The Badger Exam 3-8:
A Wisconsin Smarter Balanced Assessment

SPRING 2015

TECHNICAL MANUAL

Prepared by

Educational Testing Service

for

Office of Student Assessment
Wisconsin Department of Public Instruction

125 South Webster Street
P.O. Box 7841
Madison, WI 53707-7841

December 2015

Final

TABLE OF CONTENTS

LIST OF TABLES iv
LIST OF FIGURES viii
EXECUTIVE SUMMARY ix
CHAPTER 1: INTRODUCTION 1
1.1: Overview of This Technical Manual 2
1.2: Background of the Badger Exam 2
1.3: Appropriate Use of Test Scores, Performance Levels, and Claim Levels 3
1.4: Glossary of Abbreviations 4
CHAPTER 2: TEST DESIGN AND DEVELOPMENT 6
2.1: Smarter Balanced Content Framework 7
2.2: Test Components 7
2.3: Test Blueprint 8
2.4: Item and Form Development 9
CHAPTER 3: TEST ADMINISTRATION 19
3.1: Testing Schedule and Administration 20
3.2: Test Security and Confidentiality 20
3.3: Test Accommodations, Designated Supports, and Universal Tools 21
CHAPTER 4: SCORING AND REPORTING 32
4.1: Scoring Procedures 33
4.1.1: Types of Item Response 33
4.1.2: Machine Scoring 34
4.1.3: Hand Scoring 34
4.2: Reports to Be Produced and Data for Each Report 37
CHAPTER 5: SCORE CALIBRATION AND SCALING 42
5.1: Item Response Theory (IRT) based Calibration Procedure 43
5.1.1: Calibration Models 43
5.1.2: Test Taker Ability Estimation for Total and Claim Test Scores 45
5.2: Deriving Scale Scores for the Badger Exam 47
5.3: Score Verification Procedure 47

5.4: Samples Used for the Analyses 48
CHAPTER 6: TEST RESULTS 54
6.1: Classical Item Analyses 55
6.2: Interpreting Test Scores 57
6.2.1: Raw Score Results 57
6.2.2: Scale Score Results 58
6.2.3: Performance Level Results 60
6.2.4: Claim Level Results (Strength and Weakness) 61
CHAPTER 7: RELIABILITY 103
7.1: Overview of Reliability 104
7.2: Internal Consistency and Standard Error of Measurement 104
7.3: Interrater Agreement 107
7.4: Classification Consistency and Accuracy 108
CHAPTER 8: VALIDITY 148
8.1: Purpose of the Badger Exam 149
8.2: Evidence Based on Test Content 151
8.3: Evidence Based on Relations to Other Variables 152
8.4: Evidence Based on Response Processes 153
8.5: Evidence Based on Internal Structure 154
8.6: Evidence Based on Consequences of Testing 156
CHAPTER 9: QUALITY CONTROL PROCEDURES 161
9.1: Overview of Quality Control Process 162
9.2: Quality Control of Test Materials 162
9.3: Quality Control of System Functionality 162
9.4: Quality Control of Psychometric Analyses 163
9.5: Quality Control of Scoring and Reporting 164
REFERENCES 166
APPENDIX A: OPERATIONAL ITEM STATISTICS 168

LIST OF TABLES

Table 0.1: Summary of Internal Consistency ix
Table 0.2: Summary of Scaled Score Statistics x
Table 0.3: Percentage of Test Takers at each Performance Level x
Table 0.4: Classification Consistency and Accuracy at Proficient Cut Score x
Table 2.1: Claims Identified for ELA and Mathematics 11
Table 2.2: Test Blueprint for ELA Grade 3 12
Table 2.3: Test Blueprint for ELA Grade 4 12
Table 2.4: Test Blueprint for ELA Grade 5 13
Table 2.5: Test Blueprint for ELA Grade 6 13
Table 2.6: Test Blueprint for ELA Grade 7 14
Table 2.7: Test Blueprint for ELA Grade 8 14
Table 2.8: Test Blueprint for Mathematics Grade 3 15
Table 2.9: Test Blueprint for Mathematics Grade 4 15
Table 2.10: Test Blueprint for Mathematics Grade 5 16
Table 2.11: Test Blueprint for Mathematics Grade 6 16
Table 2.12: Test Blueprint for Mathematics Grade 7 17
Table 2.13: Test Blueprint for Mathematics Grade 8 17
Table 2.14: Summary of Assigned Classroom Activities in Mathematics Tests 18
Table 3.1: Definitions for Three Levels of Test Security Incidents 23
Table 3.2: Definitions for Accommodations, Designated Supports, and Universal Tools 23
Table 3.3: Test Accommodations 24
Table 3.4: Designated Supports 27
Table 3.5: Universal Tools 31
Table 4.1: Score Distributions of Human-Scored ELA Items 39
Table 4.2: Score Distributions of Human-Scored Mathematics Items 40
Table 5.1: Summary of Pre-Equating IRT Item Parameter Estimates 49
Table 5.2: Summary of Vertically Scaled Test Taker Proficiency Estimates 49
Table 5.3: Scaling Constants for the Reporting Metric 50
Table 5.4: Summary of Sampled Data 50

Table 6.1: Summary of ELA Classical Item Statistics 63
Table 6.2: Summary of Mathematics Classical Item Statistics..... 64
Table 6.3: Summary of Flagged items by Classical Item Statistics..... 65
Table 6.4: Raw Score Descriptive Statistics by Total..... 66
Table 6.5: Raw Score Descriptive Statistics by Gender 67
Table 6.6: Raw Score Descriptive Statistics for ELA by Race/Ethnicity..... 68
Table 6.7: Raw Score Descriptive Statistics for Mathematics by Race/Ethnicity..... 70
Table 6.8: Raw Score Descriptive Statistics by Socioeconomic Status..... 74
Table 6.9: Raw Score Descriptive Statistics by Disability 75
Table 6.10: Raw Score Descriptive Statistics by English Language Proficiency..... 76
Table 6.11: Scale Score Descriptive Statistics by Total 77
Table 6.12: The Number and Percentage of Test Takers Earning the Lowest and Highest Possible Scale Scores 78
Table 6.13: Scale Score Descriptive Statistics by Gender..... 79
Table 6.14: Scale Score Descriptive Statistics for ELA by Race/Ethnicity 80
Table 6.15: Scale Score Descriptive Statistics for Mathematics by Race/Ethnicity..... 82
Table 6.16: Scale Score Descriptive Statistics by Socioeconomic Status 84
Table 6.17: Scale Score Descriptive Statistics by Disability..... 85
Table 6.18: Scale Score Descriptive Statistics by English Language Proficiency 86
Table 6.19: Percentage of Test Takers in Performance Level, ELA Grade 3 87
Table 6.20: Percentage of Test Takers in Performance Level, ELA Grade 4 88
Table 6.21: Percentage of Test Takers in Performance Level, ELA Grade 5 89
Table 6.22: Percentage of Test Takers in Performance Level, ELA Grade 6 90
Table 6.23: Percentage of Test Takers in Performance Level, ELA Grade 7 91
Table 6.24: Percentage of Test Takers in Performance Level, ELA Grade 8 92
Table 6.25: Percentage of Test Takers in Performance Level, Mathematics Grade 3..... 93
Table 6.26: Percentage of Test Takers in Performance Level, Mathematics Grade 4..... 94
Table 6.27: Percentage of Test Takers in Performance Level, Mathematics Grade 5..... 95
Table 6.28: Percentage of Test Takers in Performance Level, Mathematics Grade 6..... 96
Table 6.29: Percentage of Test Takers in Performance Level, Mathematics Grade 7..... 97
Table 6.30: Percentage of Test Takers in Performance Level, Mathematics Grade 8..... 98

Table 6.31: Percentage of Test Takers in Claim Level, ELA 99
Table 6.32: Percentage of Test Takers in Claim Level, Mathematics 100
Table 7.1: Test Reliability Estimates for Total Group and Subgroups, ELA Grade 3 111
Table 7.2: Test Reliability Estimates for Total Group and Subgroups, ELA Grade 4 112
Table 7.3: Test Reliability Estimates for Total Group and Subgroups, ELA Grade 5 113
Table 7.4: Test Reliability Estimates for Total Group and Subgroups, ELA Grade 6 114
Table 7.5: Test Reliability Estimates for Total Group and Subgroups, ELA Grade 7 115
Table 7.6: Test Reliability Estimates for Total Group and Subgroups, ELA Grade 8 116
Table 7.7: Test Reliability Estimates for Total Group and Subgroups, Mathematics Grade 3 .. 117
Table 7.8: Test Reliability Estimates for Total Group and Subgroups, Mathematics Grade 4 .. 119
Table 7.9: Test Reliability Estimates for Total Group and Subgroups, Mathematics Grade 5 .. 121
Table 7.10: Test Reliability Estimates for Total Group and Subgroups, Mathematics Grade 6 123
Table 7.11: Test Reliability Estimates for Total Group and Subgroups, Mathematics Grade 7 125
Table 7.12: Test Reliability Estimates for Total Group and Subgroups, Mathematics Grade 8 127
Table 7.13: Test and Claim Level Reliability Estimates for ELA 129
Table 7.14: Test and Claim Level Reliability Estimates for Mathematics Grades 3-5 130
Table 7.15: Test and Claim Level Reliability Estimates for Mathematics Grades 6-8 131
Table 7.16: Interrater Agreement for ELA Human-Scored Item 132
Table 7.17: Interrater Agreement for Mathematics Human-Scored Item 133
Table 7.18: Classification Consistency and Accuracy for ELA Grade 3 135
Table 7.19: Classification Consistency and Accuracy for ELA Grade 4 135
Table 7.20: Classification Consistency and Accuracy for ELA Grade 5 136
Table 7.21: Classification Consistency and Accuracy for ELA Grade 6 136
Table 7.22: Classification Consistency and Accuracy for ELA Grade 7 137
Table 7.23: Classification Consistency and Accuracy for ELA Grade 8 137
Table 7.24: Classification Consistency and Accuracy for Mathematics Grade 3..... 138
Table 7.25: Classification Consistency and Accuracy for Mathematics Grade 4..... 139
Table 7.26: Classification Consistency and Accuracy for Mathematics Grade 5..... 140
Table 7.27: Classification Consistency and Accuracy for Mathematics Grade 6..... 141
Table 7.28: Classification Consistency and Accuracy for Mathematics Grade 7..... 142
Table 7.29: Classification Consistency and Accuracy for Mathematics Grade 8..... 143

Table 8.1: Claim Score Intercorrelations between ELA and Mathematics, Grade 3 157
Table 8.2: Claim Score Intercorrelations between ELA and Mathematics, Grade 4 157
Table 8.3: Claim Score Intercorrelations between ELA and Mathematics, Grade 5 157
Table 8.4: Claim Score Intercorrelations between ELA and Mathematics, Grade 6 158
Table 8.5: Claim Score Intercorrelations between ELA and Mathematics, Grade 7 158
Table 8.6: Claim Score Intercorrelations between ELA and Mathematics, Grade 8 158
Table 8.7: Confirmatory Factor Analysis for ELA 159
Table 8.8: Confirmatory Factor Analysis for Mathematics 160
Table A.1: Operational Item Statistics, ELA Grade 3 169
Table A.2: Operational Item Statistics, ELA Grade 4 171
Table A.3: Operational Item Statistics, ELA Grade 5 173
Table A.4: Operational Item Statistics, ELA Grade 6 175
Table A.5: Operational Item Statistics, ELA Grade 7 177
Table A.6: Operational Item Statistics, ELA Grade 8 179
Table A.7: Operational Item Statistics, Mathematics Grade 3 181
Table A.8: Operational Item Statistics, Mathematics Grade 4 183
Table A.9: Operational Item Statistics, Mathematics Grade 5 185
Table A.10: Operational Item Statistics, Mathematics Grade 6 187
Table A.11: Operational Item Statistics, Mathematics Grade 7 189
Table A.12: Operational Item Statistics, Mathematics Grade 8 191

LIST OF FIGURES

Figure 2.1: Test Components of the Badger Exam 18
Figure 5.1: Test Characteristic Curves, ELA 51
Figure 5.2: Test Characteristic Curves, Mathematics 51
Figure 5.3: Test Information Functions, ELA 52
Figure 5.4: Test Information Functions, Mathematics 52
Figure 5.5: Test Taker Proficiency Box Plots, ELA 53
Figure 5.6: Test Taker Proficiency Box Plots, Mathematics 53
Figure 6.1: Score Range and Percentage of Test Takers at each Performance Level, ELA 101
Figure 6.2: Score Range and Percentage of Test Takers at each Performance Level, Mathematics 102
Figure 7.1: Conditional Standard Error of Measurement Curves of Scale Score, ELA 144
Figure 7.2: Conditional Standard Error of Measurement Curves of Scale Score, Mathematics 146

EXECUTIVE SUMMARY

This technical manual documents the development, delivery, analyses, and results[1] of the spring 2015 Badger Exam: A Wisconsin Smarter Balanced Assessment (Badger Exam). The Badger Exam is a computer-based summative assessment administered in English language arts (ELA) and Mathematics in grades 3 through 8 during the last seven weeks of the school year.

The Badger Exam included selected response items, constructed response items, and technology enhanced items in ELA and Mathematics. Mathematics also included performance tasks which require test takers to demonstrate analytical skills and real-world problem solving. The tests were not timed. Typically, the ELA tests required about 1 1/2 hours; the Mathematics tests required 3 to 3 1/2 hours. The actual testing time varied based on test takers' ability levels.

The design of the Badger Exam followed the extensively reviewed content specifications that targeted the claims and depth of knowledge levels for each grade and subject established by the Smarter Balanced Assessment Consortium. Smarter Balanced is a group of states working together to develop next-generation assessments that accurately measure test taker progress toward college and career readiness. In general, the Badger Exam met high technical standards in terms of reliability and validity. The ELA tests had internal consistency ranging from 0.87 to 0.91, and the Mathematics tests had internal consistency ranging from 0.89 to 0.91. Table 0.1 presents a summary of the internal consistency for the Badger Exam.

Table 0.1: Summary of Internal Consistency

Grade	Internal Consistency		
	ELA	Mathematics	
		With PT1	With PT2
3	0.90	0.90	0.91
4	0.88	0.91	0.91
5	0.91	0.91	0.91
6	0.87	0.91	0.91
7	0.88	0.91	0.91
8	0.87	0.89	0.89

Note. The majority of the Mathematics tests at each grade level are identical with the exception of the performance task (PT). There are two performance tasks at each grade level (PT1 and PT2) that are spirally assigned in each grade.

Table 0.2 provides the summary of scale scores in each subject and grade. Overall, the mean scale score increased with grade level in both ELA and Mathematics because items/tests were vertically scaled across the grades during the 2014 Smarter Balanced field test standard-setting analysis. Table 0.3 provides the percentages of test takers who achieved performance levels at or above proficient (i.e., proficient and advanced). Overall, the percentage of test takers classified as proficient or advanced varied somewhat across grade levels, from 47% to 55% in ELA and from 40% to 52% in Mathematics.

[1] Unless otherwise noted, the analysis results and summaries about test performance are derived from the most recently available statewide test taker data file. Inclusion and exclusion rules to aggregate the data for purposes of these analyses may not necessarily coincide with the rules applied to produce operationally published score reports.

Table 0.2: Summary of Scale Score Statistics

Grade	ELA				Mathematics			
	Mean	*SD*	Min	Max	Mean	*SD*	Min	Max
3	2433.4	82.9	2114	2623	2436.4	73.8	2189	2621
4	2468.3	90.5	2131	2663	2475.0	80.3	2204	2659
5	2511.3	90.5	2201	2701	2504.0	89.8	2219	2700
6	2519.4	96.0	2210	2724	2525.3	98.2	2235	2748
7	2553.3	98.2	2258	2745	2544.2	105.0	2250	2778
8	2569.8	96.7	2288	2769	2559.9	106.4	2265	2802

Table 0.3: Percentage of Test Takers at each Performance Level

Grade	ELA		Mathematics	
	Below Basic + Basic	Proficient + Advanced	Below Basic + Basic	Proficient + Advanced
3	47.9	52.1	48.2	51.8
4	49.1	50.9	52.3	47.7
5	45.0	55.0	59.6	40.4
6	53.3	46.7	58.7	41.3
7	47.5	52.5	56.2	43.8
8	45.5	54.5	59.7	40.3

Classification consistency and accuracy analyses for the overall tests at the proficient cut score showed that the inferences made about whether the test takers meet the standards from the test scores are valid and defensible. Table 0.4 provides a summary of classification consistency and accuracy for the Badger Exam. From the classification of the test takers as proficient, the classification consistency ranged from 0.84 to 0.88 and the classification accuracy ranged from 0.88 to 0.90.

Table 0.4: Classification Consistency and Accuracy at Proficient Cut Score

Grade	Classification Consistency			Classification Accuracy		
	ELA	Mathematics		ELA	Mathematics	
		With PT1	With PT2		With PT1	With PT2
3	0.86	0.86	0.87	0.90	0.89	0.90
4	0.85	0.85	0.85	0.89	0.89	0.89
5	0.87	0.88	0.87	0.90	0.90	0.90
6	0.85	0.87	0.86	0.89	0.90	0.90
7	0.85	0.87	0.86	0.89	0.90	0.90
8	0.84	0.85	0.85	0.88	0.89	0.88

Note. The majority of the Mathematics tests at each grade level are identical with the exception of the performance task (PT). There are two performance tasks at each grade level (PT1 and PT2) that are spirally assigned in each grade.

Validity evidence to support the interpretation and use of the test scores was collected throughout the testing cycle of the Badger Exam. Empirical results concerning the relationships between scores on ELA and Mathematics tests provided evidence of validity because these relationships were found to be consistent with the definition of the constructs that the tests are intended to measure (See Chapter 8). The correlations of claim scores within and among Badger Exams were used for this purpose. In addition, a single-factor confirmatory factor analysis validated the assumption that the Badger Exam at each grade level is unidimensional, measuring a specific content domain.

CHAPTER 1: INTRODUCTION

1.1: Overview of This Technical Manual

1.2: Background of the Badger Exam

1.3: Appropriate Use of Test Scores, Performance Levels, and Claim Levels

1.4: Glossary of Abbreviations

1.1: Overview of This Technical Manual

This technical manual documents the test development, delivery process, scoring and reporting procedures, and statistical analysis results of the spring 2015 Badger Exam. Following the *Standards for Educational and Psychological Testing* (*Standards*; American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014), this manual also provides the evidence of the reliability of scores and the validity of their recommended interpretations for the Badger Exam.

The content of the technical manual is summarized as follows. Chapter 1 provides the overview of the Badger Exam including a brief history of Smarter Balanced. Chapter 2 describes how tests are designed based on the modified Smarter Balanced test blueprint as well as how operational tests are developed as a fixed form in each grade and subject. Chapter 3 documents the test administration procedures, test security procedures, test accommodations, designated supports, and universal tools. Chapter 4 reviews the scoring process for both machine-scored and hand-scored items as well as the reporting process for interpreting the scores at the individual, school, and district levels. Chapter 5 provides the score calibration and scaling procedures based on item response theory. In Chapter 6, reported test scores, strength and weakness levels (i.e., claim), and performance levels are summarized. Psychometric analyses are provided in Chapters 7 and 8, including reliability and validity, respectively. Finally, Chapter 9 contains a description of the quality control procedures applied.

1.2: Background of the Badger Exam

In 2010, the U.S. Department of Education awarded $330 million to two groups of states—the Smarter Balanced Assessment Consortium (SBAC) and the Partnership for Assessment of Readiness for College and Careers (PARCC)—to develop a valid, reliable, and fair system of next-generation assessments. The assessments designed by these two consortia are intended to be aligned to the Common Core State Standards (CCSS), developed by the National Governors Association and the Council of Chief State School Officers and adopted by more than 40 states. The CCSS defines the knowledge and skills test takers should take away from their K–12 schooling to be successfully prepared for college and career opportunities.

While federal funding supported the research and development work of Smarter Balanced, all policy decisions about the structure and content of the assessments are made by the member states based on input from stakeholders. At the conclusion of the federal grant in September 2014, Smarter Balanced developed an operational assessment system, supported by its member states, which measures mastery of the CCSS and provides timely information about test taker achievement and progress toward college- and career-readiness in ELA and Mathematics. All Smarter Balanced member states and territories, including Wisconsin, adopted academic standards in ELA and Mathematics that are designed to ensure that all test takers gain the knowledge and skills needed to succeed in college and the workplace.

Educational Testing Service (ETS) and its fully owned subsidiary, Computerized Assessments and Learning (CA&L), contracted with the Wisconsin Department of Public Instruction (DPI) to a) deliver the Smarter Balanced assessment system in Wisconsin, b) provide accurate test scores, and c) provide on-time, informative, and secure score reports.

1.3: Appropriate Use of Test Scores, Performance Levels, and Claim Levels

Once tests are administered, scale scores (total test) and claim scores are generated for each subject, and performance is reported at the individual, school, district, and state levels. The total test scale score is used as the basic unit of reporting to determine test takers' performance levels, indicating their level of knowledge and skill in the subject area. When aggregated, scale scores can also describe school- or district-level performance on the tests and can measure gaps in achievement among different groups of test takers.

Claim scores are used to draw inferences about a test taker's achievement in each of several specific knowledge or skill areas covered by each test. Performances on claims are reported as strength and weakness levels to provide teachers, parents, and test takers more detailed diagnostic information about test takers' performance on the test.

Smarter Balanced has also developed performance level descriptors (also known as achievement level descriptors) for ELA and Mathematics that are aligned with the CCSS and the Smarter Balanced assessment claims. The purpose of these descriptors is to specify, in content terms, the knowledge and skills that test takers display at each of the four levels of performance. Defining these levels of performance in terms of cut scores determined in a standard setting procedure is a reporting feature that is federally required under the No Child Left Behind Act. However, characterizing a test taker's performance solely in terms of four categories is an oversimplification. Performance levels should serve only as a starting point for discussion about the achievement of test takers and should not be interpreted as predictors of test takers' futures.

Although the achievement level descriptors provided by Smarter Balanced are intended to aid interpretation of performance levels, they are less precise than scale scores for describing test taker gains over time or changes in achievement gaps among groups, since they do not reveal changes of test taker scores within the bands defined by the performance levels. Furthermore, there is not a critical shift in test taker knowledge or understanding that occurs at a single cut score point defining the break between two levels. Thus, the performance levels should be understood as indicating the achievement of a typical test taker whose score is in the middle of the score range for the performance level.

While not the primary purpose of the Badger Exam, claim scores provide information about test takers' specific strengths and weaknesses. Grouping the test items by claim and then evaluating test takers' responses to this small subset of items make it possible to provide a general estimate of a test taker's strengths or weakness within ELA or Mathematics. Because claim scores are each based on only a subset of items, they are less precise than total scores.

Since each claim score only comprises approximately 10 to 20 items, the strength and weakness levels have been defined very conservatively. Most test takers are classified into the central reporting category of "Meets standard." Very strong evidence is required to describe a test taker's performance as "Does not meet standard" or "Exceeds standard" in strength and weakness levels. The rationale behind this conservative approach is based on expectations of how scores might be used. Because the classification "Does not meet standard" in a claim area suggests the need for some sort of instructional intervention, the decision was made by Smarter Balanced to report such a classification only when there is a high level of certainty.

Because the claim scores are much less precise than the total score, it is possible for a test taker's score report to provide seemingly contradictory information. For example, a test taker's total

score can be at the “Basic” performance level, even though none of the test taker’s claim scores are low enough to classify the test taker as having a weakness in that specific claim.

The information in the score reports (scale scores, performance levels, and claim levels) can be used at the school and district level to help with curriculum and instructional planning as well as for classroom instructional decisions. However, it is important to exercise caution when interpreting individual reports. Important decisions about individual test takers should be based on multiple sources of information. Decisions about the type of instruction test takers receive or their retention in a given grade level in school should be made collaboratively by people who are familiar with the test takers’ progress and achievement: parents, teachers, school counselors, school psychologists, specialist teachers, and perhaps the test takers themselves.

1.4: Glossary of Abbreviations

A glossary of abbreviations is given below for reference.

Abbreviation or Term	Meaning
CA&L	Computerized Assessments and Learning
CCSS	Common Core State Standards
CR	Constructed Response
CSEM	Conditional Standard Error of Measurement
DPI	Wisconsin Department of Public Instruction
EBSR	Evidence-Based Selected Response
ELA	English Language Arts
ELL	English Language Learners
ELP	English Language Proficiency
EQ	Equation Editor
ETS	Educational Testing Service
HOSS	Highest Obtainable Scale Score
HOT	Highest Obtainable Theta
HTQ	Hot Text (Question & Test Interoperability; QTI version)
ICC	Item Characteristic Curve
IEP	Individualized Education Plan
IPR	Individual Profile Report
IRT	Item Response Theory
LEP	Limited English Proficient
LOSS	Lowest Obtainable Scale Score

Abbreviation or Term	Meaning
LOT	Lowest Obtainable Theta
MC	Multiple-Choice Single-Select
MI	Matching Item
MLE	Maximum Likelihood Estimation
MS	Multiple-Choice Multi-Select
ONE	Online Network for Evaluation
PARCC	Partnership for Assessment of Readiness for College and Careers
PT	Performance Task
SA	Short Answer
SBAC	Smarter Balanced Assessment Consortium
SD	Standard Deviation
SE	Standard Error of Estimate
SEM	Standard Error of Measurement
SES	Socioeconomic Status
SR	Selected Response
TCC	Test Characteristic Curve
TI	Fill-in Table Item
TIF	Test Information Function
TOMS	Test Operations Management System
UAAG	Usability, Accessibility, and Accommodations Guidelines

CHAPTER 2: TEST DESIGN AND DEVELOPMENT

2.1: Smarter Balanced Content Framework

2.2: Test Components

2.3: Test Blueprint

2.4: Item and Form Development

2.1: Smarter Balanced Content Framework

The Badger Exam is Wisconsin's customized version of the Smarter Balanced assessment designed to measure test taker progress toward college- and career-readiness. This Chapter summarizes how Smarter Balanced established summative assessments based on evidence-centered design and describes how DPI, ETS, and CA&L adapted the Smarter Balanced test design and its assessment system components. The primary goal of Smarter Balanced was to ensure that its measurement properties reflected the expectations of content and performance that comprise the Common Core State Standards (CCSS). The CCSS are the content standards in ELA and Mathematics that many states have adopted. Since the CCSS were not specifically developed for assessment, Smarter Balanced content experts produced content specifications for the tests in ELA and Mathematics.

Evidence-centered design, an approach to the creation of educational assessments in terms of reasoning about evidence (arguments) concerning the intended constructs, was applied to develop the Smarter Balanced assessments. The evidence-centered design process begins with identifying the claims that users want to make concerning test taker achievement to specifying the evidence needed to support those claims, and finally, determining a specification of the items/tasks capable of eliciting that information (Mislevy, Almond, & Lukas, 2003). The alignment study conducted for the Smarter Balanced consortium discusses alignment among elements of content standards, content specifications, item specifications, and blueprints (Human Resources Research Organization [HumRRO], 2014). The study concluded that Smarter Balanced summative assessment and supporting item pools exceed levels of depth of knowledge representation recommended by Webb (2007).

There are four broad claims within each of the two subject areas in grades 3 to 8, and each claim includes several assessment targets that define more specific performance expectations within claim statements. The claims in ELA and Mathematics are given in Table 2.1. With explicit content specifications linked to CCSS, Smarter Balanced developed both summative assessments and interim assessments. In this technical manual, only summative assessment-related information is summarized. As defined by Smarter Balanced, the goals of summative assessment are: a) accurately describing test taker achievement and modeling growth of test taker learning as part of an accountability system; b) providing valid, reliable, and fair measures of test takers' progress toward, and attainment of, the knowledge and skills required to be college- and career-ready; and c) using performance tasks to provide a measure of the test taker's ability to integrate knowledge and skills. More information about the Smarter Balanced content specification can be found online at http://www.smarterbalanced.org/wordpress/wp-content/uploads/2015/03/ELA_Content_Specs.pdf for ELA, and http://www.smarterbalanced.org/wordpress/wp-content/uploads/2011/12/Mathematics-Content-Specifications_July-2015.pdf for Mathematics.

2.2: Test Components

DPI and ETS modified Smarter Balanced test blueprints/specifications to shape the composition of the two summative assessment components: Section 1 as a computer-based operational linear test, and Section 2 as performance tasks based on classroom activity. Figure 2.1 illustrates the structure of the Badger Exam. Section 1 included a variety of selected response, constructed

response, and technology enhanced items in both subject areas. The Section 1 test was administered on a computer but not tailored for each test taker. Although Smarter Balanced provided a computerized-adaptive testing delivery engine, ETS and CA&L recommended that DPI apply a linear (nonadaptive) test online delivery engine for delivering a test blueprint with a minimum degree of error, and optimizing the item usage from the currently provided Smarter Balanced item pool.

Section 2, also known as the performance task (PT) section, was intended to measure a test taker's ability to integrate knowledge and skills on a coherent task that requires using integrated skill sets. Section 2 was only administered in Mathematics tests and used to measure capacities such as depth of understanding, research skills, and complex analysis, which cannot be completely assessed with individual, discrete items. The PT was preceded by a brief teacher-led activity that is grouped into a larger theme for administration. Detailed information of Smarter Balanced PT specification is available online at http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/05/TaskItemSpecifications/PerformanceTasks/PerformanceTasksSpecifications.pdf.

The intent of the Mathematics classroom activity was to provide context for the PTs. This activity allowed test takers to demonstrate skills and knowledge without interference from lack of background knowledge or vocabulary. The classroom activity did not address the assessed skills but described the setting and provided related examples or terms. Since PTs were often applied skills that are needed in real-world settings, the classroom activity provided test takers with external information so that no test taker was given an advantage or disadvantage based on personal experience.

Teachers within each Wisconsin school district led the classroom activity. An applicable classroom activity was required before the PT portion of the test and the PT had to be administered within three days after the classroom activity. The classroom activity was not supplemented with any other content so as not to detract from the intended purpose of the classroom activity. Please note that PT sections were not provided in the ELA test in the 2015 Badger Exams. The detailed classroom activity process is described in the Badger Exam classroom activity administration guidelines available online at https://dpi.wi.gov/sites/default/files/imce/assessment/pdf/ClassroomActivityAdministrationGuidelines.pdf. The PT overview is available at https://media.dpi.wi.gov/oea/pp/2014-08-12-performance-task-overview-narrated.pptx.

2.3: Test Blueprint

The test blueprint defines the knowledge, skills, and abilities to be measured on an assessment. The test blueprint also describes the number of items of different response types, such as selected response and constructed response, as well as the score points. A test blueprint is a formal document that guides the development and assembly of an assessment by explicating the following types of essential information:

- claims to be supported by the test scores for each assessed subject and grade, across various levels of the system (test taker, classroom, school, district, and state)
- the relative emphasis of content standards, which is generally specified as the number of items or percentage of points per claim

- item types to be used to measure each claim
- the depth of knowledge[2], indicating the complexity of item types for each claim

The test blueprint also informs teachers and test takers about what knowledge and skills the test will measure and how it will measure them.

The Governing States of the Smarter Balanced consortium adopted blueprints for the summative assessments of ELA and Mathematics for grades 3 through 8. The summative test blueprints reflect the depth and breadth of the performance expectations of the CCSS and were developed with broad input from member states, partners, and stakeholders. The process used for developing the test blueprint was described in detail in Chapter 4 (Test Design) of the Smarter Balanced pilot and field test technical report which is available online at http://www.smarterbalanced.org/wordpress/wp-content/uploads/2011/12/Chapter-04-Test-Design-FINAL-20150504-mm.pdf. Final blueprints for the Smarter Balanced summative assessments are available online at http://www.smarterbalanced.org/wordpress/wp-content/uploads/2015/02/ELA_Blueprint.pdf and http://www.smarterbalanced.org/wordpress/wp-content/uploads/2015/02/Mathematics_Blueprint.pdf.

The test blueprints from Smarter Balanced were modified to create unique test blueprints to develop tests for the Badger Exam. Tables 2.2 through 2.13 provide test blueprints for each grade and subject area for the 2015 Badger Exam.

2.4: Item and Form Development

All items in the Smarter Balanced operational item pool were developed and edited during the pilot and field test periods. Item and performance task specifications provide guidance on how to translate the Smarter Balanced content specifications into actual assessment items. In addition, guidelines for bias and sensitivity, accessibility and accommodations, and style help item developers and reviewers ensure consistency and fairness across the item bank. The specifications and guidelines from Smarter Balanced were reviewed by member states, school districts, higher education professionals, and other stakeholders.

Detailed information (such as item/task specification, item writing and writer training, item review and editing) can be found in the following documents that are available online:

- Chapter 3 (Item development) of Smarter Balanced pilot and field test technical report http://www.smarterbalanced.org/wordpress/wp-content/uploads/2015/05/Chapter-03-Item_Development.pdf
- General item specification document http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/05/TaskItemSpecifications/ItemSpecifications/GeneralItemSpecifications.pdf

[2] Norman Webb's (2005) four levels of cognitive complexity (i.e., depth of knowledge) were used for this analysis. These levels were developed specifically for K–12 standards and alignment studies and are widely used in alignment studies throughout the nation.

- Alignment study report
  http://www.smarterapp.org/documents/AlignmentStudyReport.pdf

The Badger Exam grades 3-8 forms were developed in collaboration by DPI, ETS, and CA&L. DPI and ETS constructed the test forms, ensuring that the same plan was used for developing test forms within a single grade and subject area. CA&L handled the production/layout of selected items from the item pool into a fixed form test, and, together as a team, ETS and CA&L ensured that there was a consistent application of style. Although each item could have unique style characteristics, the team understood that it is essential that all innovative items are in their intended and technically correct format. All tests were designed, produced, and printed to fully support test administration and scoring operational plans.

To start this process, DPI, ETS and CA&L analyzed the operational Smarter Balanced item bank to determine the appropriate number of items for a single test form. In 2014, classical item statistics and item response theory-based item statistics were evaluated by Smarter Balanced for the quality of field test items and their eligibility for operational use. Field test items that met quality standards, statistical requirements, and content criteria were retained in the item pool for future operational use. Approved items from the pool were selected by DPI, ETS, and CA&L to construct the Badger Exam.

