full Form	Item PEID IDs	Reasons for Use of Watch Items
Number of items				

Number of easy and difficult items for preventing ceiling and floor effect

	Previous Year's Form			Current Year's Anchors			Current Year's Full Form		
	SR	CR	ER	SR	CR	ER	SR	CR	ER
Mean P-value									
No. of items: P < .30									
No. of items: .30 < P < .40									
No. of items: .80 < P < .90									
No. of items: P > .90									

Number of items flagged for point biserials (Pbis) indicating poor discrimination

	Fall 2010 Form			Fall 2011 Anchors			Fall 2011 Full Form		
	SR	CR	ER	SR	CR	ER	SR	CR	ER
No. of items: Pbis < .15									
No. of items: Pbis for distracter > 0									
No. of items: Pbis for correct choice is negative									
PEID ID of Flagged Items in Current Form:									
Reasons for Using Flagged Items in Current Form:									

Number of items near the Proficient Cut Score

	Fall 2010 Form		Anchors 2011 (SR only)		Fall 2011 Form	
	SR	CR	SR	CR	SR	CR
Proficient cut score = _____						
No. of items +/- 8 points around cut score						

TCCs overlay each other closely?

	Fall 2010 Form and Fall 2011 Anchors	Fall 2010 Form and Fall 2011 Form	Fall 2010 Anchor and Fall 2011 Form
TCCs of Selected Form			

SE curves are smoothly bow-shaped without dips, bumps, and twists?

	Fall 2010 Form and Fall 2011 Anchors	Fall 2010 Form and Fall 2011 Form	Fall 2010 Anchor and Fall 2011 Form
SE curves of Selected Form			

Expected % Max. RS Difference between any two Selected Forms ≤ 0.05 :

	Fall 2010 Form and Fall 2011 Anchors	Fall 2010 Form and Fall 2011 Form	Fall 2010 Anchor and Fall 2011 Form
Max Raw Score Difference			

Number of Items with DIF

Group	Statistic	Fall 2010 Form		Fall 2011 Anchors		Fall 2011 Full Form		PEID ID of Items Flagged using one or more DIF method
		Against	Favor	Against	Favor	Against	Favor	
Gender	Female (Linn-Harnisch)							
	Male (Linn-Harnisch)							
	Mantel-Haenszel							
Ethnicity	White (Linn-Harnisch)							
	African American (Linn-Harnisch)							
	African American (Mantel-Haenszel)							
	Hispanic (Linn-Harnisch)							
	Hispanic (Mantel-Haenszel)							
	Asian (Linn-Harnisch)							
	Asian (Mantel-Haenszel)							
	American Indian (Linn-Harnisch)							
	American Indian (Mantel-Haenszel)							

Number of Items with DIF cont'd

ELL	Proficient (Linn-Harnisch)							
	Not Proficient (Linn-Harnisch)							
	Mantel-Haenszel							
SES	Disadvantaged (Linn-Harnisch)							
	Not Disadvantaged (Linn-Harnisch)							
	Mantel-Haenszel							
Disability	Disabled (Linn-Harnisch)							
	Not Disabled (Linn-Harnisch)							
	Mantel-Haenszel							

Number of Items with Less Than Optimal Fit

	Fall 2010 Form	Fall 2011 Anchor	Fall 2011 Full Form	PEID ID of Items Flagged using one or more DIF method	Reasons for Using Flagged Items
Fit					

Items Dropped in Fall 2010 Test to Not Use in Future Tests

Grade	Subject	Item#	PEID-Item	Form

Approvals: Two independent reviews and approvals within Development are required prior to submitting to Research

	Name	Digital Signature	Date
Assessment Editor I			
Assessment Editor II			
Project Mgr/Development Lead			
Research Scientist			
WDPI			

Appendix 2: 2012 WKCE Assessment Accommodations Matrix

THE ASSESSMENT ACCOMMODATIONS MATRIX FOR STUDENTS WITH DISABILITIES - UPDATED 2012

Accommodations for Students with Disabilities

on the *Wisconsin Knowledge and Concepts Examination (WKCE)* and *Wisconsin Alternate Assessment for Students with Disabilities (WAA-SwD)*

- All accommodations for a student with a disability must be documented on an IEP or Section 504 plan in the section for statewide assessment.* Refer to page 3
- All *Allowable Test Practices for All Students* must be documented in an IEP or Section 504 plan in the section for statewide assessment.
- Accommodations should be consistent with day-to-day instructional methods and should not be first introduced during testing.
- Accommodations should enhance access without changing the skill or construct measured.
- Districts should monitor the use of accommodations by comparing assessment accommodations received with those stated in IEP or Section 504 plans.

Accommodation Description For Students with Disabilities (D)		WKCE	WAA-SwD
Test Directions			
D 1	Sign language for directions. ^{1,11}	✓	✓
D 2	Mark or highlight directions. ^{1,2,3}	✓	N/A: Test administrator reads WAA-SwD aloud.
D 3	 Provide printed copy of teacher directions (i.e. bold text following the SAY icon) from the WKCE Test Administration Manual. ¹	✓	N/A: Test administrator reads WAA-SwD aloud.
D 4	Explain or clarify directions. ¹	✓	✓
D 5	Student rereads and/or restates directions. ¹	✓	✓
Content Presentation			
D 6	Turn pages for student.	✓	✓
D 7	Braille; student responses must be transcribed into scorable test book by a licensed teacher of the visually impaired or a certified transcriber. ^{6,14}	✓	✓
D 8	DPI-provided WAA-SwD Picture Descriptions; appropriate only for a student who cannot access the printed WAA-SwD, even with magnification, or the Braille WAA-SwD. ¹³	N/A	✓
D 9	Large-print; student responses must be transcribed into scorable test book. ^{6,14}	✓	N/A: WAA-SwD is 18 pt. font, no separate large print edition.
D 10	Extra test book; answers must be recorded in one scorable test book. ¹⁴	✓	N/A: All items are presented to the student so that they view one entire item at a time.
D 11	Sign language for test passages and questions (Not allowed on Reading tests). ¹¹	✓	✓
D 12	Text talker for test passages and questions (Not allowed on Reading tests). ⁴	✓	N/A: Test administrator reads WAA-SwD aloud.
D 13	Student reads aloud to self.	✓	✓
D 14	Test administrator reads test passages and questions aloud (Not allowed on WKCE Reading test or WAA-SwD “Read-by-Student” items). ⁹	✓	N/A: Test administrator reads WAA-SwD aloud.
D 15	Student records him/herself reading aloud and plays back recording. ⁴	✓	✓

THE ASSESSMENT ACCOMMODATIONS MATRIX FOR STUDENTS WITH DISABILITIES - UPDATED 2012



Accommodation Description for <i>Students with Disabilities</i> (D)		WKCE	WAA-SwD
Content Presentation (cont.)			
D 16	Audio recording of test passages and questions in English (Not allowed on WKCE Reading test or WAA-SwD). ^{4,9}	✓	N/A: Test administrator reads WAA-SwD aloud.
D 17	Read the Reading test ONLY in the following scenarios as described in <i>Form I-7-B</i> : ^{8,9} a) For a student who is blind or visually impaired who is not yet proficient in contracted Braille, the WKCE Reading test passages and questions may be read aloud. b) For a student who is blind or visually impaired who is not yet proficient in un-contracted Braille, the WAA-SwD “Read-by-Student” Reading test items may be read aloud.	✓	✓
Response			
D 18	Manipulatives, base-ten blocks, 3-D shapes, 100’s chart (not multiplication table), whole integer number lines, number boards, etc. are allowed as long as they do not provide a definition or description.	✓	✓ Follow guidelines in WAA-SwD Manipulatives Guide. http://dpi.wi.gov/oea/pdf/maniguide.pdf
D 19	Calculator and/or multiplication table (Not allowed on sections of the Mathematics test measuring computation skills -refer to each appropriate grade’s Test Administrator’s Manual at http://dpi.wi.gov/oea/publications.html).	✓	N/A: A calculator is not allowed on the WAA-SwD.
D 20	Braille output device; transcribe student responses into scorable test book. ^{4, 6, 14}	✓	✓
D 21	Student indicates responses orally to scribe. ⁵	✓	N/A: Test administrator records all student responses.
D 22	Student signs responses to interpreter/scribe. For the Writing test, translation from American Sign Language (ASL) is not allowed; student must use English-based sign. ^{5, 11}	✓	✓
D 23	Student records responses using an audio or video device: a) Test administrator transcribes student’s responses into scorable test book. ^{6, 14} b) Student watches or listens to his/her recorded responses and transcribes into scorable test book. ^{4, 6, 14}	✓	N/A: Student is allowed to communicate responses in whichever mode is best for the student. Test administrator records student responses.
D 24	Computer or word processor; responses must be transcribed into the scorable test book. For the Language Arts and Writing tests, all spell- and grammar-checking devices must be turned off; for the Mathematics test, the calculator function must be turned off for non-calculator sessions. ^{4, 6, 14}	✓	N/A: Student is allowed to communicate responses in whichever mode is best for the student. Test administrator records student responses.
D 25	 Speech-to-text devices; responses must be transcribed into the scorable test book. For the Mathematics test, the calculator function must be turned off for non-calculator sessions (Not allowed on Language Arts or Writing tests). ^{4, 6, 14}	✓	N/A
D 26	Provide spelling assistance or a spell-check device, where appropriate (Not allowed on Language Arts or Writing tests).	✓	N/A: Student is not required to spell responses.
Setting			