ETS assessment content specialists and psychometricians jointly selected items according to the test blueprints and specifications. Operational test forms were constructed according to the requirements outlined in the test blueprints and specification. Items were selected to satisfy the test map, meet target test difficulty, and result in an overall test with balanced content. A test development checklist was used by ETS to review the initial tests assembled during the test build. Test forms were built as an iterative process to balance test content and statistical requirements.

Test specifications and blueprints guided the item selection process to ensure that all relevant claims were represented in each operational form. Items were selected by ETS staff to cover a range of difficulty levels on each of the ELA and Mathematics scales as well as the test accommodation conditions such as American Sign Language and braille. Regarding classroom activity and PT assignment, two topics were assigned and spiraled by grade/by school. Selected topics within each grade were statistically comparable (i.e., equal or similar difficulty) under balanced test content specifications. Table 2.14 provides a summary of selected PT topics for Mathematics tests. Hereafter, each PT topic is specified as PT1 or PT2 in each grade.

Table 2.1: Claims Identified for ELA and Mathematics

Subject	Claim	Description
ELA	1. Reading	Test takers can read closely and analytically to comprehend a range of increasingly complex literary and informational texts
	2. Writing	Test takers can produce effective and well-grounded writing for a range of purposes and audiences
	3. Listening	Test takers can employ effective listening skills for a range of purposes and audiences
	4. Research/Inquiry	Test takers can engage in research and inquiry to investigate topics and to analyze, integrate, and present information
Mathematics	1. Concepts and Procedures	Test takers can explain and apply mathematical concepts and interpret and carry out mathematical procedures with precision and fluency
	2. Problem Solving	Test takers can solve a range of complex well-posed problems in pure and applied mathematics, making productive use of knowledge and problem-solving strategies
	3. Communicating/Reasoning	Test takers can clearly and precisely construct viable arguments to support their own reasoning and to critique the reasoning of others
	4. Model and Data Analysis	Test takers can analyze complex, real-world scenarios and can construct and use mathematical models to interpret and solve problems

Note. In Mathematics, claims 2 and 4 are reported together, so there are only three reporting categories with four claims.

Table 2.2: Test Blueprint for ELA Grade 3

Claim/Score Reporting Category	Number of Selected Response Items						Number of Constructed Response Items						Total Number of	
	by Depth of Knowledge*				by Score Point		by Depth of Knowledge				by Score Point		Items	Points
	1	2	3	4	1	2	1	2	3	4	1	2		
1. Reading	2	10	4	1	17				1		1		18	18
2. Writing	4	2			6		3		1		3	1	10	11
3. Listening	3	3	2		8		1				1		9	9
4. Research		7			7			1			1		8	8
Total	9	22	6	1	38		4	1	2		6	1	45	46

Note. Three items have been added to the Research Claim/Score Reporting Category to compensate for removing the ELA PT.

*Webb's (2005) four levels of cognitive complexity were used: 1=Recall, 2=Basic Application, 3=Strategic Thinking, and 4=Extended Thinking.

Table 2.3: Test Blueprint for ELA Grade 4

Claim/Score Reporting Category	Number of Selected Response Items						Number of Constructed Response Items						Total Number of	
	by Depth of Knowledge*				by Score Point		by Depth of Knowledge				by Score Point		Items	Points
	1	2	3	4	1	2	1	2	3	4	1	2		
1. Reading	4	4	6		14		1	3	2		5	1	20	21
2. Writing	3	3			6		3		1		3	1	10	11
3. Listening	2	2	5		9								9	9
4. Research		7			7			1			1		8	8
Total	9	16	11		36		4	4	3		9	2	47	49

Note. Three items have been added to the Research Claim/Score Reporting Category to compensate for removing the ELA PT.

*Webb's (2005) four levels of cognitive complexity were used: 1=Recall, 2=Basic Application, 3=Strategic Thinking, and 4=Extended Thinking.

Table 2.4: Test Blueprint for ELA Grade 5

Claim/Score Reporting Category	Number of Selected Response Items						Number of Constructed Response Items						Total Number of	
	by Depth of Knowledge*				by Score Point		by Depth of Knowledge				by Score Point		Items	Points
	1	2	3	4	1	2	1	2	3	4	1	2		
1. Reading		10	2		12			2	4		4	2	18	20
2. Writing	4	3			7		3				3		10	10
3. Listening		6	3		9								9	9
4. Research		7			7			1			1		8	8
Total	4	26	5		35		3	3	4		8	2	45	47

Note. Three items have been added to the Research Claim/Score Reporting Category to compensate for removing the ELA PT.

*Webb's (2005) four levels of cognitive complexity were used: 1=Recall, 2=Basic Application, 3=Strategic Thinking, and 4=Extended Thinking.

Table 2.5: Test Blueprint for ELA Grade 6

Claim/Score Reporting Category	Number of Selected Response Items						Number of Constructed Response Items						Total Number of	
	by Depth of Knowledge*				by Score Point		by Depth of Knowledge				by Score Point		Items	Points
	1	2	3	4	1	2	1	2	3	4	1	2		
1. Reading	1	6	8	1	16			3	2		5		21	21
2. Writing	4	2			6		2		2		2	2	10	12
3. Listening		4	4		8		1				1		9	9
4. Research		8			8								8	8
Total	5	20	12	1	38		3	3	4		8	2	48	50

Note. Three items have been added to the Research Claim/Score Reporting Category to compensate for removing the ELA PT.

*Webb's (2005) four levels of cognitive complexity were used: 1=Recall, 2=Basic Application, 3=Strategic Thinking, and 4=Extended Thinking.

Table 2.6: Test Blueprint for ELA Grade 7

Claim/Score Reporting Category	Number of Selected Response Items						Number of Constructed Response Items						Total Number of	
	by Depth of Knowledge*				by Score Point		by Depth of Knowledge				by Score Point		Items	Points
	1	2	3	4	1	2	1	2	3	4	1	2		
1. Reading		11	2	2	15		1	2	1	1	4	1	20	21
2. Writing	4	1			5		2	1	2		3	2	10	12
3. Listening	2	3	4		9								9	9
4. Research		6			6			2			2		8	8
Total	6	21	6	2	35		3	5	3	1	9	3	47	50

Note. Three items have been added to the Research Claim/Score Reporting Category to compensate for removing the ELA PT.

*Webb's (2005) four levels of cognitive complexity were used: 1=Recall, 2=Basic Application, 3=Strategic Thinking, and 4=Extended Thinking.

Table 2.7: Test Blueprint for ELA Grade 8

Claim/Score Reporting Category	Number of Selected Response Items						Number of Constructed Response Items						Total Number of	
	by Depth of Knowledge*				by Score Point		by Depth of Knowledge				by Score Point		Items	Points
	1	2	3	4	1	2	1	2	3	4	1	2		
1. Reading		9	3	4	16			1	1	2	3	1	20	21
2. Writing	4	2			6		2		2		2	2	10	12
3. Listening		5	4		9								9	9
4. Research		8			8								8	8
Total	4	24	7	4	39		2	1	3	2	5	3	47	50

Note. Three items have been added to the Research Claim/Score Reporting Category to compensate for removing the ELA PT.

*Webb's (2005) four levels of cognitive complexity were used: 1=Recall, 2=Basic Application, 3=Strategic Thinking, and 4=Extended Thinking.

Table 2.8: Test Blueprint for Mathematics Grade 3

Claim/Score Reporting Category	Number of Selected Response Items							Number of Constructed Response Items							Total Number of	
	by Depth of Knowledge*				by Score Point			by Depth of Knowledge				by Score Point			Items	Points
	1	2	3	4	1	2	3	1	2	3	4	1	2	3		
1. Concepts and Procedures	3				3			10	7			17			20	20
2. Problem Solving 4. Modeling and Data Analysis		1	2		3				4-5	1-2		4-5	0-2	0-1	8-10	8-14
3. Communicating Reasoning		2	2-3		4-5				2	1		2-3	0-1		7-8	7-9
Total	3	3	4-5		10-11			10	13-14	2-3		24	1-2	0-1	36-37	37-41

*Webb's (2005) four levels of cognitive complexity were used: 1=Recall, 2=Basic Application, 3=Strategic Thinking, and 4=Extended Thinking.

Table 2.9: Test Blueprint for Mathematics Grade 4

Claim/Score Reporting Category	Number of Selected Response Items							Number of Constructed Response Items							Total Number of	
	by Depth of Knowledge*				by Score Point			by Depth of Knowledge				by Score Point			Items	Points
	1	2	3	4	1	2	3	1	2	3	4	1	2	3		
1. Concepts and Procedures	1		1		2			12	6			18			20	20
2. Problem Solving 4. Modeling and Data Analysis		1-2	1		2-3				4	2		4-5	1-2		8-9	10
3. Communicating Reasoning		2	3		5				1-2	1-2		2	1		8	9
Total	1	3-4	5		9-10			12	11-12	3-4		24-25	2-3		36-37	39

*Webb's (2005) four levels of cognitive complexity were used: 1=Recall, 2=Basic Application, 3=Strategic Thinking, and 4=Extended Thinking.

Table 2.10: Test Blueprint for Mathematics Grade 5

Claim/Score Reporting Category	Number of Selected Response Items							Number of Constructed Response Items							Total Number of	
	by Depth of Knowledge*				by Score Point			by Depth of Knowledge				by Score Point			Items	Points
	1	2	3	4	1	2	3	1	2	3	4	1	2	3		
1. Concepts and Procedures	3	2			5			10	5			15			20	20
2. Problem Solving 4. Modeling and Data Analysis			2		2				5	2		5	1-2	0-1	9	11-12
3. Communicating Reasoning		3	1		4				2	2		2-3	1-2		8	9-10
Total	3	5	3		11			10	12	4		22-23	2-4	0-1	37	41

*Webb's (2005) four levels of cognitive complexity were used: 1=Recall, 2=Basic Application, 3=Strategic Thinking, and 4=Extended Thinking.

Table 2.11: Test Blueprint for Mathematics Grade 6

Claim/Score Reporting Category	Number of Selected Response Items							Number of Constructed Response Items							Total Number of	
	by Depth of Knowledge*				by Score Point			by Depth of Knowledge				by Score Point			Items	Points
	1	2	3	4	1	2	3	1	2	3	4	1	2	3		
1. Concepts and Procedures	4	4			8			9	3			12			20	20
2. Problem Solving 4. Modeling and Data Analysis		1			1				4-5	2		5-6	1		7-8	8-9
3. Communicating Reasoning		2	1		3				2	2-3	0-1	3	2		8	10
Total	4	7	1		12			9	9-10	4-5	0-1	20-21	3		35-36	38-39

*Webb's (2005) four levels of cognitive complexity were used: 1=Recall, 2=Basic Application, 3=Strategic Thinking, and 4=Extended Thinking.

Table 2.12: Test Blueprint for Mathematics Grade 7

Claim/Score Reporting Category	Number of Selected Response Items							Number of Constructed Response Items							Total Number of	
	by Depth of Knowledge*				by Score Point			by Depth of Knowledge				by Score Point			Items	Points
	1	2	3	4	1	2	3	1	2	3	4	1	2	3		
1. Concepts and Procedures	5	1			6			9	6			15			21	21
2. Problem Solving 4. Modeling and Data Analysis		0-1	1		1-2			1	3-4	1-2	0-1	6-7			8	8
3. Communicating Reasoning			3		3				3		0-1	1-2	1-3		6-7	7-10
Total	5	1-2	4		10-11			10	12-13	1-2	0-2	22-24	1-3		35-36	36-39

*Webb's (2005) four levels of cognitive complexity were used: 1=Recall, 2=Basic Application, 3=Strategic Thinking, and 4=Extended Thinking.

Table 2.13: Test Blueprint for Mathematics Grade 8

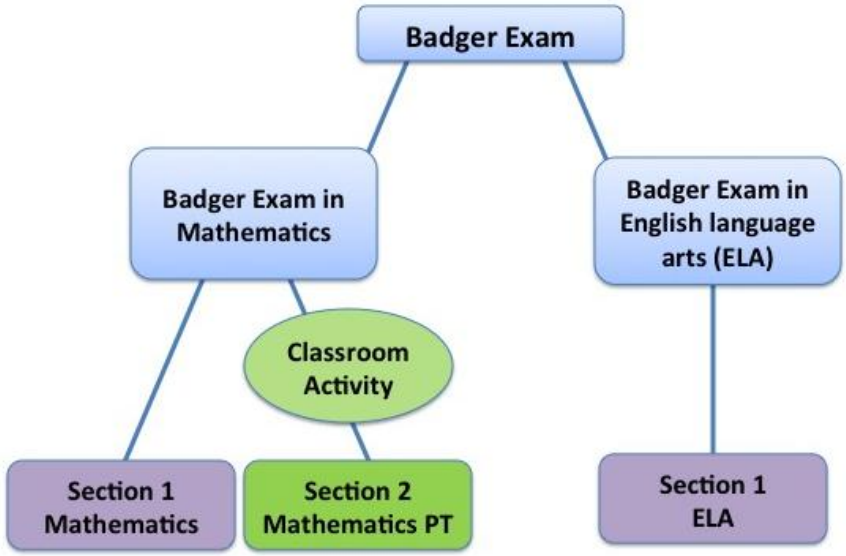
Claim/Score Reporting Category	Number of Selected Response Items							Number of Constructed Response Items							Total Number of	
	by Depth of Knowledge*				by Score Point			by Depth of Knowledge				by Score Point			Items	Points
	1	2	3	4	1	2	3	1	2	3	4	1	2	3		
1. Concepts and Procedures	7	2			9			6	5			11			20	20
2. Problem Solving 4. Modeling and Data Analysis		2	1		3				2-3	2-3		4-6			7-9	7-9
3. Communicating Reasoning		4	2		6				1-2			0-1	1		7-8	8-9
Total	7	8	3		18			6	9	2-3		16-17	1		35-36	36-37

*Webb's (2005) four levels of cognitive complexity were used: 1=Recall, 2=Basic Application, 3=Strategic Thinking, and 4=Extended Thinking.

Table 2.14: Summary of Assigned Classroom Activities in Mathematics Tests

Grade	Topic of Classroom Activity	
	Performance Task 1 (PT1)	Performance Task 2 (PT2)
3	Making Sandwiches	School Library
4	Community Garden	Soccer
5	Space Museum	School Fair
6	Talent Show	Amusement Park
7	Mini Golf	Donuts
8	Signs	Yogurt

Figure 2.1: Test Components of the Badger Exam



CHAPTER 3: TEST ADMINISTRATION

3.1: Testing Schedule and Administration

3.2: Test Security and Confidentiality

3.3: Test Accommodations, Designated Supports, and Universal Tools

3.1: Testing Schedule and Administration

This Section provides a brief overview of the first-operational-year tasks such as training of test administrators, equipment required, timing instructions, and procedures for implementation of test accommodations based on *Standard 7.8* (AERA, APA, & NCME, 2014). The administration window for the Badger Exam was seven weeks from April 13 to May 29, 2015. Specific test administration schedules within that window were determined locally. The Badger Exam was not timed and there were no predetermined session breaks within each testing component. This flexibility enabled districts/schools to tailor the administration schedule to meet their needs and their available technology resources. Guidance on establishing a testing schedule was provided in the District/School Assessment Coordinator Guide which is available at https://dpi.wi.gov/sites/default/files/imce/assessment/pdf/DAC-SAC%20Guide.pdf.

ETS conducted on-site test administration workshops in various locations throughout Wisconsin from February 23 to February 27, 2015. In addition, a number of test administration resources were provided to schools and districts. Manuals, training modules, and recordings of virtual trainings are available at http://dpi.wi.gov/assessment/smarter/resources. These resources include detailed information on topics such as technology readiness, test administration, test security, accommodations, using the test delivery system, and general testing rules. A number of ad hoc documents were also developed to support test administrators. Examples include documents describing the classroom activity process, providing guidance on using the equation tool in mathematics, and guidance on using keyboard commands.

3.2: Test Security and Confidentiality

A number of actions were taken to ensure the security of the Badger Exam and the confidentiality of test taker information in order to maintain the validity, reliability, and fairness of the results. As mentioned in *Standard 7.9* (AERA, APA, & NCME, 2014), "the documentation should explain the steps necessary to protect test materials and to prevent inappropriate exchange of information during the test administration session" (p. 128). Everyone who works with the assessments, communicates test results, and/or receives testing information is responsible for test security, including DPI staff, ETS staff, CA&L staff, district assessment coordinators, school assessment coordinators, test takers, parents, teachers, and cooperative educational service agency staff. The following paragraphs describe how potential test security incidents were prevented prior to testing and how actual security incidents were handled during and after testing.

Upon initial login to the Test Operations Management System (TOMS), all users (i.e., teachers and administrators) were required to sign an electronic confidentiality agreement. A copy of the confidentiality agreement is available at http://dpi.wi.gov/assessment/Smarter/security. By agreeing to the confidentiality agreement, all users pledged to follow the procedures in the Badger Exam manuals, participate in required training, restrict unauthorized access to the assessment, and not disclose any information related to test items, test taker responses, and/or test taker scores. In addition, staff members providing the read aloud accommodation were required to sign a confidentiality agreement to ensure nondisclosure of assessment or test taker information. Districts were directed to retain a copy of this agreement. A copy of this agreement is available at http://dpi.wi.gov/assessment/smarter/accommodations.

DPI developed and delivered online test security training. This training is available at http://dpi.wi.gov/assessment/smarter/trainings. DPI also developed a test security manual to outline test security responsibilities, expectations, and the process for reporting test security incidents. All educators participating in the administration of the Badger Exam were required to participate in the Badger Exam test security training and review the Test Security Manual. Test security training was also incorporated in the on-site Badger Exam test administration workshops. Test security practices were also incorporated into the District/School Assessment Coordinator Guide and the Test Administration Manual.

During the test administration, proctors were instructed to remain in the room and periodically walk around the room to a) ensure that test takers were not using any prohibited electronic devices, b) answer items about technology only, not content, and c) remind test takers to check their test for any items marked for review or not attempted prior to taking a long break or logging out for the day.

Whether intentional or by accident, failure to comply with security rules, either by staff or test takers, constituted a test security incident. Test security incidents were required to be reported to DPI within 24 hours of occurrence via the Incident Report Form. This form was available to all stakeholders at http://dpi.wi.gov/assessment/smarter/security. Table 3.1 lists and defines three levels of security incidents: improprieties, irregularities, and breaches. To support test administrators, a list of examples of improprieties, irregularities, and breaches was included in the Test Administration Manual and District/School Assessment Coordinator Guide.

Upon receipt of an incident report, DPI staff contacted the district assessment coordinator (if they were not the reporting individual). Depending upon the severity of the allegation, the Office of Student Assessment asked the school district to conduct a thorough fact-finding investigation of the alleged incident and report the results of its investigation to the DPI within two weeks. Because the validity and reliability of the examinations rest with the DPI, the Office of Student Assessment ultimately determined whether the incident was declared a testing violation and whether the Badger Exam results were invalidated.

Within the seven-week test window, the test delivery system was available between 7:00 a.m. and 5:00 p.m., Monday through Friday. Test takers were not able to access the assessment outside of these hours. This restriction prevented test takers from memorizing their login information and reaccessing the test in an unsecure location.

Another security feature was the use of the secure browser. All devices used for testing had to be equipped with the Badger Exam secure browser or assistive technology secure browser prior to the assessment. The secure browser was designed to support test security by prohibiting access to external applications or navigation away from the assessment.

3.3: Test Accommodations, Designated Supports, and Universal Tools

DPI is committed to the proposition that all schools, and all test takers within schools, will be held accountable to a common set of high academic content standards. Smarter Balanced is built on a framework of accessibility for all test takers, including accurate measures of achievement and growth for test takers with disabilities and for English learners. Working with member states,

educators, and experts in the field, Smarter Balanced developed a variety of accessibility tools and accommodation options.

The Smarter Balanced Assessment Consortium's Usability, Accessibility, and Accommodations Guidelines (UAAG) were used by school-level personnel and decision-making teams to inform the use of various accessibility tools. The Guidelines apply to all test takers, including test takers who have an Individualized Education Plan (IEP) or a 504 plan (under Section 504 of the Rehabilitation Act of 1973) or who are classified as limited English proficient (LEP). The Guidelines classify accessibility features into three categories — universal tools, designated supports, and accommodations — as defined in Table 3.2. Both embedded and nonembedded versions exist for each of these three categories of assessment resources. Embedded refers to resources provided as digitally delivered components of the test administration system. For example, American Sign Language was embedded in the Badger Exam and was provided as an accommodation. Nonembedded refers to resources that are provided locally or that are essentially separate from the test administration system. For example, use of a personal calculator was permitted as a nonembedded accommodation in accordance with the parameters of the Guidelines. The Usability, Accessibility, and Accommodations Guidelines, which include a comprehensive list of the various accessibility features and examples, are available at http://dpi.wi.gov/assessment/smarter/accommodations.

Prior to administration, test administrators were required to register all embedded and nonembedded designated supports and accommodations into TOMS. The only exception to the TOMS-based request process was the use of Print on Demand. This request had to be submitted to DPI at least two weeks prior to administration via the request for Print on Demand accommodation form. DPI staff reviewed the request and, if approved, staff activated the accommodation in TOMS. This process permitted the test taker to have access to printed versions of test items and/or passages. After testing was completed, test administrators were directed to destroy all printed materials using a secure process.

All designated supports and accommodations enabled in the test delivery system were noted on the footer of each test taker's Individual Profile Report (IPR), regardless of whether the test taker actually used the accessibility resources or not.

Tables 3.3 to 3.5 provide the list of the test accommodations, designated supports, and universal tools available for the Badger Exam and the numbers and percentages of test takers at each grade level for whom these features were enabled.

Table 3.1: Definitions for Three Levels of Test Security Incidents

Levels	Definition
Impropriety	An unusual circumstance that has a low impact on the individual or group and has a low risk of potentially affecting test taker performance on the test, test security, or test validity. These circumstances can be corrected and contained at the local level. An example of an impropriety might include posting a practice item to a social media site by a test taker.
Irregularity	An unusual circumstance that affects an individual or group of test takers who are testing and may potentially influence test taker performance on the test, test security, or test validity. These circumstances can be corrected and contained at the local level but must be submitted in the online system for resolution of the appeal for testing impact.
Breach	An event that poses a threat to the validity of the test. These circumstances have external implications for the Consortium and may result in a decision to remove the test item(s) from the available secure bank. A breach incident must be reported immediately.

Table 3.2: Definitions for Accommodations, Designated Supports, and Universal Tools

Types	Definition
Accommodations	Accommodations are changes in procedures or materials that increase equitable access. Assessment accommodations are designed to generate valid assessment results for test takers who need them to show what they know and can do. Accommodations provided to a test taker must be documented in a current IEP and used during routine instruction. Consortium-approved accommodations do not compromise the learning expectations, construct, grade-level standards, or intended outcome of the assessment.
Designated Supports	Access features of the assessment available for use by any test taker for whom the need has been indicated by an educator (or team of educators working with the parent/guardian and test taker). They are provided as digitally delivered components of the test administration system or separate from it (embedded or nonembedded).
Universal Tools	Access features of the assessment that are provided as digitally delivered components of the test administration system or separate from it (embedded or nonembedded). Universal tools are available to all test takers based on test taker preference and selection.

Table 3.3: Test Accommodations

Accommodation	Content Area	Grade	Number of Total Test Takers	Number of Accommodated Test Takers	Percentage of Accommodated Test Takers
American Sign Language (Embedded)	ELA	3	61,546	32	0.05
		4	61,137	24	0.04
		5	61,614	21	0.03
		6	60,893	30	0.05
		7	60,635	24	0.04
		8	60,882	24	0.04
	Mathematics	3	61,590	26	0.04
		4	61,172	24	0.04
		5	61,613	20	0.03
		6	60,709	31	0.05
		7	60,404	21	0.03
		8	60,705	24	0.04
Braille (Embedded)	ELA	3	61,546	5	0.01
		4	61,137	3	0.00
		5	61,614	7	0.01
		6	60,893	4	0.01
		7	60,635	6	0.01
		8	60,882	5	0.01
	Mathematics	3	61,590	6	0.01
		4	61,172	5	0.01
		5	61,613	8	0.01
		6	60,709	3	0.00
		7	60,404	6	0.01
		8	60,705	7	0.01
Alternate Response Options (Nonembedded)	ELA	3	61,546	85	0.14
		4	61,137	61	0.10
		5	61,614	72	0.12
		6	60,893	48	0.08
		7	60,635	48	0.08
		8	60,882	54	0.09
	Mathematics	3	61,590	61	0.10
		4	61,172	32	0.05
		5	61,613	40	0.06
		6	60,709	40	0.07
		7	60,404	25	0.04
		8	60,705	28	0.05

Table 3.3: Test Accommodations (Cont'd)

Accommodation	Content Area	Grade	Number of Total Test Takers	Number of Accommodated Test Takers	Percentage of Accommodated Test Takers
Speech-to-Text (Nonembedded)	ELA	3	61,546	298	0.48
		4	61,137	306	0.50
		5	61,614	329	0.53
		6	60,893	429	0.70
		7	60,635	385	0.63
		8	60,882	391	0.64
	Mathematics	3	61,590	335	0.54
		4	61,172	336	0.55
		5	61,613	375	0.61
		6	60,709	416	0.69
		7	60,404	375	0.62
		8	60,705	383	0.63
Closed Captioning (Embedded)	ELA Only	3	61,546	41	0.07
		4	61,137	68	0.11
		5	61,614	65	0.11
		6	60,893	74	0.12
		7	60,635	62	0.10
		8	60,882	82	0.13
Read Aloud Stimuli (Nonembedded)	ELA Only	3	61,546	2,850	4.63
		4	61,137	2,872	4.70
		5	61,614	2,721	4.42
		6	60,893	1,976	3.25
		7	60,635	1,849	3.05
		8	60,882	1,681	2.76
Scribe Items (Nonembedded)	ELA Only	3	61,546	907	1.47
		4	61,137	882	1.44
		5	61,614	827	1.34
		6	60,893	449	0.74
		7	60,635	344	0.57
		8	60,882	275	0.45

Table 3.3: Test Accommodations (Cont'd)

Accommodation	Content Area	Grade	Number of Total Test Takers	Number of Accommodated Test Takers	Percentage of Accommodated Test Takers
Text-to-Speech (Embedded)	ELA Only	3	61,546	10,783	17.52
		4	61,137	10,392	17.00
		5	61,614	10,154	16.48
		6	60,893	7,983	13.11
		7	60,635	7,648	12.61
		8	60,882	7,531	12.37
Abacus (Nonembedded)	Mathematics Only	3	61,590	33	0.05
		4	61,172	28	0.05
		5	61,613	43	0.07
		6	60,709	25	0.04
		7	60,404	27	0.04
		8	60,705	20	0.03
Multiplication Table (Nonembedded)	Mathematics Only	3	61,590	673	1.09
		4	61,172	2,382	3.89
		5	61,613	2,566	4.16
		6	60,709	2,422	3.99
		7	60,404	2,242	3.71
		8	60,705	2,054	3.38
Print On Demand (Nonembedded)	ELA	3	61,546	7	0.01
		4	61,137	4	0.01
		5	61,614	7	0.01
		6	60,893	6	0.01
		7	60,635	7	0.01
		8	60,882	7	0.01
	Mathematics	3	61,590	68	0.11
		4	61,172	68	0.11
		5	61,613	57	0.09
		6	60,709	33	0.05
		7	60,404	32	0.05
		8	60,705	23	0.04
Calculator (Nonembedded)	Mathematics Only	3	61,590	707	1.15
		4	61,172	1,138	1.86
		5	61,613	1,232	2.00
		6	60,709	1,456	2.40
		7	60,404	1,673	2.77
		8	60,705	1,657	2.73

Table 3.4: Designated Supports

Designated Support	Content Area	Grade	Number of Total Test Takers	Number of Accommodated Test Takers	Percentage of Accommodated Test Takers
Color Contrast (Embedded)	ELA	3	61,546	365	0.59
		4	61,137	531	0.87
		5	61,614	543	0.88
		6	60,893	411	0.67
		7	60,635	554	0.91
		8	60,882	511	0.84
	Mathematics	3	61,590	367	0.60
		4	61,172	535	0.87
		5	61,613	548	0.89
		6	60,709	431	0.71
		7	60,404	546	0.90
		8	60,705	511	0.84
Magnification (Nonembedded)	ELA	3	61,546	78	0.13
		4	61,137	153	0.25
		5	61,614	292	0.47
		6	60,893	78	0.13
		7	60,635	79	0.13
		8	60,882	71	0.12
	Mathematics	3	61,590	79	0.13
		4	61,172	153	0.25
		5	61,613	298	0.48
		6	60,709	97	0.16
		7	60,404	78	0.13
		8	60,705	74	0.12
Masking (Embedded)	ELA	3	61,546	1,765	2.87
		4	61,137	1,867	3.05
		5	61,614	2,163	3.51
		6	60,893	2,230	3.66
		7	60,635	2,307	3.80
		8	60,882	1,885	3.10
	Mathematics	3	61,590	1,836	2.98
		4	61,172	1,986	3.25
		5	61,613	2,187	3.55
		6	60,709	2,250	3.71
		7	60,404	2,243	3.71
		8	60,705	1,913	3.15

Table 3.4: Designated Supports (Cont'd)

Designated Support	Content Area	Grade	Number of Total Test Takers	Number of Accommodated Test Takers	Percentage of Accommodated Test Takers
Noise Buffers (Nonembedded)	ELA	3	61,546	604	0.98
		4	61,137	601	0.98
		5	61,614	706	1.15
		6	60,893	331	0.54
		7	60,635	265	0.44
		8	60,882	181	0.30
	Mathematics	3	61,590	613	1.00
		4	61,172	604	0.99
		5	61,613	713	1.16
		6	60,709	313	0.52
		7	60,404	257	0.43
		8	60,705	189	0.31
Permissive Mode[3] (Embedded)	ELA	3	61,546	407	0.66
		4	61,137	428	0.70
		5	61,614	409	0.66
		6	60,893	509	0.84
		7	60,635	492	0.81
		8	60,882	479	0.79
	Mathematics	3	61,590	442	0.72
		4	61,172	463	0.76
		5	61,613	457	0.74
		6	60,709	484	0.80
		7	60,404	470	0.78
		8	60,705	463	0.76
Read Aloud (Nonembedded)	ELA	3	61,546	3,576	5.81
		4	61,137	3,496	5.72
		5	61,614	3,299	5.35
		6	60,893	2,351	3.86
		7	60,635	2,265	3.74
		8	60,882	2,109	3.46
	Mathematics	3	61,590	488	0.79
		4	61,172	593	0.97
		5	61,613	671	1.09
		6	60,709	1,171	1.93
		7	60,404	1,060	1.75
		8	60,705	1,000	1.65

[3] Students using braille devices (Jaws and Voice-Over), speech-to-text, or other assistive technology or software must have "Permissive mode" enabled in their settings in TOMS. Permissive mode is a setting that allows assistive software and devices to function with the Assistive Technology WDPI Badger Exam Secure Browser.

Table 3.4: Designated Supports (Cont'd)

Designated Support	Content Area	Grade	Number of Total Test Takers	Number of Accommodated Test Takers	Percentage of Accommodated Test Takers
Scribe (Nonembedded)	ELA	3	61,546	497	0.81
		4	61,137	589	0.96
		5	61,614	555	0.90
		6	60,893	298	0.49
		7	60,635	247	0.41
		8	60,882	180	0.30
	Mathematics	3	61,590	958	1.56
		4	61,172	906	1.48
		5	61,613	837	1.36
		6	60,709	471	0.78
		7	60,404	338	0.56
		8	60,705	282	0.46
Separate Setting (Nonembedded)	ELA	3	61,546	6,092	9.90
		4	61,137	6,400	10.47
		5	61,614	6,277	10.19
		6	60,893	5,644	9.27
		7	60,635	5,821	9.60
		8	60,882	5,576	9.16
	Mathematics	3	61,590	6,146	9.98
		4	61,172	6,456	10.55
		5	61,613	6,327	10.27
		6	60,709	5,677	9.35
		7	60,404	5,779	9.57
		8	60,705	5,598	9.22
Text-to-Speech (Embedded)	ELA	3	61,546	128	0.21
		4	61,137	161	0.26
		5	61,614	173	0.28
		6	60,893	1,594	2.62
		7	60,635	1,699	2.80
		8	60,882	1,758	2.89
	Mathematics	3	61,590	11,079	17.99
		4	61,172	10,818	17.68
		5	61,613	10,568	17.15
		6	60,709	9,341	15.39
		7	60,404	9,018	14.93
		8	60,705	8,821	14.53

Table 3.4: Designated Supports (Cont'd)

Designated Support	Content Area	Grade	Number of Total Test Takers	Number of Accommodated Test Takers	Percentage of Accommodated Test Takers
Translated Test Directions (Embedded)	Mathematics	3	61,590	1,321	2.14
		4	61,172	975	1.59
		5	61,613	848	1.38
		6	60,709	345	0.57
		7	60,404	337	0.56
		8	60,705	388	0.64
Translation Stacked (Embedded)	Mathematics	3	61,590	1,322	2.15
		4	61,172	986	1.61
		5	61,613	900	1.46
		6	60,709	339	0.56
		7	60,404	332	0.55
		8	60,705	390	0.64
Translated Glossaries (Embedded)	Mathematics	3	61,590	1,834	2.98
		4	61,172	1,557	2.55
		5	61,613	1,229	1.99
		6	60,709	762	1.26
		7	60,404	672	1.11
		8	60,705	811	1.34
Bilingual Dictionary (Nonembedded)	ELA Only	3	61,546	278	0.45
		4	61,137	204	0.33
		5	61,614	369	0.60
		6	60,893	93	0.15
		7	60,635	108	0.18
		8	60,882	121	0.20

Table 3.5: Universal Tools

Universal Tool	Content Area	Grade	Number of Total Test Takers	Number of Accommodated Test Takers	Percentage of Accommodated Test Takers
Scratch Paper	ELA	3	61,546	5,822	9.46
		4	61,137	5,896	9.64
		5	61,614	5,866	9.52
		6	60,893	5,361	8.80
		7	60,635	5,036	8.31
		8	60,882	5,255	8.63
	Mathematics	3	61,590	6,204	10.07
		4	61,172	6,480	10.59
		5	61,613	6,124	9.94
		6	60,709	5,719	9.42
		7	60,404	5,384	8.91
		8	60,705	5,636	9.28
Thesaurus	ELA Only	3	61,546	2,527	4.11
		4	61,137	2,703	4.42
		5	61,614	2,822	4.58
		6	60,893	2,699	4.43
		7	60,635	2,532	4.18
		8	60,882	2,589	4.25

Note. Schools and districts were not required to enter universal tools requests into TOMS. These features were allowed at schools/districts' discretion.