THE ASSESSMENT ACCOMMODATIONS MATRIX FOR STUDENTS WITH DISABILITIES - UPDATED 2012

D 27	Student moves, stands, or paces during individual administration.	✓	✓
Timing/Scheduling			
D 28	Extra time; test session must be completed within the same day the student started the session. ⁷	✓	✓

Other Accommodations for Students with Disabilities			
D 29	Any accommodation not on this list must be submitted to DPI for approval, as it may represent a modification which changes the skill being measured. <ul style="list-style-type: none"> o All requests for an additional accommodation must be made to DPI at least two weeks before the test administration window begins, by completing and submitting the Request for Accommodation Form located at http://dpi.wi.gov/oea/accommtrx.html. o Requests will be reviewed by a committee to determine whether the request can be approved; approval or non-approval will be returned via fax or email. 		

*Allowable Accommodations for Students in Unique Circumstances

Some students who do *not* have an IEP or 504 plan, due to unique circumstances at the time of testing, may be able to demonstrate their learning more accurately through the use of accommodations on an **as needed basis only**. In these unique cases, please follow the guidelines outlined in the matrix for Students with Disabilities; call DPI's Office of Educational Accountability with any questions at (608) 267-1072. Examples of unique circumstances:

- o A student with a broken arm may need a scribe or be able to use a word processor to record responses.
- o A student who forgot to wear eyeglasses may need a visual magnification device.

Explanation of Footnotes - Only footnotes 1-9, 11, 13 and 14 apply to students with disabilities.

1 Test directions:

- Any portion of the WKCE test book where the word “Directions” appears in a shaded/colored box, typically at the top of a page preceding a particular section of test content. In addition, test directions refer to anything that the test administrator reads aloud to the class from the WKCE Test Administration Manual (i.e. bold text following the SAY icon).
- WKCE item stems and test questions should not be considered directions.
- Test Directions for the WAA-SwD are incorporated into the teacher test book and are read aloud to the student. These directions must be read verbatim but may be reread if a student needs further clarification.
- Directions may not be expanded.

2 Marking test book with #2 pencil: Student should not make pencil marks near answer bubbles, other than to mark one correct answer. Student should not mark in any of the following areas in the test book:

- the student Pre-ID Barcode on barcode label,
- the timing tracks (the parallel lines along the side of the test book),
- the skunk lines (the little squares and rectangles across the bottom of each page of the test book), or
- the Litho codes (the squares and numbers across the bottom of the first and last page of the test book).

3 Highlighters:

- Carefully supervise the use of highlighters as they may cause smudging of pencil marks and bubbles and, therefore, could affect scoring.
- Do not allow the highlighting of track marks, litho codes, skunk lines, barcodes, pre-slugged bubbles or any carbon black printing. The highlighters cause these black inks to blur and bleed, which could affect scoring.
- Use only a highlighter from the following list, which were tested and found to have minimal problems:
 - Avery Hi-liter (regular or thin-tipped), Bic Brite-Liner, Sanford Major Accent, or Sanford Pocket Accent (thin-tipped)

4 Using audio/video or electronic (e.g., word processor or text talker) recordings: when using accommodations that involve audio, video or electronic recordings or saved files, the test administrator must ensure that the recording or file is deleted upon completion of testing for security purposes.