CHAPTER 4: SCORING AND REPORTING

4.1: Scoring Procedures

4.1.1: Types of Item Response

4.1.2: Machine Scoring

4.1.3: Hand Scoring

4.2: Reports to Be Produced and Scores for Each Report

4.1: Scoring Procedures

This Section documents how ETS Assessment Development, Research and Statistical Analysis, Performance Assessment Scoring Service, and Information Technology groups all participated in certifying the scoring system and how each team followed procedures required by the ETS Office of Quality for operational readiness and *Standard 7.8*[4] (AERA, APA, & NCME, 2014).

ETS reviewed operational answer keys and scoring rubrics provided by Smarter Balanced and followed scoring specifications to enter scores into the ETS operational scoring system. A central aspect of the validity of Badger Exam test scores is the degree to which scoring rubrics are related to the appropriate assessment targets, depth of knowledges, and claims based on Smarter Balanced assessments. A key facet of reliability is whether scoring rules are applied accurately during scoring sessions. The test delivery system in the Test Operations Management System (TOMS) displays many item types, from the traditional selected response and short and extended response to the more advanced technology enhanced items. Some items are machine-scored, while others are human-scored, depending on item type and scoring requirements provided by Smarter Balanced item specifications.

4.1.1: Types of Item Response

ELA and Mathematics portions of the Badger Exam contained three types of items developed by Smarter Balanced: selected response (SR) items, constructed response (CR) items, and technology enhanced items. Detailed item type descriptions can be found in the Smarter Balanced item specifications available at http://www.smarterbalanced.org/smarter-balanced-assessments/.

SR items allow the opportunity to assess a broad range of content in one test. SR items are objectively scored and test taker results are collected quickly. However, SR items do not reveal a test taker's reasoning process and typically do not assess higher-order thinking skills. There are several types of SR items, including multiple-choice single-select and multiple-choice multi-select. Multiple-choice items require the test taker to choose a single best answer, while multi-select items require the test taker to mark all the answers that are correct. The partial-credit score points were awarded based on a specific scoring rule for each multiple-choice multi-select item.

The CR items on the Badger Exam are brief open-response items that focus on a particular skill or concept and require test takers to produce a short written response. CR items allow test takers to demonstrate their use of complex thinking skills such as formulating comparisons or contrasts; proposing cause and effects; identifying patterns or conflicting points of view; categorizing, summarizing, or interpreting information; and developing generalizations, explanations, justifications, or evidence-based conclusions (Darling-Hammond & Pecheone, 2010). There are several types of CR item types, including short answer, hot text, table interaction, match interaction, equation, and evidence-based selected response items. These items are scored using rubrics written specifically for each item, and these

[4] Test documentation should include detailed instructions on how a test is to be administered and scored (p.127).

items are scored on a 0 to 2 point score scale for ELA and a 0 to 3 point score scale for Mathematics.

Technology enhanced items are computer-delivered items that include specialized interactions for collecting response data. Technology enhanced items can take advantage of drag-and-drop, hot spot, and simulation technologies along with the use of online tools to measure content that was previously not assessed or was assessed through constructed response item formats requiring more elaborate scoring procedures. Note that some of the CR item types were included in technology enhanced items.

The ELA tests draw approximately 60% to 70% of total score points from SR items and 30% to 40% from technology enhanced and short answer items. The Mathematics tests draw approximately 25% to 50% of score points from SR items and 50% to 75% from technology enhanced and short answer items. Most of these items (89% to 98%) in both ELA and Mathematics are machine-scored, but some CR items are scored by trained and certified human raters.

4.1.2: Machine Scoring

Most item types (e.g., multiple-choice single-select, multiple-choice multi-select, and so on) on the Badger Exam allow for machine scoring. Smarter Balanced provided all item content and all scoring tables for these items. Test taker responses to these machine-scored items were captured via TOMS and then compared to the provided scoring tables to invoke scoring and store the score. The system scored machine-scorable items in real time. After the completion of the hand scoring, results for both item types were integrated to generate an overall score. The detailed overall test score calibration process is provided in Chapter 6.

4.1.3: Hand Scoring

Hand-scored items (mostly short-answer item type) were scored in the ETS Online Network for Evaluation (ONE) system, a distributed, web-based scoring system that enables a large number of raters to view and score assigned responses from remote locations. DPI allowed for remote scoring provided the raters signed confidentiality agreement forms. All identifying information from the responses sent to raters was removed so that neither the identity of the test taker nor the test taker's school was revealed to the rater; the rater saw only the test taker response.

Rater Recruitment/Qualifications

ETS established procedures for recruiting, training, and certifying raters for hand scoring. While there was no requirement that a specified percentage of the raters be Wisconsin educators, ETS welcomed all Wisconsin educators to apply for scoring and had a target of recruiting Wisconsin teachers to represent 20% of the total number of raters needed for hand scoring.

ETS recruited raters through social media such as LinkedIn and CareerBuilder and through nationwide teachers' associations. In addition, ETS developed a one-page electronic and printable flyer for DPI to help recruit Wisconsin educators. DPI posted the flyer on the DPI

website. Recruited teachers followed the same hiring, training, and certification process as all ETS raters. Each rater met the following minimum requirements:

- an undergraduate degree from an accredited college or university in the United States
- residency in the United States
- availability to work in the United States
- a practicing or former teacher

All potential raters who met qualifications were sent a welcome packet with new hire documentation to complete. This included completion of an application with a number of prescreening questions and an e-verify check. Pending review of the application, all hired raters were required to complete an I-9 form (i.e., Employment Eligibility Verification), certification, and training.

The total number of raters and scoring leaders that scored Wisconsin constructed responses was 1,023 individuals. Among them, a total of 59 educators were hired from Wisconsin to score the Badger Exam and 12 are currently employed in public school districts as teachers and participated in scoring.

Leadership Recruitment/Qualifications

Accurate scoring of large numbers of test taker responses requires a comprehensive scoring and leadership structure. The organizational structure for Wisconsin encompassed four levels of responsibility:

Raters. Participating members scored assigned responses, based on their availability.

Scoring leaders. The scoring leaders' primary job was to monitor and report on a team of raters. Leaders read responses that their raters had scored, to see if the raters were applying the scoring rubrics correctly, and to correct the responses if raters were not (i.e., back-read). Scoring leaders were also expected to answer questions and resolve nonroutine (but simple) responses raised by raters.

Group scoring leaders. These team members provided feedback to scoring leaders while carefully monitoring the overall quality and progress of the scoring. They scored complex, nonroutine responses and resolved any content-related issues raised by leaders.

Content scoring leaders. Working under the supervision of ETS Assessment Development content experts, team members had overall responsibility for one or more assessments. Content scoring leaders worked across the leadership team for their domain, such as the ELA upper-level grades, to escalate nonroutine issues (such as test security cases), review the performance of the group scoring leaders, and oversee the quality and progress of the scoring. The content scoring leaders worked closely with staff from ETS Assessment Development, Performance Assessment Scoring, and Human Resources groups.

In addition to the role-based formal system and content training, features within the ETS scoring management platform facilitated tight control and communication between raters and their leaders — as well as between the leadership team members.

Training

For each of the different content area assessments, raters were trained to apply scoring rules, as specified in the Smarter Balanced rubrics and test design. Raters learned to apply the

rubrics for each prompt by scoring "benchmark" responses at each score level. A benchmark response is an actual test taker-written response that illustrates the quality expected for responses receiving that score. The goal of this training is to have all raters apply the same criteria and standards so that the score a response receives will depend as little as possible on which rater scores it. After completing their training, the raters had to pass a certification test by correctly scoring a set of responses that had been previously scored by expert raters. Only after passing the certification test could they begin to score responses operationally. ETS Assessment Development staff conducted the training for raters of the Badger Exam, using sample responses provided by Smarter Balanced. Scoring leaders received additional training on grade-specific prompt types prior to the spring 2015 scoring:

- content scoring leaders – 8 hours of training
- group scoring leaders – 6 hours of training
- scoring leaders – 4 hours of training

Unscored Responses

Four types of test taker responses were left unscored:

- blank responses
- responses written in a language other than English
- responses that did not give the rater enough information to assign a valid rating
- responses that were "off topic," i.e., did not reflect an attempt to answer the item

In the ELA tests, about 0.2 to 2.2 percent of the responses to human-scored items were blank; less than 0.1 percent of the responses were written in a language other than English; about 1.2 to 3.5 percent did not provide enough information to be scored; and there were no off-topic responses. In the Mathematics tests, about 0.1 to 2.1 percent of the responses were blank; less than 0.5 percent were written in a language other than English; about 0.1 to 3.6 percent did not provide enough information to be scored; and there were no off-topic responses.

Hand Scoring Process – Certification and Calibration

Through the certification process, raters had access to the scoring rubric for each prompt, benchmark papers exemplifying each rating the rater could assign, and, as appropriate, scoring notes with information specific to the prompts. After training on certification materials, raters were provided a set of training papers to practice scoring. The certification test required the raters to assign scores on a set of prescored test taker responses. ETS set the pass threshold for certification based upon the requirements for interrater reliability for these Smarter Balanced assessments. After their training, raters were given two chances to pass the certification test. Raters who did not pass on their first attempt were given additional practice. Raters who did not pass on their second attempt were paid for their practice time but were not accepted for scoring. This structure supported the creation of the pool of qualified raters needed for completion of all scoring. In addition to the requirements listed above, inactive raters for more than 90 days were required to recertify before resuming scoring.

Calibration sets were presented to raters daily, and each calibration set had 10 responses. Before scoring responses to a particular prompt type at a particular grade level, raters were

required to pass at least one of two calibration sets of responses to that prompt type by test takers at that grade level. Raters should have at least a 70 percent exact match rate. If the rater was unsuccessful on the first attempt, the rater was required to review the Smarter Balanced materials (scoring guide, benchmark responses, etc.) with the scoring leader and then participate in a second calibration attempt. Raters who did not pass after two calibration attempts were excused from the scoring session.

Tables 4.1 and 4.2 show the distributions of scores on the human-scored items in ELA (Table 4.1) and in Mathematics (Table 4.2). All human-scored ELA items were two-point items. Most human-scored Mathematics items were two-point items, but a few were three-point items. None of the responses were classified as off topic. About 10 percent of all human-scored items were read by additional raters as a validity check (i.e., read-behind). Interrater reliability statistics were regularly analyzed to verify that raters were scoring consistently and accurately (See Chapter 7).

Quality Control

ETS created performance scoring reports so project leadership could monitor the daily human scoring process and plan the retraining activities if needed. Scoring leaders were able to read behind raters and to monitor scoring performance in real-time with a variety of performance data. Validity responses were nine percent prescored (i.e., monitor papers) within each rater's set of assigned responses in order to evaluate ongoing accuracy in real-time. Real-time management tools allowed everyone, from scoring leaders to content specialists, access to aggregate reporting on multiple levels:

- a specific rater's performance such as read rate
- check-set agreement
- overall validity response scores
- overall interrater reliability rate
- projected date for completion of the scoring for a specific prompt or task

Maintaining Consistency of Score Quality

ETS has developed a variety of procedures for controlling the quality of the rating and monitoring the consistency of the raters. These procedures specify everything from rater qualifications to rater certification and daily calibration. New raters were required to demonstrate their accuracy by a) passing a certification test before ETS assigned them to score a specific assessment and then b) passing a shorter, more focused calibration test before each scheduled scoring session. This procedure was a key tool in maintaining quality and consistency. Scoring leaders monitored raters' performance by reading responses that each rater had scored to see if the rater had assigned the correct rating. Some scoring leaders chose to read the response before finding out what score the rater had assigned; others chose to know what score the rater had assigned before reading the response.

4.2: Reports to Be Produced and Data for Each Report

ETS TOMS, a secure website designed to manage the Badger Exam, was provided to DPI. This was the system through which users interact with and inform the test delivery system.

This system used a role-specific design to restrict access to certain tools and applications based on the user's designated role. Specific functions of TOMS were:

- confirm the accuracy of test takers' test settings (i.e., accommodations) prior to testing
- generate testing sessions
- print test tickets that test takers will need during the login process
- monitor test takers' testing progress

In addition, TOMS included an online reporting system that provided authorized users with interactive and cumulative online reports for ELA and Mathematics at the individual, school, and district levels. Based on the Smarter Balanced reporting requirements for ELA and Mathematics, TOMS provided the summative reports containing information outlining test taker knowledge and skills, as well as achievement levels aligned to the assessment-specific claims. The online reports provided drill-down functionality at the district, school, classroom, and test taker levels. The online reports are available to be downloaded in pdf, Excel, and CSV format. The following reports are available in the TOMS reporting suite:

- Individual Profile Report (IPR) – IPR is provided as a printable pdf. The IPR is the official score report for the parents or guardians, telling them how their child performed on the Badger Exam. This report includes the test taker's overall performance on each test reported as a scale score, performance level, and tested topics (i.e., claims). In addition, the purpose of the Badger Exam, definition of performance levels, and information about accommodations and designated supports are described.
- School Reports
  - The school performance report provides group information by subject area including the school average scale score and percentage of test takers at or above proficient. This report also provides a list of test takers' scale scores, performance levels, and claim level information.
  - The school scale score report is presented as a dashboard to provide group information by subject area. It includes a histogram showing the distribution of test takers' scale scores.
- District Reports
  - The district performance report provides group information by subject area including the school average scale score and percentage of test takers at or above proficient. This report also lists all the proficiency information for each school, including the testing status, number of test takers who completed testing, average score, and percentage of test takers in each proficiency level.
  - The district scale score report is presented as a dashboard to provide cumulative information. The histogram shows the frequency of schools with mean scores in each score interval.

Detailed information of score reports can be found from TOMS Reporting Suite User Manual and Interpretive Guide, which is available at https://dpi.wi.gov/sites/default/files/imce/assessment/pdf/TOMS%20Reporting%20and%20Interpretative%20Guide.pdf.

Table 4.1: Score Distributions of Human-Scored ELA Items

Grade	Number of responses scored	Item Number	Percentage of Responses Receiving Score of			Number of responses not scored, for each reason				
			0	1	2	Total	Blank response	Insufficient response	Non-scorable language	Off topic
3	61,525	11	32.9	38.3	24.9	2,377	196	2,165	16	0
4	61,121	14	34.3	40.5	23.5	995	141	846	8	0
		35	63.4	27.1	6.4	1,879	369	1,505	5	0
5	61,600	27	27.8	43.3	27.0	1,174	248	919	7	0
		31	38.4	44.8	14.7	1,324	268	1,050	6	0
6	60,859	13	55.0	28.0	14.2	1,656	537	1,116	3	0
		16	51.0	34.7	10.9	2,058	672	1,380	6	0
7	60,590	5	37.6	44.5	15.9	1,211	462	743	6	0
		12	45.2	40.5	11.2	1,825	698	1,123	4	0
		21	55.4	29.9	10.9	2,352	911	1,435	6	0
8	60,842	5	37.1	45.6	14.5	1,684	691	990	3	0
		11	42.2	37.4	16.9	2,154	868	1,281	5	0
		26	35.8	38.2	20.6	3,282	1,360	1,917	5	0

Table 4.2: Score Distributions of Human-Scored Mathematics Items

Grade	Number of responses scored	Item Number	Percentage of Responses Receiving Score of				Number of responses not scored, for each reason				
			0	1	2	3	Total	Blank response	Insufficient response	Non-scorable language	Off topic
3	30,484	35 (PT1)	61.1	15.9	21.0		599	48	390	161	0
	31,087	34 (PT2)	0.0	0.0	99.8		53	53	0	0	0
		35 (PT2)	15.9	12.5	25.3	45.8	149	127	0	22	0
		36 (PT2)	44.3	53.5	0.0		672	66	468	138	0
		37 (PT2)	59.2	21.8	14.4		1,395	133	1,108	154	0
4	30,053	34 (PT1)	68.9	30.1	0.0		306	19	229	58	0
		35 (PT1)	69.5	12.5	16.5		454	44	342	68	0
		36 (PT1)	59.8	9.9	28.3		611	66	466	79	0
	31,095	34 (PT2)	64.6	34.1	0.0		424	35	269	120	0
		35 (PT2)	56.4	17.1	24.5		624	49	441	134	0
		36 (PT2)	49.0	49.3	0.0		505	57	344	104	0
		37 (PT2)	65.5	14.9	17.4		679	85	461	133	0
5	30,155	34 (PT1)	60.5	19.5	18.2		535	98	345	92	0
		35 (PT1)	79.5	7.5	9.2		1,123	134	899	90	0
		36 (PT1)	65.8	17.6	13.8		838	202	537	99	0
		37 (PT1)	55.4	8.1	33.1		1,041	150	792	99	0
	31,442	34 (PT2)	66.6	32.6	0.0		254	22	112	120	0
		35 (PT2)	29.2	31.8	37.1		581	36	404	141	0
		36 (PT2)	67.9	15.9	14.4		587	79	386	122	0
		37 (PT2)	81.1	9.9	4.3	1.9	858	161	587	110	0

Note. On Grade 3 item 34 (PT2) all responses that were not blank received full credit due to a Spanish translation issue.

Table 4.2: Score Distributions of Human-Scored Mathematics Items (Cont'd)

Grade	Number of responses scored	Item Number	Percentage of Responses Receiving Score of				Number of responses not scored, for each reason				
			0	1	2	3	Total	Blank response	Insufficient response	Non-scorable language	Off topic
6	30,167	32 (PT1)	62.2	21.2	15.6		306	63	228	15	0
		33 (PT1)	62.9	18.2	17.6		392	103	274	15	0
		34 (PT1)	78.1	21.7	0.0		72	60	12	0	0
		35 (PT1)	82.3	7.0	9.5		367	114	236	17	0
	30,523	33 (PT2)	43.5	29.8	25.9		223	63	146	14	0
		34 (PT2)	70.7	28.8	0.0		155	150	5	0	0
		35 (PT2)	25.6	21.8	51.2		411	84	302	25	0
		36 (PT2)	71.8	16.4	8.7		944	253	661	30	0
7	60,368	21	67.0	24.1	7.1		1,094	407	631	56	0
	30,386	34 (PT1)	85.1	6.5	3.6		1,464	406	1,028	30	0
		35 (PT1)	43.9	32.9	20.2		913	296	580	37	0
	29,982	34 (PT2)	86.8	11.6	0.0		486	154	308	24	0
8	28,275	34 (PT1)	74.0	16.4	7.3		651	200	433	18	0
		35 (PT1)	64.9	33.0	0.0		590	188	375	27	0
	32,389	35 (PT2)	74.3	14.1	6.8		1,573	696	860	17	0

CHAPTER 5: SCORE CALIBRATION AND SCALING

5.1: Item Response Theory (IRT) based Calibration Procedure

5.1.1: Calibration Models

5.1.2: Test Taker Ability Estimation for Total and Claim Test Scores

5.2: Deriving Scale Scores for the Badger Exam

5.3: Score Verification Procedure

5.4: Samples Used for the Analyses

5.1: Item Response Theory (IRT) based Calibration Procedure

The primary purpose of the Smarter Balanced assessments is to provide valid, reliable, and fair information about test takers' ELA and Mathematics performance with respect to the Common Core State Standards in grades 3 through 8. For federal accountability purposes, and potentially for state and local accountability systems, test takers' ELA and Mathematics proficiency must be reported. This Section explains the methods used to generate the Badger Exam overall and claim test scores using item response theory (IRT) pattern scoring.

As demonstrated by years of successful application in K–12 testing programs, IRT methods have the flexibility and strength to support the Badger Exam. IRT methods are ideally suited to the assessments and measurement goals of Smarter Balanced in both establishing a common scale and ongoing maintenance of the program. Mixed-item-format tests that consist of dichotomous (selected response) items, short answer responses, and performance tasks were combined together and scaled concurrently.

The purpose of the IRT calibration (and scaling) was to place item difficulty and test taker ability estimates onto a common theta scale in each subject area. The Common Core State Standards provided a foundation for developing Smarter Balanced assessments that support inferences concerning test taker change in achievement (i.e., growth). One approach to modeling test taker growth across grades is to report scores on a common vertical scale. A vertical scale is a single scale for scores on tests at different grade levels of the same subject area. Its purpose is to report the scores in a way that shows a test taker's growth in the subject, from one grade level to the next. For instance, comparing the means and standard deviations for scale scores across grades on the same scale is an intuitive approach for evaluating growth for a variety of test users. Vertical scales assume that increasing test taker proficiency is demonstrated across different levels of the assessment. Another advantage of vertical scaling is that growth expectations concerning the establishment of achievement levels across grades can be inspected and ordered by standard setting panelists.

The basic steps in the process of scaling the scores are:

1. Calibrate the items at each grade level onto an ability scale defined for that grade level.
2. Transform the ability scales at the different grade levels onto a common ability scale, using a separate transformation for each grade level.
3. Transform the ability scale into the reported score scale by applying a single linear transformation for all grade levels.

Test scores for the 2015 Badger Exam were calculated to the baseline scale since all items were pre-equated. The baseline scale was determined as performance level standards were defined by Smarter Balanced following the field test administration in 2014.

5.1.1: Calibration Models

Unidimensional IRT models were used to calibrate the SR and CR items. Using the criteria and results from the pilot/field tests conducted by Smarter Balanced Assessment Consortium, the two-parameter logistic (2PL) model (Birmbaum,1968) and the generalized partial credit model (GPCM; Muraki, 1992) were chosen for the Badger Exam for calibration and scoring

test takers. IRT pattern scoring was the method used to estimate test takers' abilities required for scoring. For selected response items, the 2PL model is:

$$P_i(\theta_j) = \frac{\exp[Da_i(\theta_j - b_i)]}{1 + \exp[Da_i(\theta_j - b_i)]},$$

where $\boldsymbol{P_i(\theta_j)}$ is the probability of a correct response to item *i* by a test taker *j* with ability($\boldsymbol{\theta_j}$), $\boldsymbol{a_i}$ is the discrimination parameter, $\boldsymbol{b_i}$ is the difficulty parameter for item *i*, and *D* is a constant that puts the $\boldsymbol{\theta}$ ability scale into the same metric as the normal ogive model (*D*=1.7).

For constructed response items, the GPCM is given by:

$$P_{ih}(\theta_j) = \frac{\exp \sum_{v=1}^{h}[Da_i(\theta_j - b_i + d_{iv})]}{\sum_{c=1}^{n_i} exp[\sum_{v=1}^{c} Da_i(\theta_j - b_i + d_{iv})]},$$

where $\boldsymbol{P_{ih}(\theta_j)}$ is the probability of test taker *j* obtaining a score of *h* on item *i*, $\boldsymbol{n_i}$ is the number of item score categories, $\boldsymbol{b_i}$ is the item location parameter, $\boldsymbol{d_{iv}}$ is the category parameter for item *i* for category *v*, $\boldsymbol{a_i}$ is the discrimination parameter, and *D* is a scaling constant given previously.

Item parameters used in the Badger Exam tests were estimated and linked during the Smarter Balanced field test administration. Thus, item parameter calibration software, model-to-data fit, and evaluating vertical scale anchor items are not described in the current technical manual. For more detailed information, please refer to http://www.smarterbalanced.org/wordpress/wp-content/uploads/2011/12/Chapter-6.pdf.

Table 5.1 provides summary statistics describing the distribution of item difficulty and discrimination parameters at each grade level from the field test calibration and scaling. Overall, Mathematics forms were more discriminating than ELA forms. The difficulty of the Mathematics forms increased as the grade levels increased. However, this was not the case for ELA test forms. Fourth-grade ELA test forms were more difficult, on average, than third-grade forms, and sixth-grade ELA forms were more difficult than seventh-grade forms. These same inconsistencies in ELA can be seen in the test characteristic curves (TCCs) shown in Figures 5.1 and 5.2. These curves model the relationship between test taker ability and expected scoring outcomes at the test level. The ability level required for a given percent-correct score should increase consistently from each grade level to the next. This consistent grade-to-grade increase occurred for the Mathematics tests but not for the ELA tests. Particularly, the difficulty of the 6th and 7th grade ELA tests was so similar that the TCCs for these two tests were nearly identical.

Chapter 9 from the Smarter Balanced field test technical report provides more detailed information about how Smarter Balanced assessments were both horizontally and vertically calibrated and scaled through IRT processes. The report is available at http://www.smarterbalanced.org/wordpress/wp-content/uploads/2015/06/chapter-9.pdf.

5.1.2: Test Taker Ability Estimation for Total and Claim Test Scores

Test takers' total scores were based on maximum likelihood estimates (MLE) of their overall ability within each subject; each of their claim scores was based on maximum likelihood estimates of their ability in the specific skills included in the claim. Each claim score was estimated from the test taker's responses to the items assessing that particular claim. The likelihood of the test taker j's ability (θ_j), given the observed response vector (z_j) and the item discrimination and difficulty parameters (a_i and b_i), is:

$$L\left(\theta_j \middle| z_j, a_i, b_i\right) = \prod_{i=1}^{n} P_i\left(z_j \middle| \theta_j, a_i, b_i\right),$$

where $P_i\left(z_j \middle| \theta_j, a_i, b_i\right)$ is the probability of correct response on item i. Estimated test taker ability was proportionally adjusted based on the percentage of items included in the score calculation to the number of items for a full test. For example, when a test taker answers only the first 30 items on the 40-item test, the ability estimated by 30 items is adjusted by the proportion of the test completed as 0.75.

The standard error of estimate (SE) is the reciprocal of the square root of the test information function (TIF) based on the items each test taker takes. The SE is calculated based only on the answered item(s) for both complete and incomplete tests. The TIF is the sum of information from each item on the test (See Figures 5.3 and 5.4). With MLE estimation, the SE at the test taker's ability is:

$$SE\left(\theta_j\right) = \frac{1}{\sqrt{I\left(\theta_j\right)}},$$

where $\boldsymbol{I}(\boldsymbol{\theta_j})$ is the test information for test taker j, calculated as:

$$I\left(\theta_j\right) = \sum_{i=1}^{n} I_i(\theta_j)$$

and

$$I_i\left(\theta_j\right) = \frac{\left[P_i'\left(\theta_j\right)\right]^2}{P_i\left(\theta_j\right)Q_i\left(\theta_j\right)},$$

where $P_i'(\theta_j)$ is the derivative of $P_i(\theta_j)$ with respect to θ_j, $P_i(\theta_j)$ is the item response function, and $Q_i\left(\theta_j\right) = 1 - P_i(\theta_j)$.

The standard error of estimate (SE) at the test taker's ability was also used as conditional standard error of measurement (CSEM) for reporting scores. The upper bound of CSEM is set to 2.5 in the theta metric. Any estimated SE values larger than 2.5 was reported as 2.5.

There were some special cases in which the score reported for a test taker was not based on the MLE of that test taker's ability:

- The test taker got the lowest possible score, which would lead to an estimated ability of $-\infty$.
- The test taker got the highest possible score, which would lead to an estimated ability of $+\infty$.

- The test taker's response pattern did not lead to a clear MLE of the test taker's ability.

If the test taker's response pattern did not lead to a clear MLE of the test taker's ability, the test taker's score was computed by the "inverse TCC method." The idea of the inverse TCC method is to transform the test taker's number-correct score into an ability estimate (Stocking, 1996). The transformation is the inverse of the test characteristic curve for the particular combination of items administered to the test taker. Although number-correct scores for test takers who take different items are not comparable, the ability estimates computed by this method are comparable.

The principal advantage of the inverse TCC method is that it does not require accurate estimates of the item characteristic curves (ICCs) of the individual items. Estimation errors in ICCs largely cancel out when the ICCs are summed to form a test characteristic curve. Therefore, ability estimates based on the inverse TCC is much more robust to item parameter estimation errors than ability estimates derived from maximum-likelihood estimation. However, the robustness comes at the expense of information loss by not utilizing each test taker's unique pattern of responses. As a result, this method was used only when the MLE could not be calculated.

Ability values for test takers with the highest or lowest possible scores were specified according to the following guidelines:

1. The highest obtainable theta (HOT) should be high enough so that it does not cause an unnecessary pileup of scale scores at the top of the scale. Likewise, the lowest obtainable theta (LOT) should be low enough so that it does not cause an unnecessary pileup of scale scores at the bottom part of the scale.

2. The HOT should be low enough so that the conditional standard error of measurement (CSEM) corresponding to the HOT is lower than 10 x minimum CSEM from all theta scores. The LOT should be high enough so that CSEM corresponding to the LOT is higher than 15 x minimum CSEM from all theta scores.

3. For multiple test levels placed on the same vertical scale, the HOT and LOT values should increase and transition smoothly over levels.

Table 5.2 provides some statistics describing the distribution of vertically scaled test taker ability estimates at each grade level. The minimum and maximum ability estimates at each grade level are the specified LOT and HOT values. The LOT and HOT values give the effective range of the ELA and Mathematics scales. The entire ELA scale ranges from –4.59 (the LOT for grade 3) to 3.04 (the HOT for grade 8). In Mathematics, the range was from –4.11 to 3.63. The properties of the vertical scale are consistent with the comments of Kolen (2011) that an acceptable vertical scale should display increasing mean scores from grade to grade, the amount of growth should be decelerating, and the within-grade variability should be increasing from grade to grade.

Figures 5.5 and 5.6 are box plots showing the ability distributions for ELA and Mathematics. Whiskers were extended to the 10th and 90th percentiles. The box plots show that both the median and quantiles tend to increase with grade level. No constraints were placed on the minimum and maximum theta estimates in the box plots. For additional information on the scaling methods,

readers can refer to Chapter 9 of the Smarter Balanced field test technical report, available at http://www.smarterbalanced.org/wordpress/wp-content/uploads/2015/06/chapter-9.pdf.

5.2: Deriving Scale Scores for the Badger Exam

The IRT vertical scale for the reported scores was formed by linking ability scales across grades using common items in adjacent grades. The vertical scale score was a linear transformation of the vertically scaled IRT ability estimate. Both reported total test scores and claim scores were computed as linear transformations of the MLE (or inverse TCC) scores and truncated to the lowest obtainable scale score (LOSS) and highest obtainable scale score (HOSS), which were defined by the LOT and HOT. The same linear transformation is used for all scores within a subject.

$$Scale\ Score = a * IRT\ Ability\ Score\ + b$$

The scaling constants *a* and *b* were provided by Smarter Balanced. The calculated standard error of estimate was transformed to the reported score metric. Table 5.3 lists the scaling constants for each subject for the theta-to-scaled score linear transformation. Scale scores were rounded to an integer. Detailed information regarding scale score establishment can be found in the following documents:

- Chapter 10 (Achievement Level Setting) of Smarter Balanced field test technical report http://www.smarterbalanced.org/wordpress/wp-content/uploads/2011/12/Chapter-10-Achievement-Level-Setting-121014_mm.pdf
- Smarter Balanced scoring specification document developed by American Institute of Research (AIR) http://www.smarterapp.org/deployment/TestScoringSpecs2014-2015.html

5.3: Score Verification Procedure

To provide DPI with the highest level of accurate test results, ETS conducted a thorough evaluation of all scored data. File formats and data elements were validated against client-approved layouts, specifications, and processing requirements. Detailed test scripts were executed to confirm accuracy. Some of the quality verification steps included:

- verification of answer key application
- theta-to-scale score conversion application
- scale-score comparisons to achievement levels
- processing rules for individual test taker and summary level data

The quality-assurance steps involved processing sample test taker records through the data processing and scoring system. To reduce the risk of human error, ETS' Quality Assurance

programmatic test routines written in SAS were used to thoroughly evaluate each test taker's data record.

5.4: Samples Used for the Analyses

In accordance with AERA, APA, & NCME (2014) *Standards* 7.2 and 7.5, this Section documents the target population for which a test is intended and the relevant characteristics of the individuals who participated in data collection. All valid test records were included for technical manual analyses, but test records that included total test scale scores of "not scored" were excluded. Table 5.4 provides the numbers of districts, schools, and test takers included in the analyses by grade and subject. For validity review (Chapter 8), either ELA or Mathematics claim scale scores identified as "not scored" were not included in the inter-claim correlation analysis.

Table 5.1: Summary of Pre-Equating IRT Item Parameter Estimates

Content	Grade	Number of items	Difficulty Parameter (b)				Discrimination Parameter (a)			
			Min	Max	Mean	*SD*	Min	Max	Mean	*SD*
ELA	3	45	-2.37	1.54	-0.47	0.93	0.24	1.10	0.71	0.23
	4	47	-1.92	2.94	0.30	1.17	0.15	1.17	0.59	0.25
	5	45	-1.49	2.87	0.21	1.01	0.24	1.13	0.67	0.20
	6	48	-0.89	3.62	0.87	1.04	0.19	0.94	0.50	0.19
	7	47	-1.84	3.51	0.76	1.25	0.19	0.93	0.52	0.18
	8	47	-1.54	5.38	1.27	1.29	0.18	1.11	0.49	0.21
Mathematics	3	42	-3.15	0.75	-1.17	0.92	0.13	1.43	0.93	0.28
	4	42	-3.26	2.24	-0.26	1.16	0.26	1.55	0.88	0.31
	5	43	-1.89	2.18	0.45	0.99	0.22	1.39	0.84	0.32
	6	40	-3.93	2.72	0.72	1.27	0.21	1.34	0.77	0.26
	7	40	-1.79	4.45	1.39	1.33	0.19	1.68	0.85	0.33
	8	39	-0.77	4.73	1.92	1.14	0.13	1.36	0.62	0.25

Table 5.2: Summary of Vertically Scaled Test Taker Proficiency Estimates

Content	Grade	Number of Test Takers	Min (LOT)	Max (HOT)	Mean	*SD*	Effect Size
ELA	3	61,525	-4.59	1.34	-0.87	0.97	
	4	61,121	-4.40	1.80	-0.46	1.05	0.41
	5	61,600	-3.58	2.25	0.04	1.06	0.47
	6	60,859	-3.48	2.51	0.13	1.12	0.08
	7	60,590	-2.91	2.75	0.53	1.14	0.35
	8	60,842	-2.57	3.04	0.72	1.13	0.17
Mathematics	3	61,571	-4.11	1.33	-0.99	0.93	
	4	61,148	-3.92	1.82	-0.50	1.01	0.50
	5	61,597	-3.73	2.33	-0.14	1.13	0.34
	6	60,690	-3.53	2.95	0.13	1.24	0.23
	7	60,368	-3.34	3.32	0.37	1.32	0.19
	8	60,664	-3.15	3.63	0.57	1.34	0.15

Note. LOT: Lowest obtainable theta, HOT: Highest obtainable theta, Effect size was calculated by Yen (1986).

Table 5.3: Scaling Constants for the Reporting Metric

Content	Grade	Slope (a)	Intercept (b)
ELA	3–8	85.8	2508.2
Mathematics	3–8	79.3	2514.9

Table 5.4: Summary of Sampled Data

Content	Grade	Number of		
		Districts	Schools	Test Takers
ELA	3	530	1,222	61,525
	4	529	1,208	61,121
	5	526	1,155	61,600
	6	524	806	60,859
	7	522	762	60,590
	8	517	752	60,842
Mathematics	3	529	1,221	61,571
	4	528	1,207	61,148
	5	524	1,153	61,597
	6	522	802	60,690
	7	521	761	60,368
	8	516	750	60,664

Figure 5.1: Test Characteristic Curves, ELA

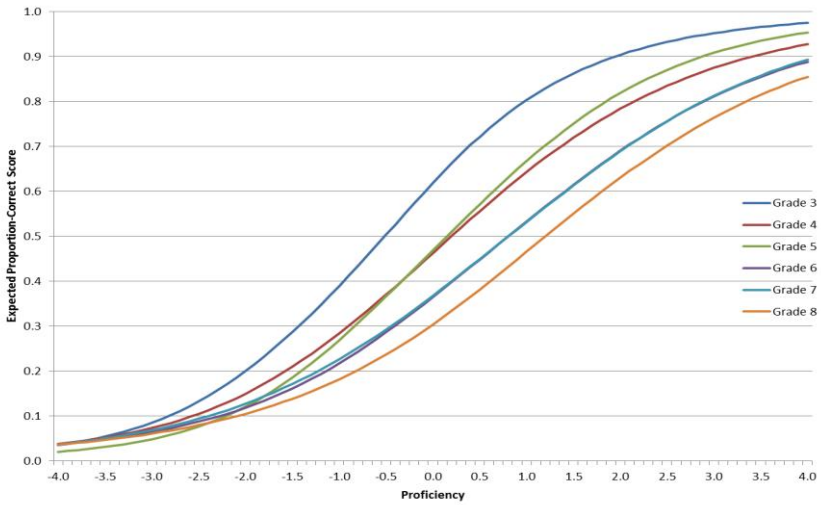
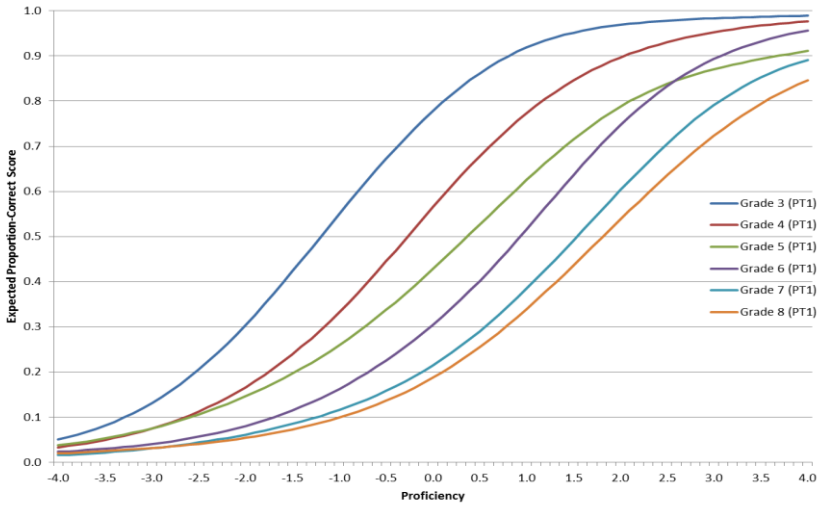


Figure 5.2: Test Characteristic Curves, Mathematics



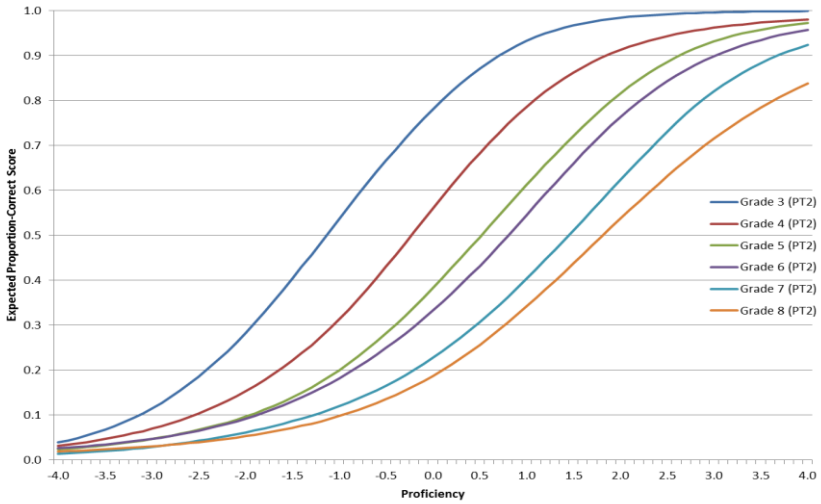


Figure 5.3: Test Information Functions, ELA

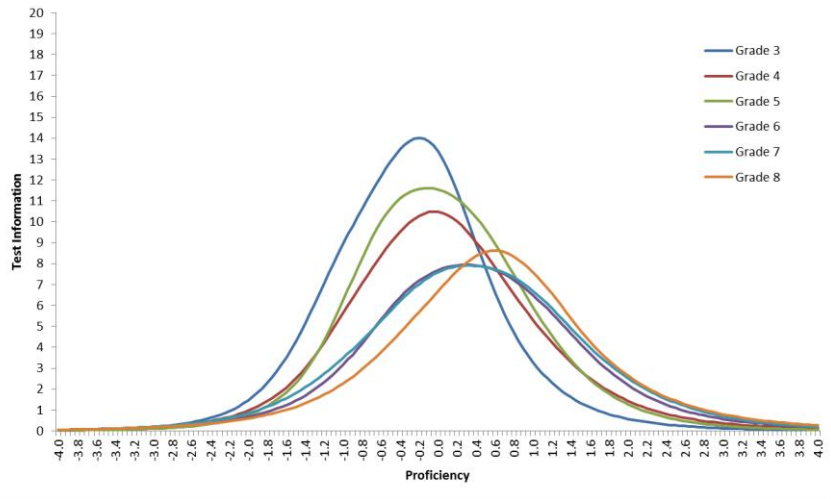
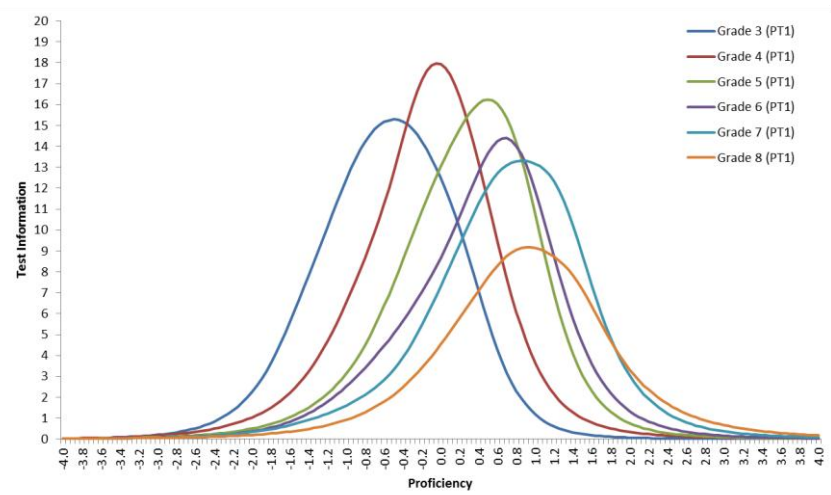


Figure 5.4: Test Information Functions, Mathematics



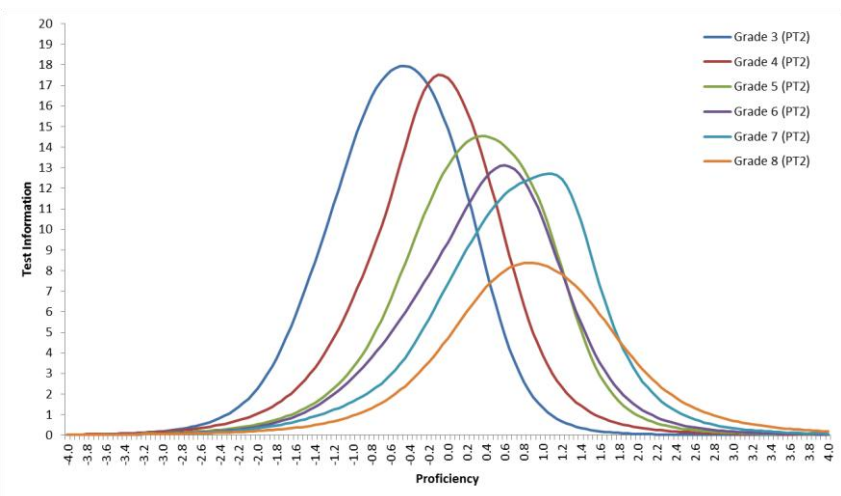


Figure 5.5: Test Taker Proficiency Box Plots, ELA

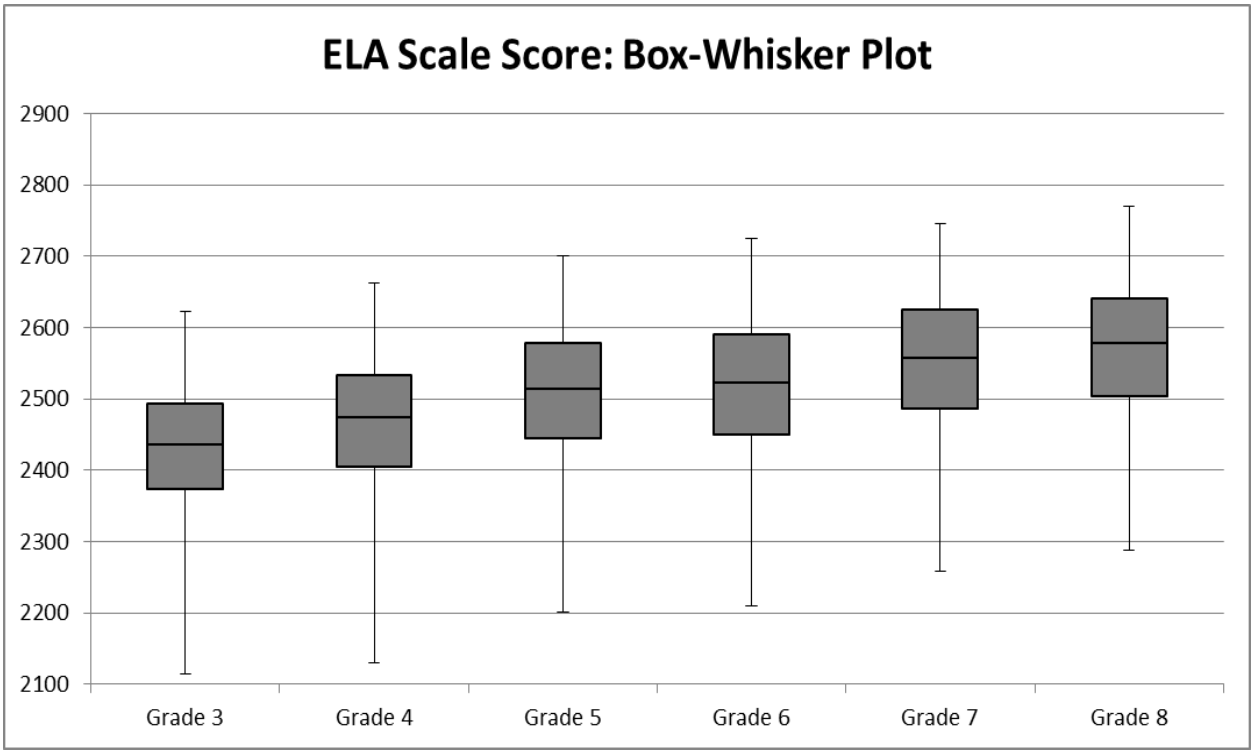
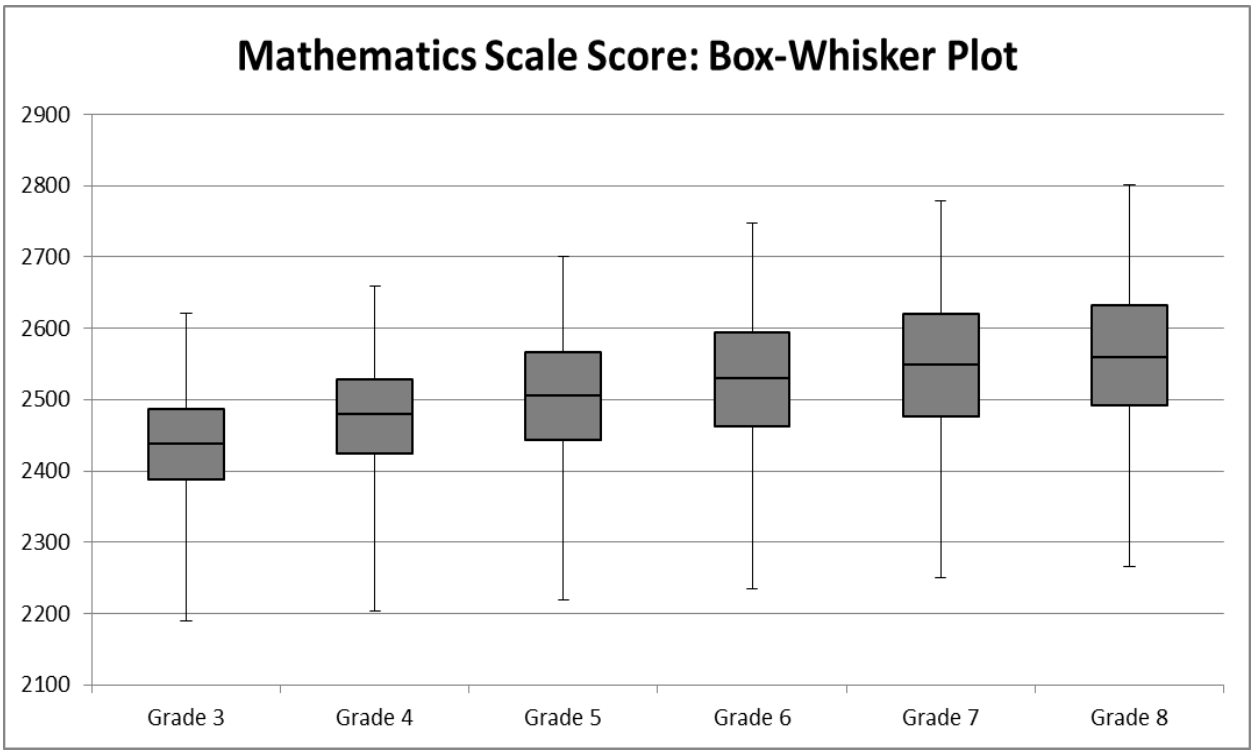


Figure 5.6: Test Taker Proficiency Box Plots, Mathematics



CHAPTER 6: TEST RESULTS

6.1: Classical Item Analyses

6.2: Interpreting Test Scores

- 6.2.1: Raw Score Results
- 6.2.2: Scale Score Results
- 6.2.3: Performance Level Results
- 6.2.4: Claim Level Results (Strength and Weakness)

6.1: Classical Item Analyses

Classical item analyses involve computing, for every item on each test, a set of statistics. Each statistic is designed to provide some key information about the quality of the item from an empirical perspective. For the Badger Exam, these statistics were the item difficulty, item discrimination, distractor analysis, and omit rate.

Item Difficulty ("p-value")
This statistic is the mean score on the item, expressed as a proportion of the maximum possible score on the item. For dichotomous items, it is equivalent to the proportion of test takers who answered the item correctly. For polytomous items scored with more than two ordered categories, item difficulty is defined as the mean score for the item divided by the maximum score point.

Item Discrimination ("item-total correlation")
Item discrimination is measured by the correlation between the item score and a criterion measure of ability, typically the total test score. The higher the correlation, the better the item is in providing information concerning test taker performance as measured by the entire test. For the Badger Exam, the criterion measure was the test taker's scale score. The item-total correlations for the Badger Exam are the point-biserial correlation for dichotomously scored items and the polyserial correlation for polytomously-scored items. Item-total correlation ranges from -1.0 to 1.0.

A large positive point-biserial correlation indicates a tendency for test takers with high scores on the overall test to choose the correct response to a dichotomously-scored item and test takers with low scores on the overall test to choose an incorrect response to the item. A large positive polyserial correlation indicates a tendency for test takers with low scores on the overall test to have low scores on a polytomously-scored item and those with high overall scores to have high scores on the item. A very small correlation, near zero, indicates no relationship between performance on the item and performance on the whole test. Negative correlations can indicate serious problems with the item content (e.g., multiple correct answers on a multiple-choice item) or indicate that test takers have not been taught the content.

Distractor Analysis ("Distractor-total correlation")
The quality of distractors is an important component of a multiple-choice item's overall quality. Distractors should be clearly incorrect, but at the same time plausible and attractive to lower-ability test takers. Several analyses are conducted and reported in the final item analysis (FIA) to evaluate the quality of distractors: the percentage of test takers at each response option, the average ability level measured by the total score for test takers selecting each response option, and distractor-total correlation analysis.

The distractor-total correlation statistic describes the relationship between selecting an incorrect response for a specific item and performance on the entire test. The point-biserial correlation is calculated for each response option in multiple-choice items. Typically, the magnitude of the correlation between an incorrect answer and total test performance is weak or negative. The values of this correlation are compared and contrasted with the item-total correlation. If the item-total correlation for an incorrect answer is stronger than for the correct answer, it may indicate

that the incorrect answer is actually correct, there may be more than one correct answer, or the item may have been keyed incorrectly.

Omit Rate

The percentage of test takers omitting an item can indicate a problem with the time limit for the test or with some feature of the item. If test takers have an adequate amount of testing time, at least 95% of the test takers should attempt to answer each item. The Badger Exams are designed as untimed tests for all test takers to respond to all of the items. Because there is no time limit for the test, an omit percentage greater than 5% for a single item could be an indication of an item difficulty problem.

Flagging an item is used to identify certain statistical characteristics of items that indicate poor functioning. Items with poor classical statistics were flagged and reviewed by ETS psychometricians. An item was flagged if it met any of the following conditions:

- low *p*-value: below 0.30
- high *p*-value: above 0.90
- low item-total correlation: below 0.20
- positive distractor-total correlation: above 0.00
- high percentage of omits: more than 5% of test takers omitted the item

Distributions and summary statistics of the *p*-value and item-total correlation statistics at each grade level are provided in Tables 6.1 (for ELA) and 6.2 (for Mathematics). Although several items were flagged as having low *p*-value, most of the items were well-distributed with item difficulty values between 0.30 and 0.90 to measure test taker ability for both ELA and Mathematics tests. Only two items were flagged for high *p*-values: one in ELA at grade 7 and one in Mathematics at grade 3. The average *p*-value for the tests ranged from 0.39 to 0.48 for ELA and from 0.31 to 0.53 for Mathematics. For item discrimination, four items in the Mathematics tests and 15 items in the ELA tests were flagged as having low item-total correlations. Overall, ELA items (average item-total correlations of 0.36–0.44) were less discriminating than Mathematics items (average item-total correlations of 0.43–0.49).

Table 6.3 shows the number of items in the spring 2015 Badger Exam operational tests that were flagged for each of four conditions: *p*-value, item-total correlation, distractor-total correlation, and omit rate. These numbers are shown separately by grade and subject. Because some items were flagged for more than one condition, the number of flags may be greater than the number of flagged items. A significant number of items were flagged for low *p*-value, indicating that they were difficult items. A total of 19 items in the ELA and Mathematics tests were flagged for low item-total correlation, and a total of 18 items were flagged for positive distractor-total correlations. Items that were flagged were reviewed and no serious issue was found during this review. Hence, none of the flagged items were excluded from scoring. There were no items flagged due to the omit rate.

Tables in Appendix A present more comprehensive results from the classical item analyses for all of the items administered in each grade and subject area. These tables provide item statistics and flags as well as item location information on test forms, item response types (multiple-choice single-select, multiple-choice multi-select, etc.), and test section types for Mathematics tests (Sections 1 and 2).

6.2: Interpreting Test Scores

The Badger Exam is a summative assessment that provides information about a test taker's achievement compared to other test takers in Wisconsin and to test takers in other states that use similar Smarter Balanced Assessments. The results from the Badger Exam provide insight into how well Wisconsin test takers are doing on a broad measure of achievement. The summary results cover three types of scores: raw scores, scale scores, and performance level by overall and claim scores (i.e., strength and weakness levels).

6.2.1: Raw Score Results

Raw scores were not used to compute scale scores because IRT pattern scoring was applied. However, descriptive statistics of raw scores are useful as a preliminary check on the comparability of test takers performance. Table 6.4 provides the descriptive statistics of raw score results based on all test takers who took the Badger Exam. Tables 6.5 through 6.11 present the same information for subgroups of interest broken down by grade and subject area.

These tables show the number of test takers taking each test form, mean, standard deviation (SD), skewness, and kurtosis of number-correct scores, mean of average item difficulty (i.e., test difficulty), test reliability as measured by Cronbach's alpha (Cronbach, Schönemann, & McKie, 1965) and standard error of measurement (SEM) for raw scores to facilitate interpretation of the raw score results. In addition, the minimum and maximum obtained scores and possible scores are compared across grades. For Mathematics tests, the raw score results were separated by test takers who took a different Section 2 (PT1 or PT2).

The results of raw score statistics, especially for mean raw score, are compared by grade and subject area. Mean raw scores from Mathematics tests were relatively lower than the scores from ELA because of different maximum possible raw score points between ELA and Mathematics. The maximum possible raw score points ranged from 47 to 51 in ELA, while it ranged from 37 to 41 in Mathematics. To standardize each mean raw score, the mean proportion correct was calculated as the mean divided by the maximum possible number of score points. Hence, this statistic ranges from 0.0 to 1.0. A larger mean proportion correct value (e.g., 0.56 in Mathematics PT2, grade 3) indicates higher performance on a test, while a smaller value (e.g., 0.30 in Mathematics PT1, grade 8) indicates lower performance. The mean proportion correct for ELA tests ranged from 0.39 to 0.46, but there was no increasing or decreasing pattern by grade. The mean proportion correct for Mathematics tests ranged from 0.30 to 0.56, with a generally increasing pattern as grade level increased.

The skewness and kurtosis statistics in Table 6.4 describes the shape of raw score distribution for each grade and subject area. The skewness measures the lack of symmetry of a distribution. A skewness value is zero when a distribution is perfectly symmetric (e.g., skewness of 0.0 for ELA, grade 5). A positive value for the skewness index indicates that the right tail of distribution is longer than the left tail. This implies that some test takers achieved raw scores far above the mean raw score, but many others achieved raw scores below, but not far below, the mean raw score (e.g., skewness of 0.8 in Mathematics PT1, grade 7). On the contrary, negative values for the skewness indicate that many test takers achieved raw scores above the mean raw score but some achieved scores far below the mean (e.g., skewness of -0.2 in Mathematics PT2, grade 3). Kurtosis measures the tendency of a distribution to be more peaked or flatter than the normal

distribution. Positive values for the kurtosis indicate that more test takers achieved raw scores in the center of the score distribution than a normal curve, while negative values indicate that distributions are flatter than a normal curve with fewer test takers near the center of the score distribution (e.g., kurtosis of -0.9 in ELA, grade 5). Overall, the score distributions from Badger Exam test takers were slightly skewed in either the positive or negative direction along the test scale in each grade and subject area. The ranges of skewness and kurtosis in ELA tests were similar across grades, but the skewness and kurtosis increased positively in Mathematics tests as the grade level increased.

Cronbach's alpha and SEM are also provided in Table 6.4 to demonstrate the reliability of the Badger Exam. Note that these statistics are affected by test length (i.e., maximum possible raw score points). Detailed descriptions of these statistics are provided further in Chapter 7 (Reliability). Overall the test reliabilities were similar across grades and relatively high for both ELA (0.87 to 0.91) and Mathematics (0.89 to 0.91). The SEM of the raw scores ranged from 2.85 to 3.09 for ELA tests and from 2.21 to 2.58 for the Mathematics tests. Note that SEMs cannot be directly comparable due to different test length. Even though the ELA tests are longer than the Mathematics tests, the overall alpha was higher in the Mathematics tests.

Overall, the raw score results by subgroups demonstrated consistent performance patterns (See Tables 6.5 through 6.11) in all grades and both subject areas. The following subgroups were defined and raw score statistics were calculated by each category of defined subgroups: gender, race/ethnicity, socioeconomic status, disability status, and English language proficiency.

Gender

In ELA, the female test takers had slightly higher mean raw scores and mean proportions correct than those of the male test takers. In Mathematics, however, the mean raw score differences and mean proportions correct between female and male groups were quite small, which implies that there was no gender difference on Mathematics tests.

Race/Ethnicity

In both ELA and Mathematics, the White test taker group, as the majority population, had the highest mean raw scores and mean proportions correct, followed by Asian, Pacific Islander, Two or more races, Hispanic, American Indian, and African American test taker groups. Because the numbers of test takers in the Pacific Islander, Two or more races, and American Indian ethnic groups were very small, the raw score variation should be interpreted with caution.

Economically Disadvantaged, Disability, and English Language Proficiency

The mean raw scores of reference groups (i.e., not economically disadvantaged, not disabled, and fully English proficient test taker groups) were always higher than those of focal groups (i.e., economically disadvantaged, disabled, and limited English proficient test taker groups). This pattern was consistent in all grades and both subject areas.

6.2.2: Scale Score Results

The Badger Exam ELA and Mathematics tests are vertically scaled and hence comparable across grade levels for grades 3 through 8. The scale scores range from 2114 to 2769 for ELA and range from 2189 to 2802 for Mathematics. The intents of a vertical scaling are to provide an

assessment system such that the claims and abilities captured by each grade level test (within subject area) reflect the same fundamental set of measured constructs and to facilitate the comparability of scores at different grade levels. In essence, each Badger Exam vertical scale reflects a single general underlying construct (i.e., overall mathematics ability from Mathematics tests, grades 3 through 8).

While vertical scaling is common practice in educational assessment, there are limits to the interpretations based on such scales (Kolen & Brennan, 2014). Where each grade-level test is based on a common blueprint design, the grade-level specifics as reflected in the test items differ from grade to grade. These differences are naturally greater as one compares over wider grade spans. It is thus important to take these underlying factors into consideration when interpreting test taker performance across grade levels, remembering that the scales for adjacent grades are linked rather than equated. The basis for vertical scaling is that there is substantial overlap from one grade to the next of the skills taught in ELA and Mathematics. For grades two or more levels apart, however, the overlap is not as great as for adjacent grade levels. Comparisons across adjacent grades are the most valid.

Test takers' total test scores and claim scores are scale scores derived from the IRT models and item-pattern scoring procedures. Overall, higher scale scores indicate higher levels of test taker performance, and lower scale scores indicate lower levels of test taker performance. Table 6.12 provides the summary of scale score statistics based on total numbers of test takers. This table includes the mean, SD, skewness, and kurtosis of the scale scores, the minimum and maximum observed scale scores, and LOSS and HOSS for all subject areas and grades, based on the total numbers of test takers.

In ELA, the mean scale score increased by grade level, ranging from 2433.4 to 2569.8. There was only a small increase between grades 5 and 6; the mean scale score difference was 8.1 points. Standard deviations in ELA tests ranged from 82.9 to 96.7. Even though there was no consistent increase from one grade to the next in the standard deviation, grades 3 to 5 had smaller SDs than grades 6 to 8. All scale score distributions from ELA tests were slightly negatively skewed and the values of kurtosis were negative. In each grade level, scale scores spanned the full scale score range from the LOSS to the HOSS.

In Mathematics, the mean scale score increased consistently from one grade level to the next (ranging from 2436.4 to 2559.9), but by a decreasing amount. Score increases were smaller at higher grade levels, ranging from 38.6 (between grades 3 and 4) to 15.7 (between grades 7 and 8). Standard deviations on Mathematics tests ranged from 73.8 to 106.4, and there was a clear increasing pattern by grade levels. All scale scores from Mathematics tests were slightly negatively skewed and the values of kurtosis were close to zero except for grade 4. In each grade level, scale scores spanned the full scale score range from the LOSS to the HOSS.

A maximum likelihood procedure cannot produce accurate ability estimates for some test takers, including those with zero scores or perfect scores. Scores for these test takers are based on an alternative procedure described in Chapter 5 (Score Calibration and Scaling). The numbers and the percentages of test takers who scored at the lowest possible scale score and at the highest, in each grade and subject area, are shown in Table 6.12. Test takers more often scored at the lowest possible score on Mathematics tests (0.14% to 1.25%) than on ELA tests (0.04% to 0.40%), while test takers scored at the highest possible score more often on ELA tests (0.84% to 2.09%) than on Mathematics tests (0.66% to 1.22%).

The results for gender, race/ethnicity, socioeconomic status, disability status, and English language proficiency are also displayed from Tables 6.13 through 6.18. Caution must be exercised when interpreting statistics for some subgroups (e.g., Pacific Islander group) since they are based on a relatively small number of cases. Overall, the scale score results by subgroups demonstrated consistent performance patterns in all grades and both subject areas the same as raw score results.

6.2.3: Performance Level Results

Overall scale scores for the Badger Exams are mapped into four performance levels per grade and subject area. After the standard setting process conducted by Smarter Balanced in 2014, the performance level categories were defined as "Below Basic," "Basic," "Proficient," and "Advanced." The combined category of "Proficient" or "Advanced" was used to define the proficiency criterion for accountability purposes. The achievement level descriptors were developed based on the feedback of reviewers who engaged in a validation process based on examining the CCSSs in each subject area and the items on the examination.

- Advanced – Test taker demonstrates thorough understanding of ELA and Mathematics and ability to apply the knowledge and skills for their grade level that are associated with college content-readiness.
- Proficient – Test taker demonstrates adequate understanding of ELA and Mathematics and ability to apply the knowledge and skills for their grade level that are associated with college content-readiness.
- Basic – Test taker demonstrates partial understanding of ELA and Mathematics and ability to apply the knowledge and skills for their grade level that are associated with college content-readiness.
- Below Basic – Test taker demonstrates minimal understanding of ELA and Mathematics and ability to apply the knowledge and skills for their grade level that are associated with college content-readiness.

For details of the standard setting and achievement level descriptor writing process, please refer to Chapter 10 (Achievement Level Setting) of the Smarter Balanced field test technical report (http://www.smarterbalanced.org/wordpress/wp-content/uploads/2011/12/Chapter-10-Achievement-Level-Setting-121014_mm.pdf) and related Smarter Balanced website located at http://www.smarterbalanced.org/achievement-levels/.

Figures 6.1 and 6.2 provide the score range of each performance level by grade and subject area, as well as the percentage of test takers in each performance level from the total test taker group in a graphical way. Overall, in ELA, the range of percentages of test takers who achieved "Proficient" or "Advanced" levels was from 46.7% (grade 6) to 55.0% (grade 5). In Mathematics, the range of percentages of test takers who achieved "Proficient" or "Advanced" levels was lower than ELA, ranging from 40.3 (grade 8) to 51.8 (grade 3).

Tables 6.19 through 6.30 show the percentages of all test takers in each performance category, as well as subgroup comparisons by gender, race/ethnicity, socioeconomic status, disability status, and English language proficiency for both ELA (6.19 to 6.24) and Mathematics (6.25 to 6.30).

6.2.4: Claim Level Results (Strength and Weakness)

Test takers received performance classifications not only for total scale score but also claim scores. As described in Chapter 2 (Test Design and Development), each claim score is based on only those items attached to a particular content knowledge or skill domains. Claim scores are intended to provide detailed information regarding strength or weakness in these specific domains. Because claim scores are each based on only a subset of items, they are less precise than total scores. The three strength and weakness levels for each claim score are as follows[5]:

- Exceeds standard – Test taker clearly understands and can successfully apply their knowledge to the standards tested in this subject area for their grade.
- Meets standard – Test taker shows understanding and can apply their knowledge to the standards tested in this subject area for their grade.
- Does not meet standard – Test taker has limited understanding and difficulty applying their knowledge to the standards tested in this subject area for their grade.

As described in Chapter 2 (Table 2.1), there are four claims in each subject area. Most claim scores are based on fewer than ten items except for Claim 1. The Smarter Balanced Assessment Consortium therefore decided to report claim scores conservatively, meaning that most test takers would be classified into the central reporting category (labeled for the Badger Exam as "Meets standard") unless strong evidence exists that the test taker in fact belongs in the lower or upper performance categories (labeled as "Does not meet standard" and "Exceeds standard"). The "Meets standard" classification is essentially declared as the "default" classification unless a test taker's pattern of performance is distinct enough to warrant reporting in the "Does not meet standard" or "Exceeds standard" category. The rationale behind this scoring system is based on expectations of how scores might be used. Because the classification "Does not meet standard" in a claim area presumably demands attention in the form of some sort of instructional intervention, the decision was made by Smarter Balanced Assessment Consortium to report such a classification only when there is a high level of certainty that performance in the claim area is indeed weak. Technically, the designation of "Meets standard" as the default means that the reported performance category may not be the one into which a test taker's numeric claim score actually falls.