5 Use of a scribe (student dictates orally to scribe):

- A scribe may be provided when a student’s documented disability, ELL status, or injury prevents them from writing their answer.
- When a student dictates responses orally to a scribe, the test must be administered in a separate, individual setting so as not to disturb other students.
- The WKCE Writing prompts measure composition, grammar, punctuation, capitalization, and spelling; therefore, a student must dictate these exactly as they are to be written.
- A scribe must be impartial and should allow the student adequate time to review and approve the response, if desired.
- All scribing should be done with a #2 pencil; responses scribed in ink will not be scored.

6 Transcribing student responses (student’s answers are documented in a manner other than in the scorable test book [e.g., large-print, Braille version, computer response, etc]):

- A translator who scribes student responses from native language to English should translate word-for-word to the extent possible for all content areas except Writing. For the Writing test, student must dictate or write responses in English (translation not allowed) exactly as they are to be written.
- The answers must be transcribed into the regular WKCE test book or WAA-SwD student Answer Document with a #2 pencil to be scored.
- Transcription of the student’s responses must be verbatim, including spelling, formatting, punctuation, etc.
- Test security must be maintained. After answers are transcribed, destroy all electronically-saved student responses, including audio tapes. All paper copies of student work (e.g., Braille tests, large-print tests, graph/lined/grid paper, printed copies of computer responses, etc.) must be returned with non-scorable test materials.

7 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, a student requiring the use of restroom facilities should be escorted by either a test administrator or other school staff. In addition, a student must not be allowed to use any form of wireless communication during these breaks.

⁸ **Student who is blind or visually impaired and is not proficient in Braille** may have the Reading portion of the WKCE and the “Read by Student” Reading items of the WAA-SwD read aloud by a test administrator.

- The WKCE is available in contracted Braille; if a student designated by his/her IEP Team, by use of *Form I-7-B* (available at <http://dpi.wi.gov/oea/dacforms.html>), to take the WKCE is not proficient in contracted Braille and is receiving instruction in reading contracted Braille, the student may have the Reading test passages and items read by a test administrator.
- The WAA-SwD is available in uncontracted Braille; if a student designated by his/her IEP Team, by use of *Form I-7-B*, to take the WAA-SwD is not proficient in uncontracted Braille, the student may have the “Read by Student” items in the Reading test read by a test administrator.

⁹ **Test Administrator Read Aloud Accommodation (not allowed on Reading test except for students qualifying for accommodation D17):**

- Test administrator must read in a pace and tone that is appropriate for each individual student. Careful attention must be given such that no changes in tone or inflection are detectable which might indicate a correct answer.
- Students may direct test administrator to reread a portion of a passage, test question, or answer choice as needed.

¹⁰ **For students who have test items and/or directions translated into native language:**

- A qualified translator and interpreter (see http://dpi.wi.gov/oea/pdf/translator_guidelines.pdf) should have a Bachelor’s Degree in Modern Languages or a certification in interpretation or translation. When this is not possible, be sure that a translator or interpreter has the following qualifications:
 1. Mastery of the target language and dialect
 2. Familiarity with both cultures
 3. Extensive general and academic vocabulary in both languages
 4. Ability to express thoughts clearly and concisely in both languages
- *Translators* work with the written word, transferring meaning from a source language into a target language. *Interpreters* work with the spoken word, transferring meaning from a source language into a target language.
- Translators and interpreters should participate in all aspects of staff training related to test administration and test security.
- For more information about state provided scripts available in Spanish and bilingual word lists in Spanish and Hmong for the WKCE, please see <http://dpi.wi.gov/oea/ells.html>.
- In order for this accommodation to be most effective, a student should have content-area knowledge in their native language.