Each test taker's strength and weakness level in a particular claim is based on the test taker's estimated ability for that claim, expressed on the IRT (theta) scale. The strength and weakness levels are defined by two standards, expressed as points on this scale. However, to be classified "Exceeds standard," the test taker's estimated ability must be above the upper standard by at least 1.5 times the SEM of the ability estimate for a test taker whose ability is at the upper standard. Similarly, to be classified "Does not meet standard," the test taker's estimated ability must be below the lower standard by at least 1.5 times the SEM of the ability estimate for a test taker whose ability is at the lower standard. Therefore, many test takers are classified as "Meets standard" even though their ability estimates would imply a higher or lower classification.

[5] Note that the original descriptors of strength and weakness level defined by Smarter Balanced were "Below standard," "At/Near standard," and "Above standard." See http://www.smarterbalanced.org/wordpress/wp-content/uploads/2015/04/indivStudentReport1.pdf

The percentage of test takers in each strength and weakness level from both ELA and Mathematics tests are provided in Tables 6.31 and 6.32, respectively. The results are summarized by subject area below.

ELA

- The percentages of test takers who achieved the "Exceeds standard" level on Claim 1 were highest compared to other claims, ranging from 23.1% (Claim 1, grade 6) to 38.0% (Claim 1, grade 5).
- The percentages of test takers who achieved the "Exceeds standard" level on other claims ranged from 9.4% (Claim 4, grade 6) to 24.7% (Claim 2, grade 7).
- The percentages of test takers who were classified at the "Does not meet standard" level on Claim 1 were, for the most part, higher than those of other claims, ranging from 21.3% (Claim 1, grade 7) to 31.8% (Claim 1, grade 6).
- The percentages of test takers who were classified at the "Does not meet standard" level on the other claims ranged from 3.2% (Claim 2, grade 6) to 26.1% (Claim 3, grade 4).
- The percentages of test takers who achieved the "Meets standard" level on Claim 1 were lower than those of other claims, which is an opposite pattern compared to the results from the "Exceeds standard" and "Does not meet standard" levels. For example, in some grades/claims, the percentages of "Meet standard" were higher than 70% and the highest percentage was 77.9% in Claim 2, grade 6.

Mathematics

- The percentages of test takers who achieved the "Exceeds standard" level on three claims were similar to each other, ranging from 15.2% (Claim 3, grade 8) to 29.3% (Claim 2, grade 3).
- The percentages of test takers who achieved the "Does not meet standard" level on Claim 1 were highest compared to other claims, ranging from 30.9% (Claim 1, grade 3) to 40.0% (Claim 1, grade 7).
- The percentages of test takers who achieved the "Does not meet standard" level on other claims ranged from 11.8% (Claim 3, grade 4) to 35.4% (Claim 3, grade 7).
- The percentages of test takers who achieved the "Meets standard" level on Claim 1 were lower than those of other claims, which is an opposite pattern compared to the results from the "Does not meet standard" level. Note that in some grades/claims, the percentages were higher than 60% and highest percentage was 64.3% in Claim 3, grade 8.

Table 6.1: Summary of ELA Classical Item Statistics

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Number of Items	45	47	45	48	47	47
p-value						
≥ 0.90	0	0	0	0	1	0
0.80 – 0.89	1	2	4	0	1	1
0.70 – 0.79	4	1	4	1	3	3
0.60 – 0.69	5	3	6	4	5	8
0.50 – 0.59	6	6	5	8	12	3
0.40 – 0.49	7	11	8	12	10	8
0.30 – 0.39	12	7	10	7	5	15
< 0.30	10	17	8	16	10	9
Mean	0.43	0.39	0.48	0.40	0.47	0.43
Median	0.40	0.39	0.44	0.42	0.47	0.39
SD	0.19	0.19	0.20	0.17	0.20	0.18
Item-Total Correlation						
≥ 0.60	1	2	2	0	0	1
0.50 – 0.59	11	9	7	6	5	7
0.40 – 0.49	17	7	23	17	15	8
0.30 – 0.39	11	15	10	8	15	19
0.20 – 0.29	5	8	3	13	11	8
< 0.20	0	6	0	4	1	4
Mean	0.42	0.38	0.44	0.37	0.38	0.36
Median	0.43	0.37	0.46	0.38	0.38	0.36
SD	0.10	0.13	0.09	0.12	0.10	0.12

Table 6.2: Summary of Mathematics Classical Item Statistics

	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Number of Items	41	42	43	40	40	39
p-value						
≥ 0.90	1	0	0	0	0	0
0.80 – 0.89	2	6	2	2	1	0
0.70 – 0.79	6	1	0	2	3	1
0.60 – 0.69	7	2	4	2	2	1
0.50 – 0.59	8	7	6	6	2	2
0.40 – 0.49	5	3	8	4	3	8
0.30 – 0.39	5	11	4	4	10	8
< 0.30	7	12	19	20	19	19
Mean	0.53	0.45	0.38	0.37	0.33	0.31
Median	0.54	0.38	0.37	0.30	0.32	0.30
SD	0.22	0.23	0.20	0.21	0.22	0.16
Item-Total Correlation						
≥ 0.60	2	5	6	5	7	4
0.50 – 0.59	18	14	14	17	8	8
0.40 – 0.49	14	13	16	9	15	12
0.30 – 0.39	6	6	5	7	5	10
0.20 – 0.29	0	4	2	2	4	3
< 0.20	1	0	0	0	1	2
Mean	0.49	0.48	0.48	0.49	0.47	0.43
Median	0.50	0.48	0.48	0.51	0.47	0.44
SD	0.10	0.11	0.11	0.11	0.13	0.13

Note. One grade 3 item was treated as full credit for all attempted responses due to a Spanish translation issue.

Table 6.3: Summary of Flagged items by Classical Item Statistics

Content	Grade	Number of Flags	Reason for Flagging			
			p-value < 0.30	Item-Total Correlation < 0.20	Distractor Correlation > 0.00	Omit > 5%
ELA	3	12	10	0	2	0
	4	25	17	6	2	0
	5	8	8	0	0	0
	6	26	16	4	6	0
	7	13	10	1	2	0
	8	16	9	4	3	0
	Total	100	70	15	15	0
Mathematics	3	9	7	1	1	0
	4	12	12	0	0	0
	5	19	19	0	0	0
	6	20	20	0	0	0
	7	20	19	1	0	0
	8	23	19	2	2	0
	Total	103	96	4	3	0

Note. An item can be flagged for more than one reason.

Table 6.4: Raw Score Descriptive Statistics by Total

Content	Grade	*N*	Mean	*SD*	Skewness	Kurtosis	Min Obtained	Max Obtained	Max Possible	Mean Prop. Correct	Alpha	*SEM*
ELA	3	61,525	19.8	9.0	0.2	-0.8	0	45	46	0.43	0.90	2.85
	4	61,121	18.9	8.4	0.2	-0.7	0	44	49	0.39	0.88	2.96
	5	61,600	22.5	9.4	0.0	-0.9	0	46	47	0.48	0.91	2.88
	6	60,859	19.9	8.6	0.2	-0.7	0	47	50	0.40	0.87	3.07
	7	60,590	23.0	8.7	0.0	-0.7	0	48	50	0.46	0.88	3.05
	8	60,842	21.2	8.6	0.1	-0.7	0	47	50	0.42	0.87	3.09
Mathematics (PT1)	3	30,484	19.5	7.9	-0.1	-0.8	0	37	37	0.53	0.90	2.45
	4	30,053	17.0	8.3	0.3	-0.7	0	39	39	0.44	0.90	2.56
	5	30,155	14.2	8.0	0.5	-0.5	0	37	41	0.35	0.91	2.38
	6	30,167	13.9	8.2	0.6	-0.5	0	38	38	0.37	0.91	2.43
	7	30,386	12.6	7.8	0.7	-0.3	0	39	39	0.32	0.91	2.34
	8	28,275	11.2	7.1	0.8	0.0	0	36	37	0.30	0.89	2.37
Mathematics (PT2)	3	31,087	23.0	8.7	-0.2	-0.8	1	41	41	0.56	0.91	2.57
	4	31,095	17.6	8.4	0.2	-0.7	0	39	39	0.45	0.91	2.54
	5	31,442	15.7	8.5	0.5	-0.5	0	41	41	0.38	0.91	2.58
	6	30,523	15.3	8.4	0.4	-0.6	0	39	39	0.39	0.91	2.52
	7	29,982	11.7	7.2	0.7	-0.3	0	36	36	0.33	0.91	2.21
	8	32,389	11.4	6.9	0.7	-0.2	0	35	38	0.30	0.89	2.33

Table 6.5: Raw Score Descriptive Statistics by Gender

Content	Grade	Male						Female					
		N	%	Mean	*SD*	Mean Prop. Correct	Alpha	*N*	%	Mean	*SD*	Mean Prop. Correct	Alpha
ELA	3	31,437	51.1	18.8	8.9	0.41	0.90	30,088	48.9	20.9	9.0	0.45	0.90
	4	31,180	51.0	17.9	8.2	0.36	0.87	29,941	49.0	20.1	8.5	0.41	0.88
	5	31,556	51.2	21.4	9.4	0.45	0.91	30,044	48.8	23.8	9.3	0.51	0.90
	6	31,125	51.1	18.7	8.3	0.37	0.87	29,734	48.9	21.2	8.6	0.42	0.87
	7	30,829	50.9	21.5	8.6	0.43	0.88	29,761	49.1	24.6	8.5	0.49	0.87
	8	31,203	51.3	19.8	8.6	0.40	0.87	29,639	48.7	22.6	8.4	0.45	0.86
Mathematics (PT1)	3	15,625	25.4	19.6	8.0	0.53	0.91	14,859	24.1	19.3	7.9	0.52	0.90
	4	15,298	25.0	17.2	8.5	0.44	0.91	14,755	24.1	16.8	8.1	0.43	0.90
	5	15,389	25.0	14.2	8.3	0.35	0.92	14,766	24.0	14.2	7.7	0.35	0.90
	6	15,461	25.5	13.8	8.3	0.36	0.92	14,706	24.2	14.0	8.0	0.37	0.91
	7	15,552	25.8	12.5	8.0	0.32	0.92	14,834	24.6	12.6	7.6	0.32	0.91
	8	14,497	23.9	11.1	7.2	0.30	0.89	13,778	22.7	11.3	7.0	0.31	0.88
Mathematics (PT2)	3	15,845	25.7	23.2	8.8	0.57	0.92	15,242	24.8	22.7	8.6	0.55	0.91
	4	15,893	26.0	18.0	8.6	0.46	0.91	15,202	24.9	17.3	8.1	0.44	0.90
	5	16,160	26.2	15.7	8.8	0.38	0.91	15,282	24.8	15.7	8.2	0.38	0.90
	6	15,575	25.7	15.2	8.6	0.39	0.91	14,948	24.6	15.5	8.3	0.40	0.91
	7	15,158	25.1	11.7	7.4	0.33	0.91	14,824	24.6	11.8	7.0	0.33	0.90
	8	16,601	27.4	11.4	7.1	0.30	0.89	15,788	26.0	11.4	6.7	0.30	0.88

Table 6.6: Raw Score Descriptive Statistics for ELA by Race/Ethnicity

Content	Grade	Race/Ethnicity	*N*	%	Mean	*SD*	Mean Prop. Correct	Alpha
ELA	3	White	41,391	67.3	21.8	8.7	0.47	0.89
		African American	6,899	11.2	12.9	7.1	0.28	0.86
		Hispanic	8,003	13.0	16.4	8.0	0.36	0.88
		Asian	2,371	3.9	19.2	9.0	0.42	0.90
		American Indian	838	1.4	16.1	7.7	0.35	0.87
		Pacific Islander	47	0.1	19.8	8.0	0.43	0.87
		Two or More Races	1,976	3.2	19.4	9.0	0.42	0.90
	4	White	42,143	69.0	20.6	8.1	0.42	0.86
		African American	6,683	10.9	12.6	7.0	0.26	0.85
		Hispanic	7,362	12.0	15.8	7.5	0.32	0.85
		Asian	2,329	3.8	19.6	8.6	0.40	0.88
		American Indian	835	1.4	15.4	7.2	0.32	0.84
		Pacific Islander	65	0.1	18.9	7.8	0.39	0.85
		Two or More Races	1,704	2.8	18.2	8.2	0.37	0.87
	5	White	42,944	69.7	24.5	8.9	0.52	0.89
		African American	6,522	10.6	14.9	7.9	0.32	0.88
		Hispanic	7,304	11.9	18.4	8.6	0.39	0.89
		Asian	2,317	3.8	22.7	9.5	0.48	0.91
		American Indian	824	1.3	18.2	8.4	0.39	0.88
		Pacific Islander	40	0.1	23.2	9.7	0.49	0.91
		Two or More Races	1,649	2.7	21.9	9.5	0.47	0.91

Table 6.6: Raw Score Descriptive Statistics for ELA by Race/Ethnicity (Cont'd)

Content	Grade	Race/Ethnicity	*N*	%	Mean	*SD*	Mean Prop. Correct	Alpha
ELA	6	White	42,833	70.4	21.6	8.3	0.43	0.86
		African American	6,368	10.5	13.3	6.7	0.27	0.82
		Hispanic	7,131	11.7	16.3	7.6	0.33	0.84
		Asian	2,185	3.6	20.3	8.8	0.41	0.88
		American Indian	796	1.3	16.0	7.4	0.32	0.84
		Pacific Islander	47	0.1	19.9	8.3	0.40	0.86
		Two or More Races	1,499	2.5	19.3	8.5	0.39	0.87
	7	White	42,926	70.8	24.6	8.3	0.49	0.87
		African American	6,328	10.4	16.4	7.7	0.33	0.86
		Hispanic	6,886	11.4	19.9	8.1	0.40	0.86
		Asian	2,127	3.5	23.6	8.8	0.47	0.88
		American Indian	812	1.3	18.9	7.8	0.38	0.85
		Pacific Islander	37	0.1	22.8	6.9	0.46	0.80
		Two or More Races	1,474	2.4	22.0	8.4	0.44	0.87
	8	White	43,495	71.5	22.6	8.4	0.45	0.86
		African American	6,257	10.3	15.0	7.2	0.30	0.83
		Hispanic	6,753	11.1	18.0	7.7	0.36	0.84
		Asian	2,155	3.5	22.6	8.6	0.45	0.87
		American Indian	786	1.3	17.3	7.5	0.35	0.84
		Pacific Islander	38	0.1	22.0	8.5	0.44	0.86
		Two or More Races	1,358	2.2	20.4	8.6	0.41	0.87

Table 6.7: Raw Score Descriptive Statistics for Mathematics by Race/Ethnicity

Content	Grade	Race/Ethnicity	*N*	%	Mean	*SD*	Mean Prop. Correct	Alpha
Mathematics (PT1)	3	White	20,567	33.4	21.3	7.4	0.58	0.88
		African American	3,477	5.6	12.9	7.0	0.35	0.88
		Hispanic	3,958	6.4	16.1	7.2	0.43	0.88
		Asian	1,180	1.9	20.0	8.0	0.54	0.91
		American Indian	299	0.5	16.6	7.3	0.45	0.88
		Pacific Islander	18	0.0	17.8	7.7	0.48	0.90
		Two or More Races	985	1.6	18.3	7.9	0.49	0.90
	4	White	20,210	33.1	19.1	7.9	0.49	0.89
		African American	3,531	5.8	10.1	6.1	0.26	0.87
		Hispanic	3,811	6.2	12.8	6.9	0.33	0.88
		Asian	1,187	1.9	17.8	8.9	0.46	0.92
		American Indian	408	0.7	13.9	6.6	0.36	0.86
		Pacific Islander	32	0.1	18.4	8.7	0.47	0.91
		Two or More Races	874	1.4	16.1	8.1	0.41	0.90
	5	White	20,488	33.3	16.0	7.9	0.39	0.90
		African American	3,483	5.7	8.0	5.4	0.20	0.85
		Hispanic	3,827	6.2	10.5	6.3	0.26	0.87
		Asian	1,172	1.9	15.5	8.4	0.38	0.92
		American Indian	361	0.6	10.3	6.2	0.25	0.87
		Pacific Islander	24	0.0	14.7	7.0	0.36	0.87
		Two or More Races	800	1.3	13.5	8.1	0.33	0.92

Table 6.7: Raw Score Descriptive Statistics for Mathematics by Race/Ethnicity (Cont'd)

Content	Grade	Race/Ethnicity	*N*	%	Mean	*SD*	Mean Prop. Correct	Alpha
Mathematics (PT1)	6	White	21,807	35.9	15.5	8.1	0.41	0.90
		African American	2,807	4.6	7.1	4.9	0.19	0.83
		Hispanic	3,296	5.4	9.7	6.3	0.26	0.88
		Asian	1,101	1.8	14.3	8.6	0.38	0.92
		American Indian	392	0.6	9.3	5.8	0.25	0.86
		Pacific Islander	18	0.0	14.1	7.5	0.37	0.89
		Two or More Races	746	1.2	12.0	7.6	0.32	0.91
	7	White	21,740	36.0	14.1	7.8	0.36	0.90
		African American	3,307	5.5	6.3	4.7	0.16	0.84
		Hispanic	3,238	5.4	9.0	6.1	0.23	0.88
		Asian	1,008	1.7	14.3	9.0	0.37	0.93
		American Indian	354	0.6	9.3	6.3	0.24	0.88
		Pacific Islander	18	0.0	13.1	7.1	0.34	0.89
		Two or More Races	721	1.2	11.7	8.0	0.30	0.92
	8	White	19,779	32.6	12.4	7.1	0.34	0.88
		African American	3,159	5.2	6.0	4.2	0.16	0.78
		Hispanic	3,293	5.4	8.3	5.5	0.22	0.83
		Asian	1,124	1.9	13.1	7.8	0.35	0.90
		American Indian	314	0.5	8.7	5.4	0.24	0.83
		Pacific Islander	17	0.0	12.2	8.3	0.33	0.92
		Two or More Races	589	1.0	10.6	6.7	0.29	0.88

Table 6.7: Raw Score Descriptive Statistics for Mathematics by Race/Ethnicity (Cont'd)

Content	Grade	Race/Ethnicity	*N*	%	Mean	*SD*	Mean Prop. Correct	Alpha
Mathematics (PT2)	3	White	20,814	33.8	25.0	8.0	0.61	0.90
		African American	3,378	5.5	15.7	7.9	0.38	0.90
		Hispanic	4,104	6.7	19.3	8.1	0.47	0.90
		Asian	1,227	2.0	23.2	8.8	0.57	0.92
		American Indian	540	0.9	19.7	7.9	0.48	0.89
		Pacific Islander	31	0.1	21.5	8.9	0.53	0.92
		Two or More Races	993	1.6	22.0	8.8	0.54	0.91
	4	White	21,924	35.9	19.3	8.0	0.49	0.90
		African American	3,105	5.1	10.4	6.6	0.27	0.88
		Hispanic	3,613	5.9	14.1	7.3	0.36	0.89
		Asian	1,167	1.9	18.6	8.9	0.48	0.92
		American Indian	426	0.7	14.3	7.0	0.37	0.88
		Pacific Islander	33	0.1	15.3	8.0	0.39	0.91
		Two or More Races	827	1.4	17.0	8.3	0.44	0.91
	5	White	22,431	36.4	17.4	8.3	0.42	0.90
		African American	2,996	4.9	8.6	5.5	0.21	0.83
		Hispanic	3,521	5.7	11.8	6.9	0.29	0.87
		Asian	1,175	1.9	16.4	9.3	0.40	0.92
		American Indian	458	0.7	11.6	6.9	0.28	0.88
		Pacific Islander	16	0.0	16.8	8.3	0.41	0.90
		Two or More Races	845	1.4	14.3	8.0	0.35	0.90

Table 6.7: Raw Score Descriptive Statistics for Mathematics by Race/Ethnicity (Cont'd)

Content	Grade	Race/Ethnicity	*N*	%	Mean	*SD*	Mean Prop. Correct	Alpha
Mathematics (PT2)	6	White	20,980	34.6	17.2	8.1	0.44	0.90
		African American	3,478	5.7	8.3	5.7	0.21	0.85
		Hispanic	3,778	6.2	11.3	6.9	0.29	0.88
		Asian	1,111	1.8	16.9	9.2	0.43	0.93
		American Indian	397	0.7	10.6	6.4	0.27	0.86
		Pacific Islander	29	0.0	13.0	7.1	0.33	0.87
		Two or More Races	750	1.2	15.1	8.4	0.39	0.91
	7	White	21,130	35.0	13.1	7.2	0.36	0.90
		African American	2,901	4.8	6.4	4.6	0.18	0.84
		Hispanic	3,600	6.0	8.8	5.9	0.24	0.88
		Asian	1,123	1.9	12.8	7.8	0.35	0.92
		American Indian	458	0.8	7.5	5.4	0.21	0.87
		Pacific Islander	19	0.0	11.5	7.5	0.32	0.91
		Two or More Races	751	1.2	10.4	6.7	0.29	0.90
	8	White	23,661	39.0	12.6	6.8	0.33	0.88
		African American	2,978	4.9	6.0	4.3	0.16	0.79
		Hispanic	3,449	5.7	8.3	5.4	0.22	0.83
		Asian	1,054	1.7	13.0	8.0	0.34	0.91
		American Indian	470	0.8	7.2	4.8	0.19	0.81
		Pacific Islander	20	0.0	12.4	6.5	0.33	0.85
		Two or More Races	757	1.2	10.7	6.6	0.28	0.88

Table 6.8: Raw Score Descriptive Statistics by Socioeconomic Status

Content	Grade	Economically Disadvantaged						Not Economically Disadvantaged					
		N	%	Mean	*SD*	Mean Prop. Correct	Alpha	*N*	%	Mean	*SD*	Mean Prop. Correct	Alpha
ELA	3	29,352	47.7	16.3	8.1	0.35	0.88	32,173	52.3	23.0	8.6	0.50	0.89
	4	27,915	45.7	15.6	7.7	0.32	0.86	33,206	54.3	21.8	8.0	0.44	0.86
	5	27,420	44.5	18.5	8.8	0.39	0.89	34,180	55.5	25.8	8.7	0.55	0.89
	6	26,430	43.4	16.2	7.6	0.32	0.84	34,429	56.6	22.7	8.2	0.45	0.85
	7	25,970	42.9	19.4	8.1	0.39	0.86	34,620	57.1	25.7	8.2	0.51	0.86
	8	25,309	41.6	17.8	7.9	0.36	0.85	35,533	58.4	23.6	8.2	0.47	0.86
Mathematics (PT1)	3	14,576	23.7	16.3	7.4	0.44	0.89	15,908	25.8	22.3	7.3	0.60	0.89
	4	13,914	22.8	13.4	7.2	0.34	0.89	16,139	26.4	20.2	7.8	0.52	0.89
	5	13,784	22.4	10.8	6.6	0.26	0.88	16,371	26.6	17.1	8.0	0.42	0.90
	6	12,901	21.3	10.1	6.5	0.27	0.88	17,266	28.4	16.7	8.2	0.44	0.90
	7	12,584	20.8	9.1	6.3	0.23	0.88	17,802	29.5	15.1	7.9	0.39	0.90
	8	12,075	19.9	8.3	5.6	0.22	0.84	16,200	26.7	13.3	7.3	0.36	0.89
Mathematics (PT2)	3	14,798	24.0	19.5	8.3	0.47	0.90	16,289	26.5	26.1	7.8	0.64	0.89
	4	14,016	22.9	14.2	7.5	0.36	0.89	17,079	27.9	20.5	8.0	0.52	0.89
	5	13,622	22.1	12.0	7.1	0.29	0.88	17,820	28.9	18.5	8.4	0.45	0.90
	6	13,356	22.0	11.4	7.1	0.29	0.89	17,167	28.3	18.3	8.2	0.47	0.90
	7	13,192	21.9	8.8	5.9	0.24	0.88	16,790	27.8	14.1	7.3	0.39	0.90
	8	13,087	21.6	8.4	5.6	0.22	0.85	19,302	31.8	13.4	7.0	0.35	0.88

Table 6.9: Raw Score Descriptive Statistics by Disability

Content	Grade	Disabled						Not Disabled					
		N	%	Mean	*SD*	Mean Prop. Correct	Alpha	*N*	%	Mean	*SD*	Mean Prop. Correct	Alpha
ELA	3	7,161	11.6	14.7	8.1	0.32	0.88	54,364	88.4	20.5	8.9	0.45	0.90
	4	7,091	11.6	13.4	7.4	0.27	0.86	54,030	88.4	19.7	8.3	0.40	0.87
	5	7,439	12.1	14.6	8.3	0.31	0.89	54,161	87.9	23.6	9.0	0.50	0.90
	6	7,171	11.8	12.7	6.5	0.25	0.82	53,688	88.2	20.9	8.3	0.41	0.86
	7	7,178	11.8	14.6	7.0	0.29	0.84	53,412	88.2	24.2	8.3	0.48	0.86
	8	7,231	11.9	13.3	6.8	0.27	0.83	53,611	88.1	22.2	8.2	0.44	0.86
Mathematics (PT1)	3	3,530	5.7	14.6	8.1	0.39	0.91	26,954	43.8	20.1	7.7	0.54	0.90
	4	3,422	5.6	11.8	7.7	0.30	0.91	26,631	43.6	17.7	8.1	0.45	0.90
	5	3,642	5.9	8.8	6.4	0.21	0.89	26,513	43.0	15.0	7.9	0.36	0.91
	6	3,556	5.9	7.6	6.1	0.20	0.89	26,611	43.8	14.7	8.0	0.39	0.91
	7	3,626	6.0	6.3	5.3	0.16	0.87	26,760	44.3	13.4	7.7	0.34	0.91
	8	3,345	5.5	5.7	4.5	0.16	0.81	24,930	41.1	11.9	7.0	0.32	0.88
Mathematics (PT2)	3	3,621	5.9	17.5	9.0	0.43	0.92	27,466	44.6	23.7	8.4	0.58	0.91
	4	3,647	6.0	12.1	7.8	0.31	0.91	27,448	44.9	18.4	8.2	0.47	0.90
	5	3,767	6.1	9.8	7.0	0.24	0.89	27,675	44.9	16.5	8.4	0.40	0.90
	6	3,545	5.8	8.4	6.6	0.21	0.89	26,978	44.5	16.2	8.2	0.42	0.90
	7	3,497	5.8	5.9	4.9	0.16	0.86	26,485	43.9	12.5	7.1	0.35	0.90
	8	3,823	6.3	6.0	4.7	0.16	0.83	28,566	47.1	12.1	6.8	0.32	0.88

Table 6.10: Raw Score Descriptive Statistics by English Language Proficiency

Content	Grade	Limited English Proficient						Fully English Proficient					
		N	%	Mean	*SD*	Mean Prop. Correct	Alpha	*N*	%	Mean	*SD*	Mean Prop. Correct	Alpha
ELA	3	5,601	9.1	15.1	7.2	0.33	0.85	55,924	90.9	20.3	9.0	0.44	0.90
	4	4,521	7.4	13.7	6.5	0.28	0.81	56,600	92.6	19.4	8.4	0.39	0.88
	5	3,248	5.3	13.9	6.5	0.30	0.82	58,352	94.7	23.0	9.3	0.49	0.90
	6	2,609	4.3	12.1	5.4	0.24	0.73	58,250	95.7	20.3	8.5	0.40	0.87
	7	2,690	4.4	15.2	6.3	0.30	0.79	57,900	95.6	23.4	8.6	0.47	0.88
	8	2,870	4.7	14.4	5.9	0.29	0.75	57,972	95.3	21.5	8.6	0.43	0.87
Mathematics (PT1)	3	2,780	4.5	15.4	6.9	0.42	0.88	27,704	45.0	19.9	7.9	0.54	0.90
	4	2,300	3.8	11.3	6.1	0.29	0.86	27,753	45.4	17.5	8.3	0.45	0.90
	5	1,764	2.9	8.3	4.9	0.20	0.81	28,391	46.1	14.6	8.0	0.36	0.91
	6	1,284	2.1	6.8	4.5	0.18	0.81	28,883	47.6	14.2	8.2	0.37	0.91
	7	1,321	2.2	6.6	4.5	0.17	0.81	29,065	48.1	12.8	7.9	0.33	0.91
	8	1,530	2.5	6.5	4.4	0.18	0.78	26,745	44.1	11.5	7.1	0.31	0.89
Mathematics (PT2)	3	2,935	4.8	18.5	7.8	0.45	0.89	28,152	45.7	23.4	8.6	0.57	0.91
	4	2,329	3.8	12.5	6.6	0.32	0.87	28,766	47.0	18.1	8.4	0.46	0.91
	5	1,577	2.6	8.9	5.3	0.22	0.81	29,865	48.5	16.1	8.5	0.39	0.91
	6	1,364	2.2	8.1	5.2	0.21	0.82	29,159	48.0	15.7	8.4	0.40	0.91
	7	1,387	2.3	6.0	4.4	0.17	0.82	28,595	47.4	12.0	7.2	0.33	0.90
	8	1,412	2.3	6.3	4.2	0.17	0.76	30,977	51.1	11.6	6.9	0.31	0.88

Table 6.11: Scale Score Descriptive Statistics by Total

Content	Grade	*N*	Mean	*SD*	Skewness	Kurtosis	Lowest		Highest	
							Observed	Possible	Observed	Possible
ELA	3	61,525	2433.4	82.9	-0.1	-0.3	2114	2114	2623	2623
	4	61,121	2468.3	90.5	-0.3	-0.2	2131	2131	2663	2663
	5	61,600	2511.3	90.5	-0.1	-0.5	2201	2201	2701	2701
	6	60,859	2519.4	96.0	-0.1	-0.4	2210	2210	2724	2724
	7	60,590	2553.3	98.2	-0.2	-0.4	2258	2258	2745	2745
	8	60,842	2569.8	96.7	-0.3	-0.3	2288	2288	2769	2769
Mathematics	3	61,571	2436.4	73.8	-0.2	0.0	2189	2189	2621	2621
	4	61,148	2475.0	80.3	-0.4	0.3	2204	2204	2659	2659
	5	61,597	2504.0	89.8	-0.2	-0.1	2219	2219	2700	2700
	6	60,690	2525.3	98.2	-0.3	0.0	2235	2235	2748	2748
	7	60,368	2544.2	105.0	-0.3	-0.1	2250	2250	2778	2778
	8	60,664	2559.9	106.4	-0.2	0.0	2265	2265	2802	2802

Table 6.12: The Number and Percentage of Test Takers Earning the Lowest and Highest Possible Scale Scores

Content	Grade	Lowest Possible	*N*	%	Highest Possible	*N*	%
ELA	3	2114	24	0.04	2623	514	0.84
	4	2131	51	0.08	2663	553	0.90
	5	2201	22	0.04	2701	958	1.56
	6	2210	88	0.14	2724	705	1.16
	7	2258	135	0.22	2745	1,265	2.09
	8	2288	245	0.40	2769	719	1.18
Mathematics	3	2189	89	0.14	2621	406	0.66
	4	2204	259	0.42	2659	603	0.99
	5	2219	177	0.29	2700	752	1.22
	6	2235	498	0.82	2748	402	0.66
	7	2250	478	0.79	2778	345	0.57
	8	2265	756	1.25	2802	743	1.22

Table 6.13: Scale Score Descriptive Statistics by Gender

Content	Grade	Male						Female					
		N	%	Mean	*SD*	Min	Max	*N*	%	Mean	*SD*	Min	Max
ELA	3	31,437	51.1	2424.1	82.7	2114	2623	30,088	48.9	2443.2	82.1	2114	2623
	4	31,180	51.0	2456.9	90.0	2131	2663	29,941	49.0	2480.2	89.4	2131	2663
	5	31,556	51.2	2500.6	90.3	2201	2701	30,044	48.8	2522.5	89.4	2201	2701
	6	31,125	51.1	2505.6	95.3	2210	2724	29,734	48.9	2533.9	94.5	2210	2724
	7	30,829	50.9	2535.8	98.0	2258	2745	29,761	49.1	2571.5	95.1	2258	2745
	8	31,203	51.3	2553.4	98.1	2288	2769	29,639	48.7	2587.1	92.1	2288	2769
Mathematics	3	31,470	51.1	2438.3	75.2	2189	2621	30,101	48.9	2434.5	72.3	2189	2621
	4	31,191	51.0	2477.3	82.6	2204	2659	29,957	49.0	2472.7	77.6	2204	2659
	5	31,549	51.2	2504.3	93.2	2219	2700	30,048	48.8	2503.7	86.1	2219	2700
	6	31,036	51.1	2523.9	101.4	2235	2748	29,654	48.9	2526.8	94.7	2235	2748
	7	30,710	50.9	2542.8	109.0	2250	2778	29,658	49.1	2545.7	100.6	2250	2778
	8	31,098	51.3	2557.5	109.2	2265	2802	29,566	48.7	2562.4	103.3	2265	2802

Table 6.14: Scale Score Descriptive Statistics for ELA by Race/Ethnicity

Content	Grade	Race/Ethnicity	*N*	%	Mean	*SD*	Min	Max
ELA	3	White	41,391	67.3	2451.7	77.9	2114	2623
		African American	6,899	11.2	2367.5	74.0	2114	2623
		Hispanic	8,003	13.0	2402.1	76.9	2114	2623
		Asian	2,371	3.9	2427.3	82.0	2130	2623
		American Indian	838	1.4	2401.0	73.3	2188	2609
		Pacific Islander	47	0.1	2434.4	74.2	2250	2612
		Two or More Races	1,976	3.2	2429.4	83.2	2146	2623
	4	White	42,143	69.0	2486.3	84.4	2131	2663
		African American	6,683	10.9	2395.6	86.6	2131	2663
		Hispanic	7,362	12.0	2435.1	84.9	2131	2663
		Asian	2,329	3.8	2475.3	90.8	2181	2663
		American Indian	835	1.4	2432.1	82.2	2131	2663
		Pacific Islander	65	0.1	2469.7	82.4	2325	2646
		Two or More Races	1,704	2.8	2461.1	88.8	2131	2663
	5	White	42,944	69.7	2530.1	85.2	2201	2701
		African American	6,522	10.6	2437.2	78.9	2201	2701
		Hispanic	7,304	11.9	2472.1	82.9	2201	2701
		Asian	2,317	3.8	2512.3	90.5	2201	2701
		American Indian	824	1.3	2469.9	80.7	2240	2701
		Pacific Islander	40	0.1	2516.2	96.0	2282	2689
		Two or More Races	1,649	2.7	2505.1	90.5	2239	2701

Table 6.14: Scale Score Descriptive Statistics for ELA by Race/Ethnicity (Cont'd)