¹¹ **Sign Language and Oral Interpreters**

- An interpreter needs to be able to translate in the same method of sign language typically used by the student (e.g., American Sign Language [ASL] or English-based Sign Language). The interpreters must not clarify, elaborate, or provide assistance with the meaning of words, intent of test questions, or responses to test items.
 - E.g. The sign for many math symbols often defines for the student what the item is intending to measure and would therefore invalidate the item.

¹² **Simplified English:** The test administrator providing an accommodation in which English is simplified for words not related to content or vocabulary should be familiar with the content area being tested. The WAA-SwD is already in simplified language.

Example (Grade 5 WKCE Released Item) of a simplified English test item:

The sales receipt below shows the groceries that José purchased from the supermarket. What is the estimated cost of José’s groceries?

Simplified English: The receipt below shows the food that José bought from the store. Estimate how much money José spent on the food.

Note: It is important that “estimate” remain in this test item because it is part of the standard which is being tested.

¹³ **DPI-provided Picture Descriptions** are descriptions of the graphic found within an item. Picture descriptions are intended to replace, *not* supplement graphics for a student who is blind or is visually impaired who is not able to access the printed WAA-SwD, even with magnification, or the Braille WAA-SwD. Ordering information can be found at: <http://dpi.wi.gov/oea/dacforms.html>.

¹⁴ **Scorable Test Books** are the documents that are returned to the test vendor for scoring. For the WKCE, this is the test book itself. For the WAA-SwD, this is the student Answer Document. All student responses must be recorded on these documents in order to be scored.

Appendix 3: WKCE Glossary

Glossary

Abbreviations used in the WKCE Technical Report

2PPC: The two-parameter partial credit (2PPC) 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 (3PL) item response theory model. A mathematical model that shows the relationship between student achievement on a test and a single MC 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.

AYP: Adequate Yearly Progress. A state-defined criteria of educational accountability required as an outcome of the Federal NCLB law.

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. DIF is the degree to which an item performs differently for one group of examinees than it performs for another group of equally able examinees. DIF 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, and 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.

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.

FT: Field test item. A field test item is a newly developed item in a content area that is being administered to students for the first time. It does not contribute to student's score in a content area on WKCE.

HOSS: Highest obtainable scale score. The highest possible scale score on a test.

ICC: Item characteristic curve. ICCs show the mathematical probabilities of students of varying degrees of achievement answering an item correctly as well as the characteristics of the item (e.g., item difficulty, item discrimination, item guessing).

IRT: Item response theory. IRT is a mathematic model that shows the relationship between student achievement on a test and the performance on a test item.

LA: Language Arts. A content area in the WKCE.

LH: Linn-Harnisch. A DIF statistic that utilizes information provided by the three-parameter IRT model for multiple-choice and constructed-response items.

LOSS: Lowest obtainable scale score. The lowest possible scale score on a test.

MA: Mathematics. A content area in the WKCE.

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. The Mantel-Haenszel (MH χ^2_{MH}) statistic is a commonly used DIF statistic for multiple-choice items.

NCLB: No Child Left Behind Act. The name of Federal Public Law No. 107-110, 115 Stat. 1425

NCME: National Council on Measurement in Education. A professional organization centered in assessment, evaluation, testing, and educational measurement.

OP: Operational item. An operational item is one that has previously undergone field testing so it contributes to a student's score in a specific content area on the WKCE.

RD: Reading. A content area in the WKCE.

SC: Science. A content area in the WKCE.

SD: Standard deviation. The SD is a measure of the variability of observations from the mean.

SEM: Standard error of measurement. The SEM is an estimated average standard deviation of the observed score.

SES: Socioeconomic status. A student population subgroup category describing students as economically disadvantaged or not economically disadvantaged.

SMD: Standardized mean difference. SMD is commonly used DIF statistic for constructed-choice items.

SPI: Standardized performance indicator score. A subcontent area reporting score based on the items from a single content standard within given content area.

SS: Social Studies. A content area in the WKCE.

TCC: Test characteristic curve. TCCs show the mathematical relationship between students with varying degrees of achievement and their estimated overall test performance.

WKCE: Wisconsin Knowledge and Concepts Examinations. A criterion-referenced test designed to measure student achievement on the Wisconsin Model Academic Standards

WR: Writing. A content area in the WKCE.