Content	Grade	Race/Ethnicity	*N*	%	Mean	*SD*	Min	Max
ELA	6	White	42,833	70.4	2538.9	90.1	2210	2724
		African American	6,368	10.5	2441.0	84.6	2210	2724
		Hispanic	7,131	11.7	2477.7	89.0	2210	2724
		Asian	2,185	3.6	2522.8	97.8	2210	2724
		American Indian	796	1.3	2475.7	87.8	2210	2724
		Pacific Islander	47	0.1	2521.2	90.7	2309	2713
		Two or More Races	1,499	2.5	2512.6	96.1	2210	2724
	7	White	42,926	70.8	2571.2	92.3	2258	2745
		African American	6,328	10.4	2476.5	94.3	2258	2745
		Hispanic	6,886	11.4	2518.0	93.3	2258	2745
		Asian	2,127	3.5	2560.2	98.0	2258	2745
		American Indian	812	1.3	2507.7	90.3	2258	2739
		Pacific Islander	37	0.1	2555.3	74.7	2372	2728
		Two or More Races	1,474	2.4	2542.6	94.9	2258	2745
	8	White	43,495	71.5	2585.5	92.0	2288	2769
		African American	6,257	10.3	2498.8	92.1	2288	2769
		Hispanic	6,753	11.1	2536.0	91.2	2288	2769
		Asian	2,155	3.5	2587.1	94.2	2288	2769
		American Indian	786	1.3	2526.5	90.6	2288	2759
		Pacific Islander	38	0.1	2580.2	97.9	2338	2744
		Two or More Races	1,358	2.2	2560.9	97.8	2288	2769

Table 6.15: Scale Score Descriptive Statistics for Mathematics by Race/Ethnicity

Content	Grade	Race/Ethnicity	*N*	%	Mean	*SD*	Min	Max
Mathematics	3	White	41,381	67.2	2453.3	68.0	2189	2621
		African American	6,855	11.1	2374.7	69.1	2189	2621
		Hispanic	8,062	13.1	2405.9	67.4	2189	2621
		Asian	2,407	3.9	2441.4	75.3	2189	2621
		American Indian	839	1.4	2412.1	66.0	2189	2621
		Pacific Islander	49	0.1	2425.7	76.8	2207	2610
		Two or More Races	1,978	3.2	2426.5	74.6	2189	2621
	4	White	42,134	68.9	2493.6	71.8	2204	2659
		African American	6,636	10.9	2400.5	79.3	2204	2659
		Hispanic	7,424	12.1	2437.8	74.2	2204	2659
		Asian	2,354	3.8	2484.0	84.4	2204	2659
		American Indian	834	1.4	2447.1	68.3	2204	2648
		Pacific Islander	65	0.1	2469.0	81.1	2255	2659
		Two or More Races	1,701	2.8	2469.1	78.4	2204	2659
	5	White	42,919	69.7	2524.1	83.1	2219	2700
		African American	6,479	10.5	2423.0	78.4	2219	2700
		Hispanic	7,348	11.9	2462.2	80.6	2219	2700
		Asian	2,347	3.8	2514.4	90.6	2219	2700
		American Indian	819	1.3	2459.8	82.6	2219	2681
		Pacific Islander	40	0.1	2509.8	84.6	2277	2680
		Two or More Races	1,645	2.7	2492.9	88.5	2219	2700

Table 6.15: Scale Score Descriptive Statistics for Mathematics by Race/Ethnicity (Cont'd)

Content	Grade	Race/Ethnicity	*N*	%	Mean	*SD*	Min	Max
Mathematics	6	White	42,787	70.5	2547.4	89.9	2235	2748
		African American	6,285	10.4	2435.2	86.9	2235	2748
		Hispanic	7,074	11.7	2476.9	89.7	2235	2748
		Asian	2,212	3.6	2536.8	100.6	2235	2748
		American Indian	789	1.3	2469.9	89.8	2235	2708
		Pacific Islander	47	0.1	2516.0	86.7	2235	2745
		Two or More Races	1,496	2.5	2513.1	97.7	2235	2748
	7	White	42,870	71.0	2565.9	96.9	2250	2778
		African American	6,208	10.3	2450.6	94.8	2250	2772
		Hispanic	6,838	11.3	2498.2	97.6	2250	2778
		Asian	2,131	3.5	2561.5	108.7	2250	2778
		American Indian	812	1.3	2487.4	100.0	2250	2778
		Pacific Islander	37	0.1	2547.5	101.8	2346	2703
		Two or More Races	1,472	2.4	2527.9	104.3	2250	2778
	8	White	43,440	71.6	2579.7	100.6	2265	2802
		African American	6,137	10.1	2469.2	92.6	2265	2802
		Hispanic	6,742	11.1	2515.0	95.8	2265	2802
		Asian	2,178	3.6	2585.3	110.9	2265	2802
		American Indian	784	1.3	2505.2	92.5	2265	2802
		Pacific Islander	37	0.1	2575.4	104.4	2361	2789
		Two or More Races	1,346	2.2	2550.5	102.1	2265	2802

Table 6.16: Scale Score Descriptive Statistics by Socioeconomic Status

Content	Grade	Economically Disadvantaged						Not Economically Disadvantaged					
		N	%	Mean	*SD*	Min	Max	*N*	%	Mean	*SD*	Min	Max
ELA	3	29,352	47.7	2401.1	77.9	2114	2623	32,173	52.3	2462.9	76.1	2114	2623
	4	27,915	45.7	2432.2	87.2	2131	2663	33,206	54.3	2498.7	81.5	2131	2663
	5	27,420	44.5	2472.9	84.7	2201	2701	34,180	55.5	2542.0	82.9	2201	2701
	6	26,430	43.4	2477.9	89.7	2210	2724	34,429	56.6	2551.3	88.1	2210	2724
	7	25,970	42.9	2512.8	94.7	2258	2745	34,620	57.1	2583.8	89.4	2258	2745
	8	25,309	41.6	2531.8	94.0	2288	2769	35,533	58.4	2596.9	89.2	2288	2769
Mathematics	3	29,374	47.7	2407.4	70.1	2189	2621	32,197	52.3	2463.0	66.8	2189	2621
	4	27,930	45.7	2440.8	77.9	2204	2659	33,218	54.3	2503.9	70.3	2204	2659
	5	27,406	44.5	2464.3	83.7	2219	2700	34,191	55.5	2535.8	81.5	2219	2700
	6	26,257	43.3	2479.9	92.3	2235	2748	34,433	56.7	2559.9	87.9	2235	2748
	7	25,776	42.7	2497.6	100.1	2250	2778	34,592	57.3	2579.0	94.5	2250	2778
	8	25,162	41.5	2513.8	98.7	2265	2802	35,502	58.5	2592.6	99.3	2265	2802

Table 6.17: Scale Score Descriptive Statistics by Disability

Content	Grade	Disabled						Not Disabled					
		N	%	Mean	*SD*	Min	Max	*N*	%	Mean	*SD*	Min	Max
ELA	3	7,161	11.6	2383.9	81.4	2114	2623	54,364	88.4	2439.9	80.9	2114	2623
	4	7,091	11.6	2406.1	89.3	2131	2663	54,030	88.4	2476.5	87.4	2131	2663
	5	7,439	12.1	2435.2	83.5	2201	2701	54,161	87.9	2521.7	86.4	2201	2701
	6	7,171	11.8	2434.9	83.5	2210	2724	53,688	88.2	2530.7	91.8	2210	2724
	7	7,178	11.8	2453.2	88.1	2258	2745	53,412	88.2	2566.8	91.5	2258	2745
	8	7,231	11.9	2475.4	89.1	2288	2769	53,611	88.1	2582.6	90.4	2288	2769
Mathematics	3	7,151	11.6	2389.8	79.8	2189	2621	54,420	88.4	2442.6	70.8	2189	2621
	4	7,069	11.6	2416.3	88.8	2204	2659	54,079	88.4	2482.7	75.8	2204	2659
	5	7,409	12.0	2434.0	89.5	2219	2700	54,188	88.0	2513.6	85.5	2219	2700
	6	7,101	11.7	2433.1	101.1	2235	2748	53,589	88.3	2537.5	91.1	2235	2748
	7	7,123	11.8	2441.8	102.1	2250	2778	53,245	88.2	2557.9	97.5	2250	2778
	8	7,168	11.8	2460.6	98.1	2265	2802	53,496	88.2	2573.2	100.3	2265	2802

Table 6.18: Scale Score Descriptive Statistics by English Language Proficiency

Content	Grade	Limited English Proficient						Fully English Proficient					
		N	%	Mean	*SD*	Min	Max	*N*	%	Mean	*SD*	Min	Max
ELA	3	5,601	9.1	2389.6	71.2	2114	2623	55,924	90.9	2437.8	82.7	2114	2623
	4	4,521	7.4	2412.7	76.3	2131	2663	56,600	92.6	2472.8	90.0	2131	2663
	5	3,248	5.3	2429.1	64.9	2201	2699	58,352	94.7	2515.8	89.5	2201	2701
	6	2,609	4.3	2427.4	71.4	2210	2710	58,250	95.7	2523.5	94.8	2210	2724
	7	2,690	4.4	2463.2	79.2	2258	2745	57,900	95.6	2557.5	97.0	2258	2745
	8	2,870	4.7	2493.7	77.9	2288	2769	57,972	95.3	2573.6	96.0	2288	2769
Mathematics	3	5,715	9.3	2399.6	65.3	2189	2621	55,856	90.7	2440.2	73.6	2189	2621
	4	4,629	7.6	2422.7	71.4	2204	2659	56,519	92.4	2479.3	79.4	2204	2659
	5	3,341	5.4	2431.0	72.0	2219	2700	58,256	94.6	2508.2	88.9	2219	2700
	6	2,648	4.4	2433.6	81.1	2235	2741	58,042	95.6	2529.5	96.9	2235	2748
	7	2,708	4.5	2453.3	90.1	2250	2778	57,660	95.5	2548.5	103.7	2250	2778
	8	2,942	4.8	2481.1	87.6	2265	2802	57,722	95.2	2563.9	105.7	2265	2802

Table 6.19: Percentage of Test Takers in Performance Level, ELA Grade 3

	Number of Test Takers	Performance Levels			
		Below Basic	Basic	Proficient	Advanced
Total	61,525	22.4	25.5	25.5	26.6
Gender					
Male	31,437	25.8	26.8	24.4	23.0
Female	30,088	18.8	24.1	26.7	30.4
Race/Ethnicity					
White	41,391	14.8	23.7	28.4	33.2
African American	6,899	53.0	27.2	13.8	6.0
Hispanic/Latino	8,003	33.6	31.0	22.0	13.4
Asian	2,371	23.7	30.0	23.1	23.1
American Indian	838	32.3	34.2	20.9	12.5
Pacific Islander	47	19.1	27.7	31.9	21.3
Two or More Races	1,976	24.2	26.0	25.1	24.6
ELP					
Limited English Proficient	5,601	38.1	33.5	20.4	8.0
Fully English Proficient	55,924	20.8	24.7	26.0	28.5
Disability					
Disabled	7,161	43.9	28.1	17.3	10.7
Not Disabled	54,364	19.6	25.1	26.6	28.7
SES					
Economically Disadvantaged	29,352	34.4	30.2	21.8	13.6
Not Economically Disadvantaged	32,173	11.4	21.2	28.9	38.5

Note. ELP=English language proficiency; SES=Socioeconomic status

Table 6.20: Percentage of Test Takers in Performance Level, ELA Grade 4

	Number of Test Takers	Performance Levels			
		Below Basic	Basic	Proficient	Advanced
Total	61,121	28.1	20.9	25.4	25.5
Gender					
Male	31,180	32.3	22.0	24.6	21.1
Female	29,941	23.8	19.9	26.3	30.0
Race/Ethnicity					
White	42,143	20.3	20.4	28.3	30.9
African American	6,683	60.8	19.3	13.5	6.4
Hispanic/Latino	7,362	41.2	24.7	21.2	12.9
Asian	2,329	26.9	21.2	23.3	28.7
American Indian	835	42.0	24.6	22.2	11.3
Pacific Islander	65	32.3	16.9	27.7	23.1
Two or More Races	1,704	30.9	22.1	24.1	22.9
ELP					
Limited English Proficient	4,521	51.5	26.5	16.3	5.7
Fully English Proficient	56,600	26.3	20.5	26.2	27.1
Disability					
Disabled	7,091	56.1	20.8	14.6	8.6
Not Disabled	54,030	24.5	21.0	26.9	27.7
SES					
Economically Disadvantaged	27,915	42.7	23.3	21.2	12.8
Not Economically Disadvantaged	33,206	15.9	18.9	29.0	36.1

Note. ELP=English language proficiency; SES=Socioeconomic status

Table 6.21: Percentage of Test Takers in Performance Level, ELA Grade 5

	Number of Test Takers	Performance Levels			
		Below Basic	Basic	Proficient	Advanced
Total	61,600	23.8	21.1	31.3	23.7
Gender					
Male	31,556	27.4	22.2	30.3	20.1
Female	30,044	20.1	20.0	32.4	27.5
Race/Ethnicity					
White	42,944	16.1	19.8	34.8	29.2
African American	6,522	56.2	22.9	16.4	4.5
Hispanic/Latino	7,304	38.3	26.2	25.2	10.3
Asian	2,317	23.6	22.1	30.4	24.0
American Indian	824	38.3	27.4	25.0	9.2
Pacific Islander	40	20.0	17.5	40.0	22.5
Two or More Races	1,649	26.7	21.2	30.4	21.6
ELP					
Limited English Proficient	3,248	59.8	26.8	11.9	1.5
Fully English Proficient	58,352	21.8	20.8	32.4	24.9
Disability					
Disabled	7,439	58.1	21.0	15.1	5.8
Not Disabled	54,161	19.1	21.2	33.5	26.2
SES					
Economically Disadvantaged	27,420	38.2	24.9	25.9	11.0
Not Economically Disadvantaged	34,180	12.3	18.1	35.7	33.9

Note. ELP=English language proficiency; SES=Socioeconomic status

Table 6.22: Percentage of Test Takers in Performance Level, ELA Grade 6

	Number of Test Takers	Performance Levels			
		Below Basic	Basic	Proficient	Advanced
Total	60,859	27.0	26.3	30.3	16.4
Gender					
Male	31,125	31.7	27.1	28.4	12.8
Female	29,734	22.1	25.4	32.3	20.2
Race/Ethnicity					
White	42,833	19.0	25.9	34.7	20.3
African American	6,368	61.1	23.7	12.5	2.7
Hispanic/Latino	7,131	42.4	29.8	21.4	6.5
Asian	2,185	26.4	28.1	27.7	17.8
American Indian	796	44.2	28.4	22.4	5.0
Pacific Islander	47	25.5	27.7	31.9	14.9
Two or More Races	1,499	29.6	26.4	29.8	14.3
ELP					
Limited English Proficient	2,609	67.6	24.6	7.1	0.7
Fully English Proficient	58,250	25.2	26.4	31.3	17.1
Disability					
Disabled	7,171	64.7	22.5	10.1	2.7
Not Disabled	53,688	22.0	26.8	33.0	18.2
SES					
Economically Disadvantaged	26,430	42.7	29.0	21.9	6.5
Not Economically Disadvantaged	34,429	15.0	24.2	36.7	24.0

Note. ELP=English language proficiency; SES=Socioeconomic status

Table 6.23: Percentage of Test Takers in Performance Level, ELA Grade 7

	Number of Test Takers	Performance Levels			
		Below Basic	Basic	Proficient	Advanced
Total	60,590	22.9	24.6	34.9	17.6
Gender					
Male	30,829	28.3	26.4	32.3	13.0
Female	29,761	17.4	22.6	37.6	22.4
Race/Ethnicity					
White	42,926	16.5	23.5	38.6	21.4
African American	6,328	52.3	25.6	18.4	3.7
Hispanic/Latino	6,886	34.6	28.9	28.3	8.2
Asian	2,127	19.6	26.8	34.0	19.6
American Indian	812	39.2	29.1	24.5	7.3
Pacific Islander	37	10.8	37.8	37.8	13.5
Two or More Races	1,474	26.4	25.6	33.9	14.2
ELP					
Limited English Proficient	2,690	58.0	29.3	11.3	1.4
Fully English Proficient	57,900	21.3	24.3	36.0	18.4
Disability					
Disabled	7,178	64.4	21.9	11.2	2.5
Not Disabled	53,412	17.4	24.9	38.1	19.6
SES					
Economically Disadvantaged	25,970	36.4	28.4	27.4	7.8
Not Economically Disadvantaged	34,620	12.9	21.7	40.5	25.0

Note. ELP=English language proficiency; SES=Socioeconomic status

Table 6.24: Percentage of Test Takers in Performance Level, ELA Grade 8

	Number of Test Takers	Performance Levels			
		Below Basic	Basic	Proficient	Advanced
Total	60,842	20.3	25.2	39.0	15.5
Gender					
Male	31,203	25.5	27.1	35.5	12.0
Female	29,639	14.8	23.2	42.7	19.3
Race/Ethnicity					
White	43,495	15.1	23.5	42.7	18.7
African American	6,257	45.9	29.4	21.7	3.0
Hispanic/Latino	6,753	29.3	31.9	32.1	6.7
Asian	2,155	15.0	23.8	41.0	20.2
American Indian	786	34.7	30.7	28.9	5.7
Pacific Islander	38	15.8	23.7	39.5	21.1
Two or More Races	1,358	23.4	25.6	36.6	14.4
ELP					
Limited English Proficient	2,870	45.9	36.2	16.8	1.1
Fully English Proficient	57,972	19.0	24.6	40.1	16.3
Disability					
Disabled	7,231	58.2	26.0	13.6	2.2
Not Disabled	53,611	15.2	25.1	42.4	17.3
SES					
Economically Disadvantaged	25,309	32.1	30.1	31.0	6.8
Not Economically Disadvantaged	35,533	11.9	21.7	44.7	21.8

Note. ELP=English language proficiency; SES=Socioeconomic status

Table 6.25: Percentage of Test Takers in Performance Level, Mathematics Grade 3

	Number of Test Takers	Performance Levels			
		Below Basic	Basic	Proficient	Advanced
Total	61,571	21.9	26.3	32.5	19.3
Gender					
Male	31,470	21.5	25.7	32.2	20.6
Female	30,101	22.4	26.9	32.8	17.9
Race/Ethnicity					
White	41,381	13.9	24.5	37.2	24.4
African American	6,855	53.2	27.3	16.3	3.2
Hispanic/Latino	8,062	35.0	32.5	24.6	7.9
Asian	2,407	20.4	28.3	29.6	21.7
American Indian	839	31.7	34.2	25.0	9.1
Pacific Islander	49	22.4	34.7	28.6	14.3
Two or More Races	1,978	26.2	28.1	30.1	15.6
ELP					
Limited English Proficient	5,715	37.5	34.0	22.6	5.9
Fully English Proficient	55,856	20.3	25.5	33.5	20.7
Disability					
Disabled	7,151	46.0	25.3	19.9	8.8
Not Disabled	54,420	18.8	26.4	34.1	20.7
SES					
Economically Disadvantaged	29,374	34.4	30.9	25.8	9.0
Not Economically Disadvantaged	32,197	10.6	22.1	38.6	28.7

Note. ELP=English language proficiency; SES=Socioeconomic status

Table 6.26: Percentage of Test Takers in Performance Level, Mathematics Grade 4

	Number of Test Takers	Performance Levels			
		Below Basic	Basic	Proficient	Advanced
Total	61,148	19.9	32.4	30.4	17.3
Gender					
Male	31,191	19.6	31.3	30.1	19.0
Female	29,957	20.1	33.7	30.8	15.4
Race/Ethnicity					
White	42,134	11.8	30.9	35.6	21.6
African American	6,636	53.8	32.0	11.8	2.4
Hispanic/Latino	7,424	34.4	39.3	20.0	6.3
Asian	2,354	17.8	33.4	26.7	22.1
American Indian	834	28.8	41.0	23.5	6.7
Pacific Islander	65	18.5	36.9	27.7	16.9
Two or More Races	1,701	21.9	35.8	28.0	14.3
ELP					
Limited English Proficient	4,629	41.3	40.1	15.3	3.3
Fully English Proficient	56,519	18.1	31.8	31.7	18.4
Disability					
Disabled	7,069	46.3	31.5	15.8	6.5
Not Disabled	54,079	16.4	32.6	32.3	18.7
SES					
Economically Disadvantaged	27,930	32.9	37.6	22.3	7.1
Not Economically Disadvantaged	33,218	8.9	28.1	37.2	25.8

Note. ELP=English language proficiency; SES=Socioeconomic status

Table 6.27: Percentage of Test Takers in Performance Level, Mathematics Grade 5

	Number of Test Takers	Performance Levels			
		Below Basic	Basic	Proficient	Advanced
Total	61,597	28.9	30.7	19.2	21.2
Gender					
Male	31,549	29.7	29.2	18.6	22.5
Female	30,048	28.1	32.3	19.8	19.8
Race/Ethnicity					
White	42,919	19.6	31.2	22.6	26.5
African American	6,479	67.4	23.6	6.2	2.8
Hispanic/Latino	7,348	46.8	33.1	12.5	7.6
Asian	2,347	26.2	32.1	17.3	24.4
American Indian	819	48.5	29.9	14.2	7.4
Pacific Islander	40	17.5	37.5	25.0	20.0
Two or More Races	1,645	33.8	31.9	16.4	17.9
ELP					
Limited English Proficient	3,341	63.9	27.9	6.2	2.1
Fully English Proficient	58,256	26.9	30.9	19.9	22.3
Disability					
Disabled	7,409	61.4	24.2	7.9	6.5
Not Disabled	54,188	24.5	31.6	20.7	23.2
SES					
Economically Disadvantaged	27,406	45.5	32.3	13.5	8.7
Not Economically Disadvantaged	34,191	15.6	29.4	23.7	31.2

Note. ELP=English language proficiency; SES=Socioeconomic status

Table 6.28: Percentage of Test Takers in Performance Level, Mathematics Grade 6

	Number of Test Takers	Performance Levels			
		Below Basic	Basic	Proficient	Advanced
Total	60,690	28.3	30.3	21.1	20.2
Gender					
Male	31,036	29.2	29.1	21.0	20.7
Female	29,654	27.3	31.7	21.3	19.7
Race/Ethnicity					
White	42,787	18.9	30.9	25.0	25.2
African American	6,285	67.6	23.4	6.7	2.3
Hispanic/Latino	7,074	47.5	32.3	13.3	6.9
Asian	2,212	25.9	32.0	17.8	24.3
American Indian	789	49.3	33.2	12.7	4.8
Pacific Islander	47	25.5	42.6	17.0	14.9
Two or More Races	1,496	33.3	31.1	18.4	17.2
ELP					
Limited English Proficient	2,648	69.4	24.3	4.5	1.8
Fully English Proficient	58,042	26.4	30.6	21.9	21.0
Disability					
Disabled	7,101	65.6	22.3	7.8	4.3
Not Disabled	53,589	23.4	31.4	22.9	22.3
SES					
Economically Disadvantaged	26,257	45.6	32.2	14.7	7.5
Not Economically Disadvantaged	34,433	15.1	28.9	26.1	29.9

Note. ELP=English language proficiency; SES=Socioeconomic status

Table 6.29: Percentage of Test Takers in Performance Level, Mathematics Grade 7

	Number of Test Takers	Performance Levels			
		Below Basic	Basic	Proficient	Advanced
Total	60,368	27.3	28.9	23.2	20.6
Gender					
Male	30,710	28.7	27.0	22.8	21.5
Female	29,658	25.8	30.9	23.6	19.7
Race/Ethnicity					
White	42,870	19.1	28.8	26.8	25.3
African American	6,208	63.9	25.1	8.1	2.9
Hispanic/Latino	6,838	43.4	32.8	15.6	8.2
Asian	2,131	22.3	29.0	23.3	25.4
American Indian	812	49.8	28.2	13.9	8.1
Pacific Islander	37	27.0	27.0	18.9	27.0
Two or More Races	1,472	33.2	31.1	20.2	15.4
ELP					
Limited English Proficient	2,708	63.1	27.0	7.8	2.2
Fully English Proficient	57,660	25.6	29.0	23.9	21.5
Disability					
Disabled	7,123	67.9	20.0	8.2	4.0
Not Disabled	53,245	21.9	30.1	25.2	22.8
SES					
Economically Disadvantaged	25,776	43.4	31.5	16.8	8.3
Not Economically Disadvantaged	34,592	15.3	27.0	27.9	29.8

Note. ELP=English language proficiency; SES=Socioeconomic status

Table 6.30: Percentage of Test Takers in Performance Level, Mathematics Grade 8

	Number of Test Takers	Performance Levels			
		Below Basic	Basic	Proficient	Advanced
Total	60,664	28.6	31.1	21.0	19.3
Gender					
Male	31,098	30.1	30.2	20.1	19.5
Female	29,566	27.0	32.1	21.8	19.1
Race/Ethnicity					
White	43,440	21.0	31.1	24.4	23.5
African American	6,137	65.0	25.8	6.6	2.6
Hispanic/Latino	6,742	43.9	35.4	13.5	7.2
Asian	2,178	21.4	30.4	20.5	27.7
American Indian	784	49.0	32.3	13.0	5.7
Pacific Islander	37	21.6	35.1	16.2	27.0
Two or More Races	1,346	31.8	33.6	19.0	15.6
ELP					
Limited English Proficient	2,942	59.7	30.8	7.1	2.4
Fully English Proficient	57,722	27.0	31.1	21.7	20.2
Disability					
Disabled	7,168	68.8	21.6	6.5	3.1
Not Disabled	53,496	23.2	32.4	22.9	21.5
SES					
Economically Disadvantaged	25,162	44.6	33.4	14.2	7.7
Not Economically Disadvantaged	35,502	17.2	29.5	25.7	27.5

Note. ELP=English language proficiency; SES=Socioeconomic status

Table 6.31: Percentage of Test Takers in Claim Level, ELA

Grade	Claim	Number of Test Takers	Claim Levels		
			Does Not Meet Standard	Meets Standard	Exceeds Standard
3	1	61,446	24.8	44.1	31.1
	2	61,510	17.4	58.8	23.9
	3	61,369	8.8	69.4	21.8
	4	61,394	19.1	65.4	15.5
4	1	61,076	24.1	46.3	29.7
	2	61,109	17.5	64.9	17.6
	3	61,002	26.1	54.2	19.7
	4	61,028	12.6	73.0	14.4
5	1	61,564	22.4	39.6	38.0
	2	61,592	22.2	58.5	19.4
	3	61,490	14.5	65.8	19.7
	4	61,541	19.0	61.4	19.6
6	1	60,783	31.8	45.2	23.1
	2	60,845	3.2	77.9	19.0
	3	60,644	18.1	69.2	12.7
	4	60,709	13.3	77.3	9.4
7	1	60,473	21.3	50.3	28.5
	2	60,571	17.1	58.2	24.7
	3	60,335	23.7	63.6	12.7
	4	60,398	12.4	69.3	18.4
8	1	60,733	21.6	43.2	35.2
	2	60,775	18.4	63.8	17.8
	3	60,635	18.4	70.8	10.8
	4	60,699	10.4	74.2	15.4

Note. Numbers of test takers by claim can vary because a claim score is provided only when a test taker answered all items in each claim.

Table 6.32: Percentage of Test Takers in Claim Level, Mathematics

Grade	Claim	Number of Test Takers	Claim Levels		
			Does Not Meet Standard	Meets Standard	Exceeds Standard
3	1	61,535	30.9	41.5	27.7
	2&4	61,557	15.5	55.2	29.3
	3	61,536	15.8	62.5	21.8
4	1	61,115	35.3	38.1	26.6
	2&4	61,118	17.0	60.2	22.7
	3	61,106	11.8	63.8	24.4
5	1	61,563	38.4	39.4	22.2
	2&4	61,575	32.3	48.5	19.2
	3	61,553	24.1	55.4	20.5
6	1	60,623	39.5	36.2	24.3
	2&4	60,635	32.6	46.0	21.4
	3	60,594	19.6	60.4	20.0
7	1	60,298	40.0	32.4	27.6
	2&4	60,292	21.1	56.1	22.8
	3	60,286	35.4	41.8	22.9
8	1	60,595	34.9	43.3	21.8
	2&4	60,634	17.6	62.4	20.0
	3	60,582	20.5	64.3	15.2

Note. Numbers of test takers by claim can vary because a claim score is provided only when a test taker answered all items in each claim.

Figure 6.1: Score Range and Percentage of Test Takers at each Performance Level, ELA

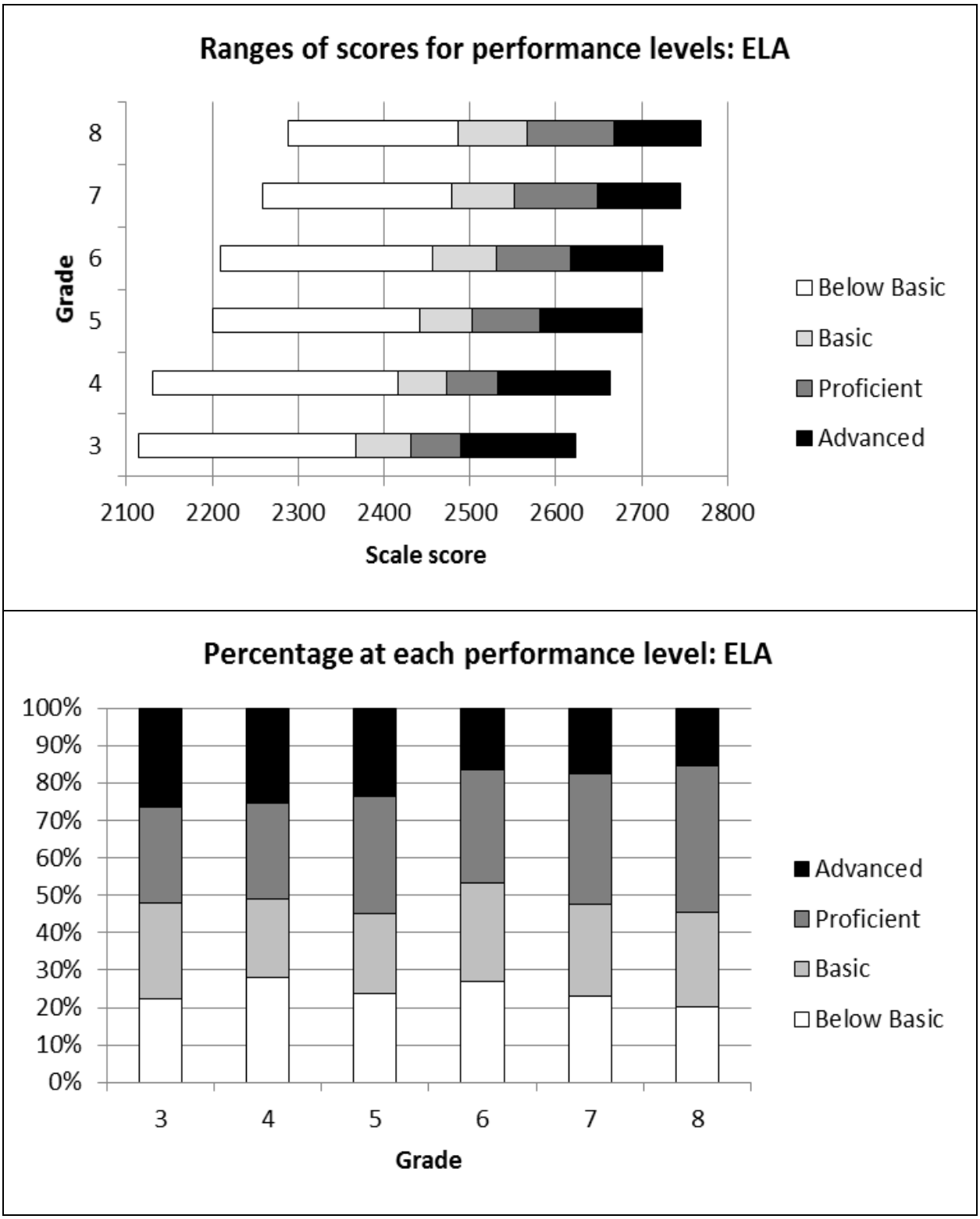
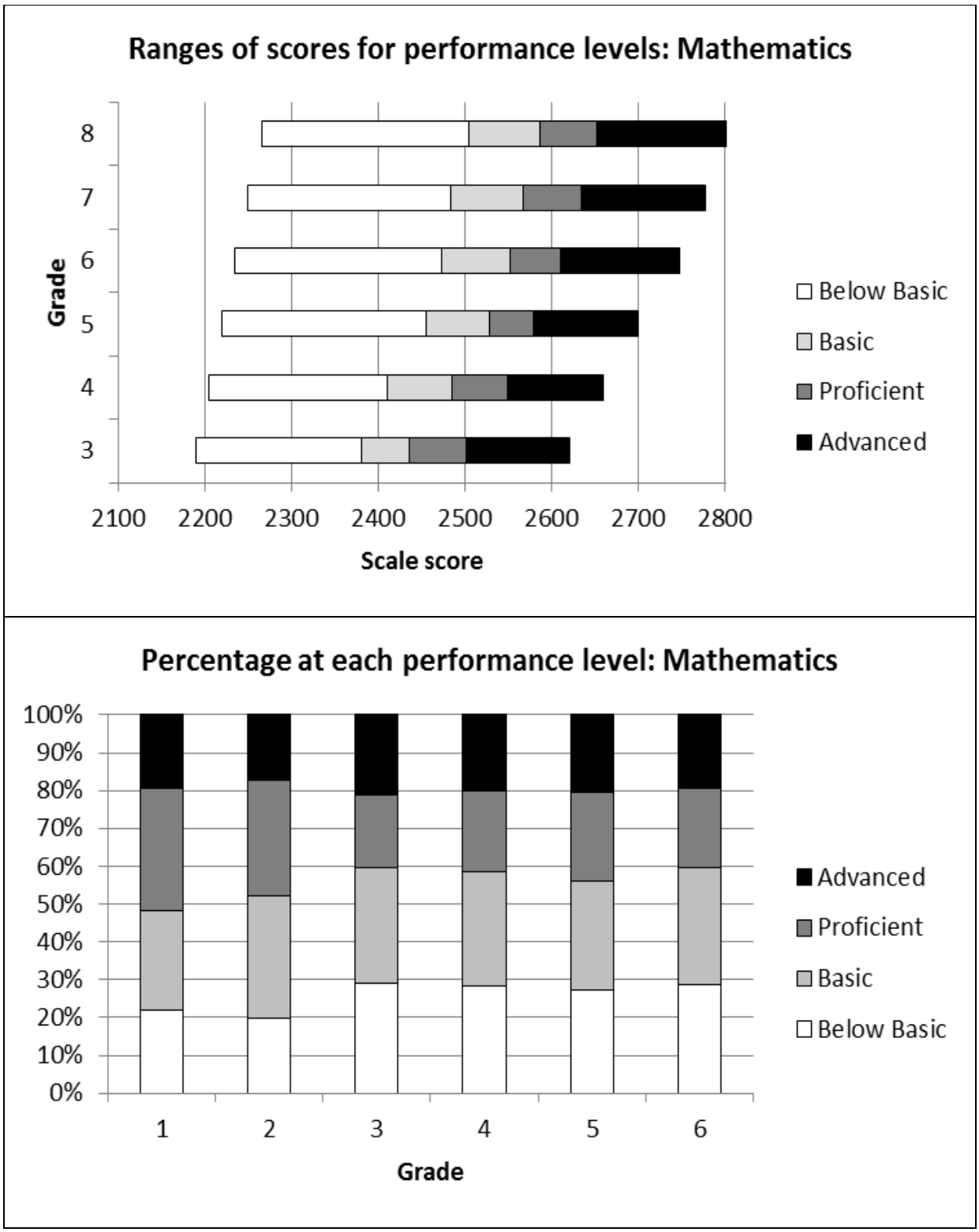


Figure 6.2: Score Range and Percentage of Test Takers at each Performance Level, Mathematics



CHAPTER 7: RELIABILITY

7.1: Overview of Reliability

7.2: Internal Consistency and Standard Error of Measurement

7.3: Interrater Agreement

7.4: Classification Consistency and Accuracy

7.1: Overview of Reliability

There are many definitions of reliability (Haertel, 2006) that have their genesis in classical test theory, and a variety of methods can be used to estimate reliability. The general concept of reliability concerns the consistency of a test score. It is the extent to which the information provided by the scores does not change, when there are changes in things that can reasonably be assumed affect the scores randomly. Two of these random effects are: (1) the particular form of the test that a test taker takes and (2) the particular raters who score the test taker's responses to constructed response tasks.

The reliability coefficient, as a measure of the general concept of reliability, has some limitations. The reliability coefficient applies only to the scores of a group, not to the score of an individual test taker. Also, it takes into account only the variation within that group, not to the absolute level of the scores or the extent to which they differ from some other set of scores (e.g., the scores of the same group tested at a different stage of their instruction).

This Chapter documents how DPI and ETS accumulated appropriate reliability evidence for the Badger Exam in several components for the interpretation for each intended score use (AERA, APA, & NCME, 2014). The components of reliability examined in this Chapter are internal consistency, standard error of measurement, interrater agreement, and classification consistency and accuracy.

7.2: Internal Consistency and Standard Error of Measurement

An internal consistency reliability coefficient is an indicator of how similar test takers perform across items measuring similar knowledge and skills; that is, how consistently does each test taker perform on all of the items within a test. In classical test theory, the reliability coefficient is defined in four different ways in terms of two statistical bases (i.e., proportion of variance and correlations) and two conceptual bases (i.e., true scores and measurement error)[6]. One of the definitions of reliability coefficient is the proportion of observed total score variance that is true-score variance. The variance in the test score distributions is partially due to real differences in the knowledge or skills being tested and partially due to random variation in the measurement process, which is typically referred to as "measurement error." Several different ways of estimating this proportion exist. One of the commonly used estimates of reliability is Cronbach's alpha (Cronbach, 1951), an internal consistency measure. It is derived from analysis of the consistency of performance over items within a test and provides a lower-bound estimate of a test's reliability. There are two requirements to estimate this score reliability coefficient:

1. The test should be homogeneous; i.e., the items should all measure the same set of knowledge and skills, and
2. The number of items should be sufficient to obtain stable estimates of test takers' achievement.

[6] a) the ratio of true score variance to observed score variance, b) the square of the correlation between observed scores and true scores, c) the lack of error variance, and d) the lack of correlation between observed scores and error scores.

The Badger Exams are complex measures that combine selected response, constructed response, and technology enhanced items. Test taker performance may differ from one item to another due to interactions with prior knowledge, educational experiences, and exposure to similar content or item format. The heterogeneity of items in the ELA and Mathematics tests may result in an underestimate of the reliability of test scores as estimated by Cronbach's alpha. The number of total score points and test structure of the Badger Exam are sufficient to justify the use of Cronbach's alpha to compute an internal consistency estimate of reliability. Cronbach's alpha is represented by:

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_{i=1}^{n}\sigma_i^2}{\sigma_x^2}\right),$$

where n is the number of items, σ_i^2 is the variance of scores on the ith item, and σ_x^2 is the variance of the total score (sum of scores on the individual items). Sample estimates are substituted for the population variances in this formula to provide reliability estimates.

Another way to interpret the reliability of test scores is with the standard error of measurement (SEM). It is useful to take into account the size of measurement errors because no assessment measures test taker ability with perfect consistency. The SEM is an estimate of the typical difference between test takers' actual scores and the scores they would get if the test (and the scoring procedure) were perfectly reliable. The SEM is inversely related to the reliability — the higher the reliability, the lower the SEM and the more confidence one may have in the precision of the observed test score. The measurement error is commonly expressed in terms of standard deviation units; that is, the SEM is the standard deviation of the measurement error distribution.

Under classical test theory and traditional item analysis, SEM is estimated by:

$$\sigma_e = \sigma_x\sqrt{1-r},$$

where σ_x is the observed score standard deviation of the raw score distribution, and r is reliability (i.e., Cronbach's alpha).

In the IRT framework, SEM can be estimated in a different way and is often referred to as a conditional standard error of measurement (CSEM, see Chapter 5). CSEMs are smaller in scale score units toward the center of the scale where there are more items and, therefore, more test information. In contrast, CSEMs are larger at the extremes where there are fewer items and less test information.

Tables 7.1 through 7.12 show Cronbach's alpha and SEMs of the 2015 Badger Exam at all grades and subject areas for the total test taker group and the various subgroups such as gender, race/ethnicity, socioeconomic status, disability status, and English language proficiency subgroups. Coefficient alphas for total test takers were, for the most part, close to 0.90 and not lower than 0.87 for any grade/subject combination. The overall reliability of ELA and Mathematics tests ranged from 0.87 to 0.91 and from 0.89 to 0.91, respectively. These observed reliabilities meet generally accepted industry levels and benchmarks for large-scale assessments. The test scores from ELA and Mathematics all exhibit relatively high coefficient alphas suggesting the construct is being measured consistently.

The results also show that the subgroup reliability coefficients were quite similar or slightly lower than the total reliability coefficients. In general, subgroups usually had relatively smaller

standard deviations than the total test taker group (See Chapter 6), and reliability coefficients are sensitive to the score distribution and variance.

The differences in reliability among the subgroups were mostly quite small. The differences between male and female test takers were within 0.01 for all grades and subject areas. The differences among the race/ethnicity groups were moderate and less than 0.10 except for the Mathematics grade 8 test. The largest reliability difference across ethnicity groups was 0.14 between the African American (0.78) and Pacific Islander (0.92) groups in Mathematics grade 8 PT1 form. This large reliability difference between African American and Pacific Islander groups can be explained by the unusually large variance (possibly due to the small sample size) of the Pacific Islander group (N = 17). The estimates for disabled and nondisabled test takers and for economically disadvantaged and nondisadvantaged test takers were within 0.07 of one another. For the most part, there were large differences found between English language proficiency groups. The reliability coefficients from the Limited English proficient group were consistently lower than from the fully English proficient group with differences ranging from 0.05 to 0.14 for ELA and from 0.02 to 0.12 for Mathematics. There was no notable difference between the two Mathematics forms in each grade (i.e., classified by different Section 2 – PT) except for the Pacific Islander group difference in the Mathematics grade 8 test. As mentioned previously, the sample sizes of Pacific Islander groups were quite small (N = 17 for PT1 and N = 20 for PT2), and the variances were quite different.

Tables 7.1 through 7.12 also present the raw score SEM for the total population and for the subgroups described previously. Overall the differences in raw score SEM for the different subgroups were small and within 0.5 score point. In addition, there was no difference between the two Mathematics forms (by PTs) in each grade; the differences being within 0.3 score point.

The conditional SEMs for individual scale scores based on the IRT framework are also provided in Figures 7.1 and 7.2, with the locations of the cut scores shown in each plot so that the associated SEMs can be easily located. These Figures include the truncated scale score SEM value (i.e., SEM = 25) as red horizontal lines. When a test taker's CSEM was higher than 25, the value was truncated and converted to 25 in the score report. The CSEM at the "Proficient" cut score was low in all grades and subject areas, although values were not always the lowest on the SEM curves.

Cronbach's alpha and SEM for various claim scores are provided in Tables 7.13 through 7.15. The maximum possible raw scores for the total and each claim are also provided to demonstrate the relationship between the number of score points and reliability – a smaller number of score points is generally associated with lower reliability. Overall, Claim 1 had the largest number of score points compared to other claims (18 to 21 score points for ELA and 20 to 21 score points for Mathematics). Other claims had 8 to 14 score points for each claim. Therefore, lower reliability coefficients for these claims are expected especially for Claims 2 to 4 from both ELA and Mathematics. For ELA, reliability indices by claim ranged from 0.50 (for Claim 3 in grade 8, with 9 score points) to 0.82 (for Claim 1 in grade 5, with 20 score points). For Mathematics, reliability indices by claim ranged from 0.48 (for Claim 3 in grade 8 PT2, with 8 score points) to 0.86 (for Claim 1 in grade 3 PT1, with 20 score points). SEMs were smaller than those for the total test and are generally consistent with the number of items within each claim.

7.3: Interrater Agreement

Rater agreement or consistency is critical for valid test score interpretation of assessments requiring human raters to score the performance of test takers. When two trained raters independently assign the same score to a test taker's item response, there is evidence that the scoring standard is being applied consistently. To monitor and evaluate the accuracy of rating, 10% of the responses to each human-scored item were scored twice. Interrater reliability was evaluated empirically in three ways: a) Percentage agreement between two raters, b) Intraclass correlation, and c) Weighted kappa coefficient.

Percentage agreement between two raters is frequently defined as the percentage of exact score and adjacent score agreement. Percentage of exact score agreement is a stringent criterion which tends to decrease with increasing numbers of item score points. The fewer the item score points, the fewer degrees of freedom on which two raters can vary, and the higher the percentage of agreement.

The intraclass correlation is defined as the percentage of overall score variance accounted for by the variance of mean response scores of two raters. The range of intraclass correlation is from 0.0 to 1.0, with perfect agreement equaling 1.0. Suppose that N is the number of responses that are scored twice, $X_{n1}$and X_{n2} are the two scores of response *n*, when *n*=1, 2, … *N*):

$$\rho_{IC} = \frac{\frac{1}{N-1}\sum_{n=1}^{N}[(\bar{X}_{n.}-\bar{X}_{..})^2]}{\frac{1}{2(N-1)}\sum_{n=1}^{N}[(X_{n1}-\bar{X}_{..})^2+(X_{n2}-\bar{X}_{..})^2]},$$

where

$$\bar{X}_{n.} = (X_{n1} + X_{n2})/2,$$

and

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

Weighted kappa is selected since kappa does not take into account the degree of disagreement between raters. It is a generalization of the simple kappa coefficient using weights to quantify the relative difference between categories. The range of weighted kappa is from 0.0 to 1.0, with perfect agreement equaling 1.0.

For a human-scored item with *m* categories, one can construct an *m* x *m* rating table with scores provided by two raters A and B. Suppose *m* is maximum obtainable score for each item, n_{ij} is the number of responses for which rater A's score = *i* and rater B's score = *j*, n_{i+} is the number of responses for which Rater A = *i*, n_{+j} is the number of responses for which Rater B = *j*, and n_{++} is the number of all responses from either rater A or rater B. The weighted kappa coefficient is defined as:

$$k_{ij} = \frac{\left(\sum_{i=0}^{m}\sum_{j=0}^{m} w_{ij}\frac{n_{ij}}{n_{++}}\right)-\left(\sum_{i=0}^{m}\sum_{j=0}^{m} w_{ij}\frac{n_{i+}n_{+j}}{n_{++}^2}\right)}{1-\left(\sum_{i=0}^{m}\sum_{j=0}^{m} w_{ij}\frac{n_{i+}n_{+j}}{n_{++}^2}\right)},$$

where

$$w_{ij} = 1 - \frac{(i-j)^2}{m^2}.$$

Tables 7.16 and 7.17 provide the agreement rates, intraclass correlations, and weighted kappas for the Badger Exam human-scored items. There were 13 human-scored items in ELA and 35 human-scored items in Mathematics (See Tables 4.1 and 4.2 for score distribution of human-scored items). In ELA, approximately 6,000 responses (10% of the total number of test takers in each grade) were scored twice. In Mathematics, a similar total number of responses were also scored by two raters, but half of them were on PT1 forms and half on PT2 forms.

Interrater reliability results for ELA items are shown in Table 7.16. Each item had a maximum possible score of 2. Overall, the agreement rates were moderately high. When the percentage of "No Difference" and "1 Point Difference" were summed, it ranged from 95.9% to 98.7%. The largest percentage of "2 or More Point Difference" was found in grade 6 (3.3%). The percentage of responses with condition codes (e.g., blank, insufficient response, or nonscorable language) ranged from 0.5% to 1.6% across all items. The average intraclass correlation was 0.80, ranging from 0.74 to 0.86. The average weighted kappa was 0.61, ranging from 0.49 to 0.73.

Table 7.17 provides the interrater reliability results for Mathematics items. Among 35 items, nine items had a maximum possible score of 1, 24 items had a maximum possible score of 2, and two items had a maximum possible score of 3. Overall, the agreement rates were high and the summed percentage of "No Difference" and "1 Point Difference" ranged from 92.8% to 99.9%. The largest percentage of "2 or More Point Difference" was found for grade 5 (5.6%). The percentage of responses with condition codes ranged from 0.1% to 5.4% across all items. The average intraclass correlation was 0.89, ranging from 0.70 to 0.98. The average weighted kappa was 0.78, ranging from 0.41 to 0.96. The unweighted kappa was used for nine items because the maximum score point for these items was 1.

7.4: Classification Consistency and Accuracy

The accuracy of decisions (classifications) based on specified cut scores for the Badger Exam was assessed as a measure of the reliability of performance level classifications. Every discrete test administration will result in some error in the classification of test takers. When an assessment uses performance classifications as the primary method to report test results, accuracy and consistency of decisions become important indicators about the quality of the assessment.

The methodology used for estimating the reliability of classification and decision accuracy is described in Livingston and Lewis (1995) and is implemented using the computer program RELCLASS (Version 4.12). RELCLASS provides two statistics (i.e., classification consistency and accuracy) that estimate the reliability of classifications based on test scores.

The accuracy of classifications is represented by the agreement between the classifications based on test takers' observed scores on the actual test form and the classifications that would have been made based on test takers' true scores. Classification consistency is the agreement between two nonoverlapping and equally difficult forms of the test.

For each Badger Exam, the classification consistency and accuracy table includes estimates of the proportion of:

- overall consistent and accurate classifications
- consistency and accuracy around all cut points

A classification consistency table is a cross tabulation of the observed score by hypothetical alternate form score classifications. A classification accuracy table is a cross tabulation of the true score by observed score classifications. The cells of the cross tabulation tables contain the proportions of cases falling in the cells. The proportion of overall consistency and accuracy classifications is estimated as the sum of the diagonal cell entries (proportional agreement between observed and hypothetical alternate form score decisions for consistency; proportional agreement between observed and true score decisions for accuracy).

Consistency and accuracy classifications around all cut points (e.g., “Below Basic” and “Basic” versus “Proficient” and “Advanced”) are similarly computed by collapsing all classification decisions into a dichotomized distribution around the specified cut point.

Note that in all cases the classification accuracy indices are somewhat higher than the classification consistency indices. This is due to differences in the quantities being estimated. Classification consistency is the agreement between two classifications that are both affected by random variation in the selection of test items. Classification accuracy is the agreement between one classification that is affected by random variation in the selection of test items and another classification that is not (Livingston & Lewis, 1995).

Tables 7.18 through 7.29 provide two tables for each grade and subject area. The first table is a 4 x 4 contingency table with all three cut scores. In the first table, the rows represent the observed test takers’ test results, and the columns represent the simulated parallel test results using estimated test reliability. The value in each cell represents the probability of belonging to a particular pair of performance levels in the observed and simulated performance level classification results. For example, in Table 7.18 (ELA grade 3), 0.17 represents the probability of belonging to “Below Basic” in both the observed and simulated results. “Total” is obtained simply by adding the four row values or the four column values. The “Total” value 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 provides a summary of indices for classification consistency and accuracy. Because there are four performance levels for the Badger Exam, there are three cut scores. The values in “Agreement on All Categories” were obtained by applying all three cuts together. For example, in Table 7.18 (ELA grade 3), when all three cuts were used for the computation, classification consistency is 0.66 and classification accuracy is 0.75 for “Agreement on All Categories.” The indices for “Cut 1,” “Cut 2,” or “Cut 3” were obtained by applying only the specified one 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 indices for all three cuts are smaller than those for any single cut point. The probability of assigning test takers to the incorrect performance level commonly increases with the number of cut scores.

Although classification accuracy indices were provided for each cut score, the results from the second cut is more important because the “Proficient” cut score is a criterion for school accountability reports. In ELA, the classification consistency of the “Proficient” cut score ranged

from 0.84 to 0.87 (mean of 0.85), and the classification accuracy ranged from 0.88 to 0.90 (mean of 0.89). For Mathematics for the “Proficient” cut score, the classification consistency ranged from 0.85 to 0.88 (mean of 0.86), and the classification accuracy ranged from 0.88 to 0.90 (mean of 0.90). When two Mathematics forms classified by PTs were compared, the classification consistency and accuracy were quite similar.

Table 7.1: Test Reliability Estimates for Total Group and Subgroups, ELA Grade 3

	Number of Test Takers	Max. Possible Raw Score	Reliability Estimate	Raw Score *SEM*
Total	61,525	46	0.90	2.85
Gender				
Male	31,437	46	0.90	2.83
Female	30,088	46	0.90	2.85
Race/Ethnicity				
White	41,391	46	0.89	2.87
African American	6,899	46	0.86	2.69
Hispanic/Latino	8,003	46	0.88	2.81
Asian	2,371	46	0.90	2.83
American Indian	838	46	0.87	2.80
Pacific Islander	47	46	0.87	2.91
Two or More Races	1,976	46	0.90	2.84
ELP				
Fully English Proficient	55,924	46	0.90	2.85
Limited English Proficient	5,601	46	0.85	2.78
Disability				
Disabled	7,161	46	0.89	2.75
Not Disabled	54,364	46	0.90	2.86
SES				
Economically Disadvantaged	29,352	46	0.88	2.80
Non-economically Disadvantaged	32,173	46	0.89	2.88

Note. ELP=English language proficiency; SES=Socioeconomic status

Table 7.2: Test Reliability Estimates for Total Group and Subgroups, ELA Grade 4

	Number of Test Takers	Max. Possible Raw Score	Reliability Estimate	Raw Score *SEM*
Total	61,121	49	0.88	2.96
Gender				
Male	31,180	49	0.87	2.93
Female	29,941	49	0.88	2.98
Race/Ethnicity				
White	42,143	49	0.86	3.00
African American	6,683	49	0.85	2.72
Hispanic/Latino	7,362	49	0.85	2.88
Asian	2,329	49	0.88	2.99
American Indian	835	49	0.84	2.86
Pacific Islander	65	49	0.85	3.00
Two or More Races	1,704	49	0.87	2.95
ELP				
Fully English Proficient	56,600	49	0.88	2.97
Limited English Proficient	4,521	49	0.81	2.82
Disability				
Disabled	7,091	49	0.86	2.76
Not Disabled	54,030	49	0.87	2.98
SES				
Economically Disadvantaged	27,915	49	0.86	2.87
Non-economically Disadvantaged	33,206	49	0.86	3.02

Note. ELP=English language proficiency; SES=Socioeconomic status

Table 7.3: Test Reliability Estimates for Total Group and Subgroups, ELA Grade 5

	Number of Test Takers	Max. Possible Raw Score	Reliability Estimate	Raw Score *SEM*
Total	61,600	47	0.91	2.88
Gender				
Male	31,556	47	0.91	2.87
Female	30,044	47	0.90	2.89
Race/Ethnicity				
White	42,944	47	0.89	2.90
African American	6,522	47	0.88	2.76
Hispanic/Latino	7,304	47	0.89	2.86
Asian	2,317	47	0.91	2.89
American Indian	824	47	0.89	2.85
Pacific Islander	40	47	0.91	2.87
Two or More Races	1,649	47	0.91	2.89
ELP				
Fully English Proficient	58,352	47	0.90	2.89
Limited English Proficient	3,248	47	0.82	2.77
Disability				
Disabled	7,439	47	0.89	2.73
Not Disabled	54,161	47	0.90	2.90
SES				
Economically Disadvantaged	27,420	47	0.89	2.85
Non-economically Disadvantaged	34,180	47	0.89	2.90

Note. ELP=English language proficiency; SES=Socioeconomic status

Table 7.4: Test Reliability Estimates for Total Group and Subgroups, ELA Grade 6

	Number of Test Takers	Max. Possible Raw Score	Reliability Estimate	Raw Score *SEM*
Total	60,859	50	0.87	3.07
Gender				
Male	31,125	50	0.87	3.03
Female	29,734	50	0.87	3.11
Race/Ethnicity				
White	42,833	50	0.86	3.11
African American	6,368	50	0.82	2.86
Hispanic/Latino	7,131	50	0.84	3.00
Asian	2,185	50	0.88	3.09
American Indian	796	50	0.84	2.98
Pacific Islander	47	50	0.86	3.06
Two or More Races	1,499	50	0.87	3.06
ELP				
Fully English Proficient	58,250	50	0.87	3.08
Limited English Proficient	2,609	50	0.73	2.82
Disability				
Disabled	7,171	50	0.82	2.81
Not Disabled	53,688	50	0.86	3.10
SES				
Economically Disadvantaged	26,430	50	0.85	2.99
Non-economically Disadvantaged	34,429	50	0.85	3.13

Note. ELP=English language proficiency; SES=Socioeconomic status

Table 7.5: Test Reliability Estimates for Total Group and Subgroups, ELA Grade 7

	Number of Test Takers	Max. Possible Raw Score	Reliability Estimate	Raw Score *SEM*
Total	60,590	50	0.88	3.05
Gender				
Male	30,829	50	0.88	3.01
Female	29,761	50	0.87	3.07
Race/Ethnicity				
White	42,926	50	0.87	3.07
African American	6,328	50	0.86	2.91
Hispanic/Latino	6,886	50	0.86	3.01
Asian	2,127	50	0.88	3.06
American Indian	812	50	0.86	2.98
Pacific Islander	37	50	0.80	3.10
Two or More Races	1,474	50	0.87	3.05
ELP				
Fully English Proficient	57,900	50	0.88	3.05
Limited English Proficient	2,690	50	0.79	2.90
Disability				
Disabled	7,178	50	0.84	2.84
Not Disabled	53,412	50	0.86	3.07
SES				
Economically Disadvantaged	25,970	50	0.86	3.00
Non-economically Disadvantaged	34,620	50	0.86	3.07

Note. ELP=English language proficiency; SES=Socioeconomic status

Table 7.6: Test Reliability Estimates for Total Group and Subgroups, ELA Grade 8

	Number of Test Takers	Max. Possible Raw Score	Reliability Estimate	Raw Score *SEM*
Total	60,842	50	0.87	3.09
Gender				
Male	31,203	50	0.87	3.04
Female	29,639	50	0.86	3.12
Race/Ethnicity				
White	43,495	50	0.86	3.11
African American	6,257	50	0.83	2.94
Hispanic/Latino	6,753	50	0.84	3.05
Asian	2,155	50	0.87	3.14
American Indian	786	50	0.84	3.03
Pacific Islander	38	50	0.86	3.15
Two or More Races	1,358	50	0.87	3.08
ELP				
Fully English Proficient	57,972	50	0.87	3.10
Limited English Proficient	2,870	50	0.75	2.93
Disability				
Disabled	7,231	50	0.83	2.81
Not Disabled	53,611	50	0.86	3.12
SES				
Economically Disadvantaged	25,309	50	0.85	3.03
Non-economically Disadvantaged	35,533	50	0.86	3.12

Note. ELP=English language proficiency; SES=Socioeconomic status

Table 7.7: Test Reliability Estimates for Total Group and Subgroups, Mathematics Grade 3

Mathematics (PT1)	Number of Test Takers	Max. Possible Raw Score	Reliability Estimate	Raw Score *SEM*
Total	30,484	37	0.90	2.45
Gender				
Male	15,625	37	0.91	2.44
Female	14,859	37	0.90	2.47
Race/Ethnicity				
White	20,567	37	0.89	2.46
African American	3,477	37	0.88	2.37
Hispanic/Latino	3,958	37	0.88	2.46
Asian	1,180	37	0.91	2.43
American Indian	299	37	0.89	2.45
Pacific Islander	18	37	0.90	2.48
Two or More Races	985	37	0.90	2.46
ELP				
Fully English Proficient	27,704	37	0.90	2.45
Limited English Proficient	2,780	37	0.88	2.44
Disability				
Disabled	3,530	37	0.91	2.40
Not Disabled	26,954	37	0.90	2.46
SES				
Economically Disadvantaged	14,576	37	0.89	2.45
Non-economically Disadvantaged	15,908	37	0.89	2.45

Note. ELP=English language proficiency; SES=Socioeconomic status

Table 7.7: Test Reliability Estimates for Total Group and Subgroups, Mathematics Grade 3 (Cont'd)

Mathematics (PT2)	Number of Test Takers	Max. Possible Raw Score	Reliability Estimate	Raw Score *SEM*
Total	31,087	41	0.91	2.57
Gender				
Male	15,845	41	0.92	2.56
Female	15,242	41	0.91	2.59
Race/Ethnicity				
White	20,814	41	0.90	2.55
African American	3,378	41	0.90	2.53
Hispanic/Latino	4,104	41	0.90	2.61
Asian	1,227	41	0.92	2.55
American Indian	540	41	0.89	2.60
Pacific Islander	31	41	0.92	2.57
Two or More Races	993	41	0.91	2.58
ELP				
Fully English Proficient	28,152	41	0.91	2.57
Limited English Proficient	2,935	41	0.89	2.59
Disability				
Disabled	3,621	41	0.92	2.58
Not Disabled	27,466	41	0.91	2.57
SES				
Economically Disadvantaged	14,798	41	0.90	2.60
Non-economically Disadvantaged	16,289	41	0.89	2.52

Note. ELP=English language proficiency; SES=Socioeconomic status

Table 7.8: Test Reliability Estimates for Total Group and Subgroups, Mathematics Grade 4

Mathematics (PT1)	Number of Test Takers	Max. Possible Raw Score	Reliability Estimate	Raw Score *SEM*
Total	30,053	39	0.91	2.56
Gender				
Male	15,298	39	0.91	2.55
Female	14,755	39	0.90	2.56
Race/Ethnicity				
White	20,210	39	0.89	2.63
African American	3,531	39	0.87	2.20
Hispanic/Latino	3,811	39	0.88	2.40
Asian	1,187	39	0.92	2.53
American Indian	408	39	0.86	2.45
Pacific Islander	32	39	0.91	2.64
Two or More Races	874	39	0.90	2.51
ELP				
Fully English Proficient	27,753	39	0.90	2.57
Limited English Proficient	2,300	39	0.86	2.30
Disability				
Disabled	3,422	39	0.91	2.34
Not Disabled	26,631	39	0.90	2.58
SES				
Economically Disadvantaged	13,914	39	0.89	2.42
Non-economically Disadvantaged	16,139	39	0.89	2.65

Note. ELP=English language proficiency; SES=Socioeconomic status

Table 7.8: Test Reliability Estimates for Total Group and Subgroups, Mathematics Grade 4 (Cont'd)

Mathematics (PT2)	Number of Test Takers	Max. Possible Raw Score	Reliability Estimate	Raw Score *SEM*
Total	31,095	39	0.91	2.54
Gender				
Male	15,893	39	0.91	2.54
Female	15,202	39	0.90	2.55
Race/Ethnicity				
White	21,924	39	0.90	2.59
African American	3,105	39	0.89	2.23
Hispanic/Latino	3,613	39	0.89	2.45
Asian	1,167	39	0.92	2.53
American Indian	426	39	0.88	2.44
Pacific Islander	33	39	0.91	2.46
Two or More Races	827	39	0.91	2.52
ELP				
Fully English Proficient	28,766	39	0.91	2.55
Limited English Proficient	2,329	39	0.87	2.37
Disability				
Disabled	3,647	39	0.91	2.33
Not Disabled	27,448	39	0.90	2.57
SES				
Economically Disadvantaged	14,016	39	0.89	2.44
Non-economically Disadvantaged	17,079	39	0.89	2.60

Note. ELP=English language proficiency; SES=Socioeconomic status

Table 7.9: Test Reliability Estimates for Total Group and Subgroups, Mathematics Grade 5

Mathematics (PT1)	Number of Test Takers	Max. Possible Raw Score	Reliability Estimate	Raw Score *SEM*
Total	30,155	41	0.91	2.38
Gender				
Male	15,389	41	0.92	2.37
Female	14,766	41	0.91	2.38
Race/Ethnicity				
White	20,488	41	0.90	2.44
African American	3,483	41	0.85	2.07
Hispanic/Latino	3,827	41	0.87	2.24
Asian	1,172	41	0.92	2.38
American Indian	361	41	0.87	2.21
Pacific Islander	24	41	0.87	2.54
Two or More Races	800	41	0.92	2.34
ELP				
Fully English Proficient	28,391	41	0.91	2.39
Limited English Proficient	1,764	41	0.81	2.12
Disability				
Disabled	3,642	41	0.89	2.13
Not Disabled	26,513	41	0.91	2.40
SES				
Economically Disadvantaged	13,784	41	0.88	2.27
Non-economically Disadvantaged	16,371	41	0.91	2.46

Note. ELP=English language proficiency; SES=Socioeconomic status

Table 7.9: Test Reliability Estimates for Total Group and Subgroups, Mathematics Grade 5 (Cont'd)

Mathematics (PT2)	Number of Test Takers	Max. Possible Raw Score	Reliability Estimate	Raw Score *SEM*
Total	31,442	41	0.91	2.58
Gender				
Male	16,160	41	0.91	2.57
Female	15,282	41	0.90	2.58
Race/Ethnicity				
White	22,431	41	0.90	2.62
African American	2,996	41	0.83	2.26
Hispanic/Latino	3,521	41	0.87	2.46
Asian	1,175	41	0.92	2.57
American Indian	458	41	0.88	2.44
Pacific Islander	16	41	0.90	2.60
Two or More Races	845	41	0.90	2.55
ELP				
Fully English Proficient	29,865	41	0.91	2.59
Limited English Proficient	1,577	41	0.81	2.29
Disability				
Disabled	3,767	41	0.89	2.33
Not Disabled	27,675	41	0.90	2.60
SES				
Economically Disadvantaged	13,622	41	0.88	2.47
Non-economically Disadvantaged	17,820	41	0.90	2.64

Note. ELP=English language proficiency; SES=Socioeconomic status

Table 7.10: Test Reliability Estimates for Total Group and Subgroups, Mathematics Grade 6

Mathematics (PT1)	Number of Test Takers	Max. Possible Raw Score	Reliability Estimate	Raw Score *SEM*
Total	30,167	38	0.91	2.43
Gender				
Male	15,461	38	0.92	2.42
Female	14,706	38	0.91	2.44
Race/Ethnicity				
White	21,807	38	0.90	2.51
African American	2,807	38	0.83	2.00
Hispanic/Latino	3,296	38	0.88	2.22
Asian	1,101	38	0.92	2.40
American Indian	392	38	0.86	2.20
Pacific Islander	18	38	0.89	2.49
Two or More Races	746	38	0.91	2.33
ELP				
Fully English Proficient	28,883	38	0.91	2.45
Limited English Proficient	1,284	38	0.81	1.98
Disability				
Disabled	3,556	38	0.89	2.05
Not Disabled	26,611	38	0.91	2.47
SES				
Economically Disadvantaged	12,901	38	0.88	2.24
Non-economically Disadvantaged	17,266	38	0.90	2.55

Note. ELP=English language proficiency; SES=Socioeconomic status

Table 7.10: Test Reliability Estimates for Total Group and Subgroups, Mathematics Grade 6 (Cont'd)

Mathematics (PT2)	Number of Test Takers	Max. Possible Raw Score	Reliability Estimate	Raw Score *SEM*
Total	30,523	39	0.91	2.52
Gender				
Male	15,575	39	0.91	2.52
Female	14,948	39	0.91	2.52
Race/Ethnicity				
White	20,980	39	0.90	2.58
African American	3,478	39	0.85	2.19
Hispanic/Latino	3,778	39	0.88	2.39
Asian	1,111	39	0.93	2.51
American Indian	397	39	0.86	2.39
Pacific Islander	29	39	0.87	2.50
Two or More Races	750	39	0.91	2.52
ELP				
Fully English Proficient	29,159	39	0.91	2.54
Limited English Proficient	1,364	39	0.82	2.18
Disability				
Disabled	3,545	39	0.89	2.20
Not Disabled	26,978	39	0.90	2.55
SES				
Economically Disadvantaged	13,356	39	0.89	2.39
Non-economically Disadvantaged	17,167	39	0.90	2.60

Note. ELP=English language proficiency; SES=Socioeconomic status

Table 7.11: Test Reliability Estimates for Total Group and Subgroups, Mathematics Grade 7

Mathematics (PT1)	Number of Test Takers	Max. Possible Raw Score	Reliability Estimate	Raw Score *SEM*
Total	30,386	39	0.91	2.34
Gender				
Male	15,552	39	0.92	2.32
Female	14,834	39	0.91	2.35
Race/Ethnicity				
White	21,740	39	0.90	2.41
African American	3,307	39	0.84	1.91
Hispanic/Latino	3,238	39	0.88	2.14
Asian	1,008	39	0.93	2.38
American Indian	354	39	0.88	2.18
Pacific Islander	18	39	0.89	2.35
Two or More Races	721	39	0.92	2.28
ELP				
Fully English Proficient	29,065	39	0.91	2.35
Limited English Proficient	1,321	39	0.81	1.96
Disability				
Disabled	3,626	39	0.87	1.91
Not Disabled	26,760	39	0.91	2.38
SES				
Economically Disadvantaged	12,584	39	0.88	2.15
Non-economically Disadvantaged	17,802	39	0.91	2.44

Note. ELP=English language proficiency; SES=Socioeconomic status

Table 7.11: Test Reliability Estimates for Total Group and Subgroups, Mathematics Grade 7 (Cont'd)

Mathematics (PT2)	Number of Test Takers	Max. Possible Raw Score	Reliability Estimate	Raw Score *SEM*
Total	29,982	36	0.91	2.21
Gender				
Male	15,158	36	0.91	2.20
Female	14,824	36	0.90	2.22
Race/Ethnicity				
White	21,130	36	0.90	2.27
African American	2,901	36	0.84	1.87
Hispanic/Latino	3,600	36	0.88	2.07
Asian	1,123	36	0.92	2.24
American Indian	458	36	0.87	1.98
Pacific Islander	19	36	0.91	2.30
Two or More Races	751	36	0.90	2.15
ELP				
Fully English Proficient	28,595	36	0.91	2.22
Limited English Proficient	1,387	36	0.83	1.85
Disability				
Disabled	3,497	36	0.86	1.83
Not Disabled	26,485	36	0.90	2.25
SES				
Economically Disadvantaged	13,192	36	0.88	2.07
Non-economically Disadvantaged	16,790	36	0.90	2.30

Note. ELP=English language proficiency; SES=Socioeconomic status

Table 7.12: Test Reliability Estimates for Total Group and Subgroups, Mathematics Grade 8

Mathematics (PT1)	Number of Test Takers	Max. Possible Raw Score	Reliability Estimate	Raw Score *SEM*
Total	28,275	36	0.89	2.37
Gender				
Male	14,497	36	0.89	2.36
Female	13,778	36	0.88	2.39
Race/Ethnicity				
White	19,779	36	0.88	2.44
African American	3,159	36	0.78	2.01
Hispanic/Latino	3,293	36	0.84	2.23
Asian	1,124	36	0.90	2.43
American Indian	314	36	0.83	2.26
Pacific Islander	17	36	0.92	2.39
Two or More Races	589	36	0.88	2.34
ELP				
Fully English Proficient	26,745	36	0.89	2.39
Limited English Proficient	1,530	36	0.78	2.07
Disability				
Disabled	3,345	36	0.81	1.97
Not Disabled	24,930	36	0.88	2.42
SES				
Economically Disadvantaged	12,075	36	0.84	2.22
Non-economically Disadvantaged	16,200	36	0.89	2.47

Note. ELP=English language proficiency; SES=Socioeconomic status

Table 7.12: Test Reliability Estimates for Total Group and Subgroups, Mathematics Grade 8 (Cont'd)

Mathematics (PT2)	Number of Test Takers	Max. Possible Raw Score	Reliability Estimate	Raw Score *SEM*
Total	32,389	37	0.89	2.34
Gender				
Male	16,601	37	0.89	2.33
Female	15,788	37	0.88	2.34
Race/Ethnicity				
White	23,661	37	0.88	2.40
African American	2,978	37	0.79	1.97
Hispanic/Latino	3,449	37	0.83	2.18
Asian	1,054	37	0.91	2.35
American Indian	470	37	0.81	2.10
Pacific Islander	20	37	0.85	2.48
Two or More Races	757	37	0.88	2.29
ELP				
Fully English Proficient	30,977	37	0.88	2.35
Limited English Proficient	1,412	37	0.76	2.03
Disability				
Disabled	3,823	37	0.83	1.96
Not Disabled	28,566	37	0.88	2.37
SES				
Economically Disadvantaged	13,087	37	0.85	2.19
Non-economically Disadvantaged	19,302	37	0.88	2.42

Note. ELP=English language proficiency; SES=Socioeconomic status

Table 7.13: Test and Claim Level Reliability Estimates for ELA

Grade	Test/Claim	Max. Possible Raw Score	Reliability Estimate	Raw Score *SEM*
3	Total Test	46	0.90	2.85
	Claim 1	18	0.79	1.74
	Claim 2	11	0.69	1.39
	Claim 3	9	0.66	1.24
	Claim 4	8	0.61	1.23
4	Total Test	49	0.88	2.96
	Claim 1	21	0.75	1.90
	Claim 2	11	0.57	1.49
	Claim 3	9	0.72	1.23
	Claim 4	8	0.51	1.15
5	Total Test	47	0.91	2.88
	Claim 1	20	0.82	1.86
	Claim 2	10	0.68	1.37
	Claim 3	9	0.64	1.26
	Claim 4	8	0.63	1.14
6	Total Test	50	0.87	3.07
	Claim 1	21	0.79	1.90
	Claim 2	12	0.52	1.56
	Claim 3	9	0.56	1.27
	Claim 4	8	0.55	1.30
7	Total Test	50	0.88	3.05
	Claim 1	21	0.75	1.95
	Claim 2	12	0.63	1.48
	Claim 3	9	0.62	1.32
	Claim 4	8	0.57	1.19
8	Total Test	50	0.87	3.09
	Claim 1	21	0.79	1.96
	Claim 2	12	0.59	1.53
	Claim 3	9	0.50	1.28
	Claim 4	8	0.52	1.27

Table 7.14: Test and Claim Level Reliability Estimates for Mathematics Grades 3–5

Grade	PT Set	Test/Claim	Max. Possible Raw Score	Reliability Estimate	Raw Score *SEM*
3	PT1	Total Test	37	0.90	2.45
		Claim 1	20	0.86	1.66
		Claim 2&4	8	0.70	1.16
		Claim 3	9	0.65	1.35
	PT2	Total Test	41	0.91	2.57
		Claim 1	20	0.86	1.66
		Claim 2&4	14	0.73	1.61
		Claim 3	7	0.71	1.09
4	PT1	Total Test	39	0.91	2.56
		Claim 1	20	0.85	1.71
		Claim 2&4	10	0.67	1.31
		Claim 3	9	0.71	1.34
	PT2	Total Test	39	0.91	2.54
		Claim 1	20	0.84	1.72
		Claim 2&4	10	0.72	1.29
		Claim 3	9	0.70	1.33
5	PT1	Total Test	41	0.91	2.38
		Claim 1	20	0.84	1.71
		Claim 2&4	11	0.79	0.93
		Claim 3	10	0.66	1.45
	PT2	Total Test	41	0.91	2.58
		Claim 1	20	0.84	1.72
		Claim 2&4	12	0.74	1.39
		Claim 3	9	0.67	1.31

Table 7.15: Test and Claim Level Reliability Estimates for Mathematics Grades 6–8

Grade	PT Set	Test/Claim	Max. Possible Raw Score	Reliability Estimate	Raw Score *SEM*
6	PT1	Total Test	38	0.91	2.43
		Claim 1	20	0.86	1.79
		Claim 2&4	8	0.71	1.05
		Claim 3	10	0.69	1.25
	PT2	Total Test	39	0.91	2.52
		Claim 1	20	0.86	1.80
		Claim 2&4	9	0.72	1.17
		Claim 3	10	0.64	1.32
7	PT1	Total Test	39	0.91	2.34
		Claim 1	21	0.86	1.74
		Claim 2&4	8	0.64	1.02
		Claim 3	10	0.71	1.16
	PT2	Total Test	36	0.91	2.21
		Claim 1	21	0.86	1.73
		Claim 2&4	8	0.66	1.01
		Claim 3	7	0.64	0.91
8	PT1	Total Test	36	0.89	2.37
		Claim 1	20	0.82	1.83
		Claim 2&4	7	0.67	0.92
		Claim 3	9	0.58	1.19
	PT2	Total Test	37	0.89	2.34
		Claim 1	20	0.82	1.85
		Claim 2&4	9	0.72	1.10
		Claim 3	8	0.48	0.93

Table 7.16: Interrater Agreement for ELA Human-Scored Item

Grade	Item Number	Max. Possible Score	Mean Score	Number of Responses Read Twice	Percentage Absolute Difference				Intraclass Correlation	Weighted Kappa
					No Difference	1 Point	2 or more Points	Codes		
3	11	2	0.93	5,908	67.4	29.8	1.2	1.6	0.86	0.71
4	14	2	0.91	6,024	63.0	34.0	2.0	1.0	0.83	0.65
	35	2	0.42	5,958	69.1	27.4	2.0	1.5	0.76	0.53
5	27	2	1.01	6,032	62.9	34.6	1.7	0.8	0.81	0.62
	31	2	0.76	6,014	64.3	33.6	1.3	0.9	0.80	0.60
6	13	2	0.59	5,939	67.4	28.5	3.2	0.9	0.81	0.62
	16	2	0.57	5,900	61.5	34.5	3.3	0.7	0.74	0.49
7	5	2	0.77	5,942	63.6	33.6	2.2	0.6	0.79	0.58
	12	2	0.64	5,867	61.5	35.9	2.0	0.6	0.76	0.53
	21	2	0.53	5,847	73.9	24.2	1.1	0.8	0.85	0.70
8	5	2	0.77	5,919	60.3	36.4	2.7	0.6	0.75	0.51
	11	2	0.74	5,867	71.9	26.8	0.7	0.5	0.86	0.73
	26	2	0.84	5,779	65.6	31.4	1.5	1.5	0.84	0.67

Table 7.17: Interrater Agreement for Mathematics Human-Scored Item

Grade	Item Number	Max. Possible Score	Mean Score	Number of Responses Read Twice	Percentage Absolute Difference				Intraclass Correlation	Weighted Kappa
					No Difference	1 Point	2 or more Points	Codes		
3	35 (PT1)	2	0.56	2,996	75.6	18.1	5.2	1.0	0.85	0.70
	34 (PT2)[a]	2								
	35 (PT2)	3	1.98	3,099	89.4	9.6	0.7	0.3	0.97	0.95
	36 (PT2)	1	0.55	3,050	88.2	10.7		1.0	0.89	0.78
	37 (PT2)	2	0.54	2,970	84.1	12.3	1.1	2.5	0.93	0.85
4	34 (PT1)	1	0.31	2,987	92.1	7.3		0.6	0.91	0.83
	35 (PT1)	2	0.46	2,945	90.0	8.9	0.4	0.6	0.95	0.91
	36 (PT1)	2	0.68	2,953	82.4	12.8	3.8	1.0	0.91	0.82
	34 (PT2)	1	0.35	3,064	82.4	17.3		0.4	0.81	0.62
	35 (PT2)	2	0.66	3,060	88.8	8.8	1.3	1.1	0.95	0.90
	36 (PT2)	1	0.51	3,081	94.9	4.0		1.1	0.96	0.91
	37 (PT2)	2	0.51	3,043	89.8	7.6	1.5	1.1	0.94	0.89
5	34 (PT1)	2	0.56	2,971	87.3	11.1	0.8	0.8	0.94	0.88
	35 (PT1)	2	0.27	2,934	83.4	11.5	2.6	2.4	0.86	0.71
	36 (PT1)	2	0.46	3,005	83.9	12.3	1.4	2.5	0.92	0.83
	37 (PT1)	2	0.76	2,921	81.2	11.6	5.6	1.6	0.90	0.80
	34 (PT2)	1	0.33	3,095	87.2	12.4		0.3	0.86	0.72
	35 (PT2)	2	1.08	3,091	79.2	19.8	0.4	0.6	0.92	0.84
	36 (PT2)	2	0.47	3,094	84.6	12.9	1.8	0.7	0.91	0.82
	37 (PT2)	3	0.26	3,060	80.0	14.2	4.3	1.4	0.78	0.56

[a]On Grade 3 item 34 (PT2) all responses that were not blank received full credit due to a Spanish translation issue.

Table 7.17: Interrater Agreement for Mathematics Human-Scored Item (Cont'd)

Grade	Item Number	Max. Possible Score	Mean Score	Number of Responses Read Twice	Percentage Absolute Difference				Intraclass Correlation	Weighted Kappa
					No Difference	1 Point	2 or more Points	Codes		
6	32 (PT1)	2	0.53	2,982	81.3	17.5	0.9	0.3	0.91	0.82
	33 (PT1)	2	0.55	2,983	75.1	22.2	2.2	0.5	0.87	0.74
	34 (PT1)	1	0.21	3,006	96.2	3.7		0.1	0.94	0.89
	35 (PT1)	2	0.27	3,001	85.7	11.0	2.7	0.6	0.86	0.76
	33 (PT2)	2	0.84	3,031	87.0	11.8	0.6	0.5	0.95	0.89
	34 (PT2)	1	0.29	3,055	97.7	1.4		0.8	0.98	0.96
	35 (PT2)	2	1.27	3,013	82.0	17.1	0.4	0.6	0.93	0.87
	36 (PT2)	2	0.37	2,966	75.8	19.6	3.0	1.6	0.81	0.63
7	21	2	0.39	5,934	83.1	15.0	1.2	0.7	0.87	0.74
	34 (PT1)	2	0.16	2,899	85.0	10.6	2.9	1.6	0.76	0.52
	35 (PT1)	2	0.75	2,971	73.8	24.7	0.3	1.2	0.89	0.78
	34 (PT2)	1	0.12	2,966	87.0	12.4		0.6	0.70	0.41
8	34 (PT1)	2	0.31	2,792	80.3	18.1	0.1	1.5	0.87	0.74
	35 (PT1)	1	0.34	2,780	76.7	22.7		0.6	0.75	0.49
	35 (PT2)	2	0.29	3,158	82.0	11.8	0.7	5.4	0.89	0.79

Table 7.18: Classification Consistency and Accuracy for ELA Grade 3

Contingency Table with All Cut Scores

	Below Basic	Basic	Proficient	Advanced	Total
Below Basic	0.17	0.05	0.00	0.00	0.22
Basic	0.05	0.14	0.06	0.00	0.25
Proficient	0.01	0.06	0.13	0.06	0.26
Advanced	0.00	0.00	0.05	0.22	0.27
Total	0.23	0.25	0.24	0.28	

Indexes for Classification Consistency and Accuracy

Cut Score Level	Classification Consistency	Classification Accuracy
Cut 1	0.89	0.92
Cut 2	0.86	0.90
Cut 3	0.89	0.92
Agreement on All Categories	0.66	0.75

Note. Cut 1=Below Basic vs. Basic, Proficient, or Advanced; Cut 2=Below Basic or Basic vs. Proficient or Advanced; Cut 3=Below Basic, Basic, or Proficient vs. Advanced.

Table 7.19: Classification Consistency and Accuracy for ELA Grade 4

Contingency Table with All Cut Scores

	Below Basic	Basic	Proficient	Advanced	Total
Below Basic	0.22	0.05	0.01	0.00	0.28
Basic	0.05	0.09	0.06	0.01	0.21
Proficient	0.01	0.06	0.12	0.06	0.25
Advanced	0.00	0.00	0.05	0.20	0.25
Total	0.28	0.21	0.24	0.27	

Indices for Classification Consistency and Accuracy

Cut Score Level	Classification Consistency	Classification Accuracy
Cut 1	0.87	0.91
Cut 2	0.85	0.89
Cut 3	0.87	0.91
Agreement on All Categories	0.62	0.71

Note. Cut 1=Below Basic vs. Basic, Proficient, or Advanced; Cut 2=Below Basic or Basic vs. Proficient or Advanced; Cut 3=Below Basic, Basic, or Proficient vs. Advanced.

Table 7.20: Classification Consistency and Accuracy for ELA Grade 5

Contingency Table with All Cut Scores

	Below Basic	Basic	Proficient	Advanced	Total
Below Basic	0.19	0.05	0.00	0.00	0.24
Basic	0.05	0.10	0.06	0.00	0.21
Proficient	0.01	0.06	0.19	0.06	0.31
Advanced	0.00	0.00	0.04	0.19	0.24
Total	0.24	0.21	0.29	0.25	

Indices for Classification Consistency and Accuracy

Cut Score Level	Classification Consistency	Classification Accuracy
Cut 1	0.89	0.92
Cut 2	0.87	0.90
Cut 3	0.90	0.93
Agreement on All Categories	0.67	0.76

Note. Cut 1=Below Basic vs. Basic, Proficient, or Advanced; Cut 2=Below Basic or Basic vs. Proficient or Advanced; Cut 3=Below Basic, Basic, or Proficient vs. Advanced.

Table 7.21: Classification Consistency and Accuracy for ELA Grade 6

Contingency Table with All Cut Scores

	Below Basic	Basic	Proficient	Advanced	Total
Below Basic	0.21	0.06	0.01	0.00	0.27
Basic	0.06	0.13	0.07	0.00	0.26
Proficient	0.01	0.07	0.18	0.05	0.30
Advanced	0.00	0.00	0.04	0.12	0.16
Total	0.27	0.26	0.29	0.18	

Indices for Classification Consistency and Accuracy

Cut Score Level	Classification Consistency	Classification Accuracy
Cut 1	0.87	0.91
Cut 2	0.85	0.89
Cut 3	0.90	0.93
Agreement on All Categories	0.63	0.73

Note. Cut 1=Below Basic vs. Basic, Proficient, or Advanced; Cut 2=Below Basic or Basic vs. Proficient or Advanced; Cut 3=Below Basic, Basic, or Proficient vs. Advanced.

Table 7.22: Classification Consistency and Accuracy for ELA Grade 7

Contingency Table with All Cut Scores

	Below Basic	Basic	Proficient	Advanced	Total
Below Basic	0.18	0.05	0.00	0.00	0.23
Basic	0.06	0.12	0.07	0.00	0.25
Proficient	0.01	0.07	0.22	0.06	0.35
Advanced	0.00	0.00	0.05	0.13	0.18
Total	0.24	0.23	0.34	0.19	

Indices for Classification Consistency and Accuracy

Cut Score Level	Classification Consistency	Classification Accuracy
Cut 1	0.89	0.92
Cut 2	0.85	0.89
Cut 3	0.89	0.93
Agreement on All Categories	0.64	0.73

Note. Cut 1=Below Basic vs. Basic, Proficient, or Advanced; Cut 2=Below Basic or Basic vs. Proficient or Advanced; Cut 3=Below Basic, Basic, or Proficient vs. Advanced.

Table 7.23: Classification Consistency and Accuracy for ELA Grade 8

Contingency Table with All Cut Scores

	Below Basic	Basic	Proficient	Advanced	Total
Below Basic	0.15	0.05	0.00	0.00	0.20
Basic	0.05	0.13	0.07	0.00	0.25
Proficient	0.01	0.07	0.24	0.06	0.39
Advanced	0.00	0.00	0.05	0.11	0.16
Total	0.21	0.25	0.37	0.18	

Indices for Classification Consistency and Accuracy

Cut Score Level	Classification Consistency	Classification Accuracy
Cut 1	0.89	0.93
Cut 2	0.84	0.88
Cut 3	0.89	0.92
Agreement on All Categories	0.63	0.73

Note. Cut 1=Below Basic vs. Basic, Proficient, or Advanced; Cut 2=Below Basic or Basic vs. Proficient or Advanced; Cut 3=Below Basic, Basic, or Proficient vs. Advanced.

Table 7.24: Classification Consistency and Accuracy for Mathematics Grade 3

Contingency Table with All Cut Scores (PT1)

	Below Basic	Basic	Proficient	Advanced	Total
Below Basic	0.20	0.04	0.00	0.00	0.25
Basic	0.06	0.14	0.07	0.01	0.27
Proficient	0.00	0.06	0.18	0.05	0.30
Advanced	0.00	0.00	0.04	0.14	0.18
Total	0.27	0.24	0.29	0.20	

Contingency Table with All Cut Scores (PT2)

	Below Basic	Basic	Proficient	Advanced	Total
Below Basic	0.15	0.04	0.00	0.00	0.19
Basic	0.05	0.14	0.06	0.00	0.25
Proficient	0.01	0.07	0.22	0.06	0.34
Advanced	0.00	0.00	0.04	0.17	0.21
Total	0.21	0.25	0.32	0.23	

Indices for Classification Consistency and Accuracy (PT1)

Cut Score Level	Classification Consistency	Classification Accuracy
Cut 1	0.90	0.92
Cut 2	0.86	0.89
Cut 3	0.90	0.93
Agreement on All Categories	0.66	0.75

Note. Cut 1=Below Basic vs. Basic, Proficient, or Advanced; Cut 2=Below Basic or Basic vs. Proficient or Advanced; Cut 3=Below Basic, Basic, or Proficient vs. Advanced.

Indices for Classification Consistency and Accuracy (PT2)

Cut Score Level	Classification Consistency	Classification Accuracy
Cut 1	0.91	0.93
Cut 2	0.87	0.90
Cut 3	0.91	0.93
Agreement on All Categories	0.68	0.77

Note. Cut 1=Below Basic vs. Basic, Proficient, or Advanced; Cut 2=Below Basic or Basic vs. Proficient or Advanced; Cut 3=Below Basic, Basic, or Proficient vs. Advanced.

Table 7.25: Classification Consistency and Accuracy for Mathematics Grade 4

Contingency Table with All Cut Scores (PT1)

	Below Basic	Basic	Proficient	Advanced	Total
Below Basic	0.17	0.04	0.00	0.00	0.20
Basic	0.06	0.19	0.07	0.01	0.33
Proficient	0.00	0.07	0.18	0.06	0.31
Advanced	0.00	0.00	0.04	0.13	0.16
Total	0.23	0.29	0.29	0.19	

Contingency Table with All Cut Scores (PT2)

	Below Basic	Basic	Proficient	Advanced	Total
Below Basic	0.16	0.03	0.00	0.00	0.19
Basic	0.06	0.19	0.07	0.01	0.32
Proficient	0.00	0.07	0.18	0.06	0.30
Advanced	0.00	0.00	0.04	0.15	0.18
Total	0.22	0.29	0.29	0.21	

Indices for Classification Consistency and Accuracy (PT1)

Cut Score Level	Classification Consistency	Classification Accuracy
Cut 1	0.91	0.94
Cut 2	0.85	0.89
Cut 3	0.90	0.93
Agreement on All Categories	0.67	0.76

Note. Cut 1=Below Basic vs. Basic, Proficient, or Advanced; Cut 2=Below Basic or Basic vs. Proficient or Advanced; Cut 3=Below Basic, Basic, or Proficient vs. Advanced.

Indices for Classification Consistency and Accuracy (PT2)

Cut Score Level	Classification Consistency	Classification Accuracy
Cut 1	0.91	0.94
Cut 2	0.85	0.89
Cut 3	0.90	0.93
Agreement on All Categories	0.67	0.76

Note. Cut 1=Below Basic vs. Basic, Proficient, or Advanced; Cut 2=Below Basic or Basic vs. Proficient or Advanced; Cut 3=Below Basic, Basic, or Proficient vs. Advanced.

Table 7.26: Classification Consistency and Accuracy for Mathematics Grade 5

Contingency Table with All Cut Scores (PT1)

	Below Basic	Basic	Proficient	Advanced	Total
Below Basic	0.26	0.05	0.00	0.00	0.31
Basic	0.06	0.17	0.06	0.01	0.30
Proficient	0.00	0.05	0.09	0.04	0.19
Advanced	0.00	0.00	0.04	0.16	0.20
Total	0.32	0.28	0.19	0.22	

Contingency Table with All Cut Scores (PT2)

	Below Basic	Basic	Proficient	Advanced	Total
Below Basic	0.22	0.05	0.00	0.00	0.27
Basic	0.06	0.17	0.06	0.01	0.31
Proficient	0.00	0.06	0.09	0.05	0.20
Advanced	0.00	0.00	0.04	0.18	0.22
Total	0.29	0.28	0.20	0.24	

Indices for Classification Consistency and Accuracy (PT1)

Cut Score Level	Classification Consistency	Classification Accuracy
Cut 1	0.89	0.92
Cut 2	0.88	0.90
Cut 3	0.91	0.93
Agreement on All Categories	0.68	0.76

Note. Cut 1=Below Basic vs. Basic, Proficient, or Advanced; Cut 2=Below Basic or Basic vs. Proficient or Advanced; Cut 3=Below Basic, Basic, or Proficient vs. Advanced.

Indices for Classification Consistency and Accuracy (PT2)

Cut Score Level	Classification Consistency	Classification Accuracy
Cut 1	0.89	0.92
Cut 2	0.87	0.90
Cut 3	0.90	0.93
Agreement on All Categories	0.67	0.75

Note. Cut 1=Below Basic vs. Basic, Proficient, or Advanced; Cut 2=Below Basic or Basic vs. Proficient or Advanced; Cut 3=Below Basic, Basic, or Proficient vs. Advanced.

Table 7.27: Classification Consistency and Accuracy for Mathematics Grade 6

Contingency Table with All Cut Scores (PT1)

	Below Basic	Basic	Proficient	Advanced	Total
Below Basic	0.24	0.05	0.00	0.00	0.28
Basic	0.06	0.16	0.06	0.01	0.30
Proficient	0.00	0.06	0.10	0.05	0.21
Advanced	0.00	0.00	0.04	0.16	0.20
Total	0.30	0.27	0.21	0.23	

Contingency Table with All Cut Scores (PT2)

	Below Basic	Basic	Proficient	Advanced	Total
Below Basic	0.24	0.04	0.00	0.00	0.28
Basic	0.06	0.17	0.07	0.01	0.31
Proficient	0.00	0.06	0.10	0.05	0.21
Advanced	0.00	0.00	0.04	0.16	0.20
Total	0.30	0.27	0.21	0.22	

Indices for Classification Consistency and Accuracy (PT1)

Cut Score Level	Classification Consistency	Classification Accuracy
Cut 1	0.90	0.92
Cut 2	0.87	0.90
Cut 3	0.90	0.93
Agreement on All Categories	0.67	0.75

Note. Cut 1=Below Basic vs. Basic, Proficient, or Advanced; Cut 2=Below Basic or Basic vs. Proficient or Advanced; Cut 3=Below Basic, Basic, or Proficient vs. Advanced.

Indices for Classification Consistency and Accuracy (PT2)

Cut Score Level	Classification Consistency	Classification Accuracy
Cut 1	0.90	0.92
Cut 2	0.86	0.90
Cut 3	0.90	0.93
Agreement on All Categories	0.67	0.75

Note. Cut 1=Below Basic vs. Basic, Proficient, or Advanced; Cut 2=Below Basic or Basic vs. Proficient or Advanced; Cut 3=Below Basic, Basic, or Proficient vs. Advanced.

Table 7.28: Classification Consistency and Accuracy for Mathematics Grade 7

Contingency Table with All Cut Scores (PT1)

	Below Basic	Basic	Proficient	Advanced	Total
Below Basic	0.23	0.04	0.00	0.00	0.27
Basic	0.06	0.15	0.06	0.01	0.29
Proficient	0.00	0.06	0.12	0.05	0.23
Advanced	0.00	0.00	0.04	0.17	0.21
Total	0.29	0.25	0.23	0.24	

Contingency Table with All Cut Scores (PT2)

	Below Basic	Basic	Proficient	Advanced	Total
Below Basic	0.23	0.05	0.00	0.00	0.28
Basic	0.06	0.16	0.07	0.01	0.29
Proficient	0.00	0.06	0.12	0.05	0.23
Advanced	0.00	0.00	0.04	0.15	0.20
Total	0.29	0.26	0.23	0.22	

Indices for Classification Consistency and Accuracy (PT1)

Cut Score Level	Classification Consistency	Classification Accuracy
Cut 1	0.90	0.93
Cut 2	0.87	0.90
Cut 3	0.89	0.92
Agreement on All Categories	0.67	0.75

Note. Cut 1=Below Basic vs. Basic, Proficient, or Advanced; Cut 2=Below Basic or Basic vs. Proficient or Advanced; Cut 3=Below Basic, Basic, or Proficient vs. Advanced.

Indices for Classification Consistency and Accuracy (PT2)

Cut Score Level	Classification Consistency	Classification Accuracy
Cut 1	0.90	0.92
Cut 2	0.86	0.90
Cut 3	0.89	0.92
Agreement on All Categories	0.66	0.75

Note. Cut 1=Below Basic vs. Basic, Proficient, or Advanced; Cut 2=Below Basic or Basic vs. Proficient or Advanced; Cut 3=Below Basic, Basic, or Proficient vs. Advanced.

Table 7.29: Classification Consistency and Accuracy for Mathematics Grade 8

Contingency Table with All Cut Scores (PT1)

	Below Basic	Basic	Proficient	Advanced	Total
Below Basic	0.24	0.05	0.00	0.00	0.29
Basic	0.07	0.15	0.07	0.02	0.31
Proficient	0.00	0.06	0.10	0.05	0.20
Advanced	0.00	0.00	0.04	0.15	0.19
Total	0.31	0.27	0.21	0.21	

Contingency Table with All Cut Scores (PT2)

	Below Basic	Basic	Proficient	Advanced	Total
Below Basic	0.23	0.05	0.00	0.00	0.28
Basic	0.07	0.15	0.07	0.02	0.31
Proficient	0.00	0.06	0.10	0.05	0.22
Advanced	0.00	0.00	0.04	0.15	0.19
Total	0.30	0.26	0.21	0.22	

Indices for Classification Consistency and Accuracy (PT1)

Cut Score Level	Classification Consistency	Classification Accuracy
Cut 1	0.87	0.91
Cut 2	0.85	0.89
Cut 3	0.90	0.93
Agreement on All Categories	0.64	0.73

Note. Cut 1=Below Basic vs. Basic, Proficient, or Advanced; Cut 2=Below Basic or Basic vs. Proficient or Advanced; Cut 3=Below Basic, Basic, or Proficient vs. Advanced.

Indices for Classification Consistency and Accuracy (PT2)

Cut Score Level	Classification Consistency	Classification Accuracy
Cut 1	0.88	0.91
Cut 2	0.85	0.88
Cut 3	0.89	0.92
Agreement on All Categories	0.64	0.72

Note. Cut 1=Below Basic vs. Basic, Proficient, or Advanced; Cut 2=Below Basic or Basic vs. Proficient or Advanced; Cut 3=Below Basic, Basic, or Proficient vs. Advanced.

Figure 7.1: Conditional Standard Error of Measurement Curves of Scale Score, ELA

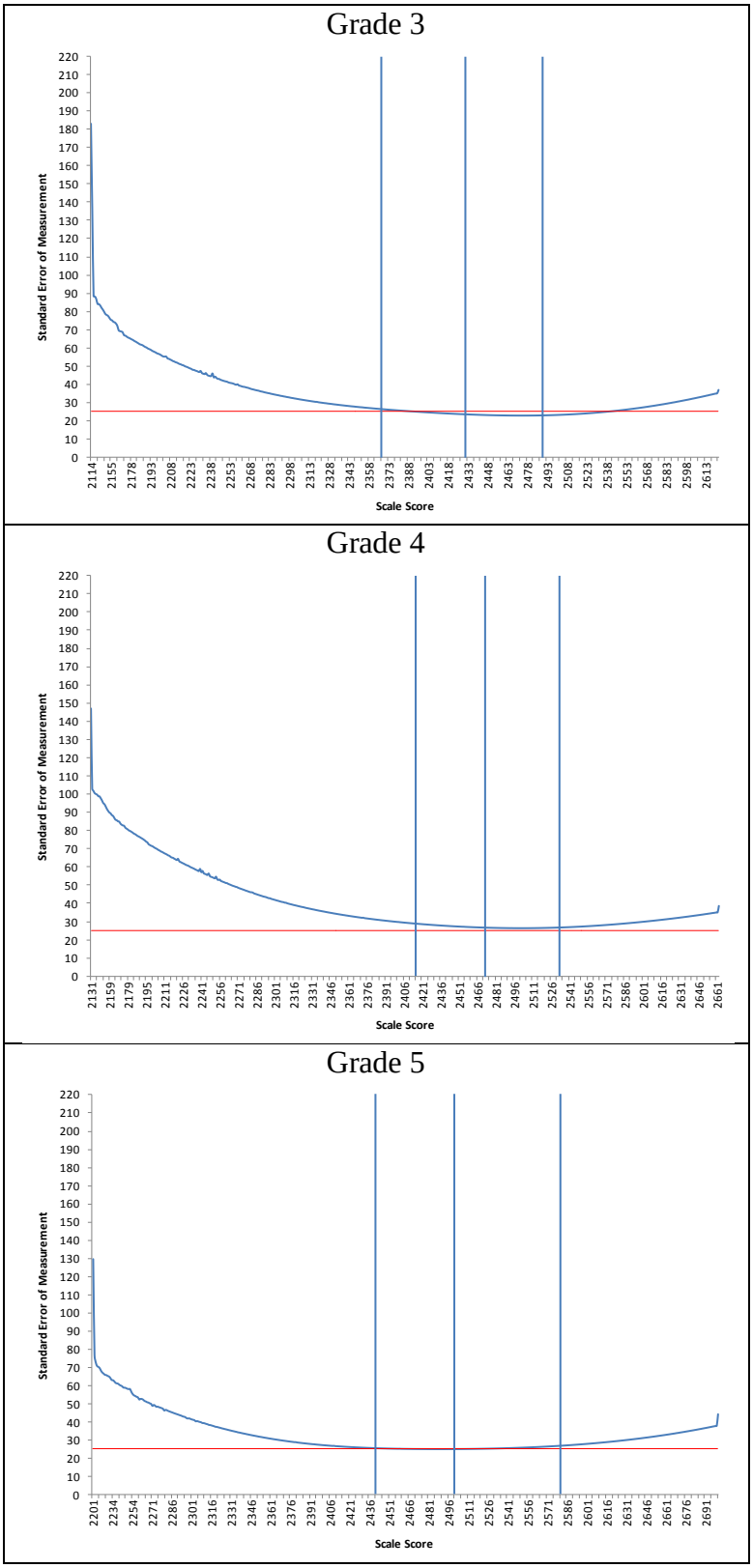


Figure 7.1: Conditional Standard Error of Measurement Curves of Scale Score, ELA (Cont'd)

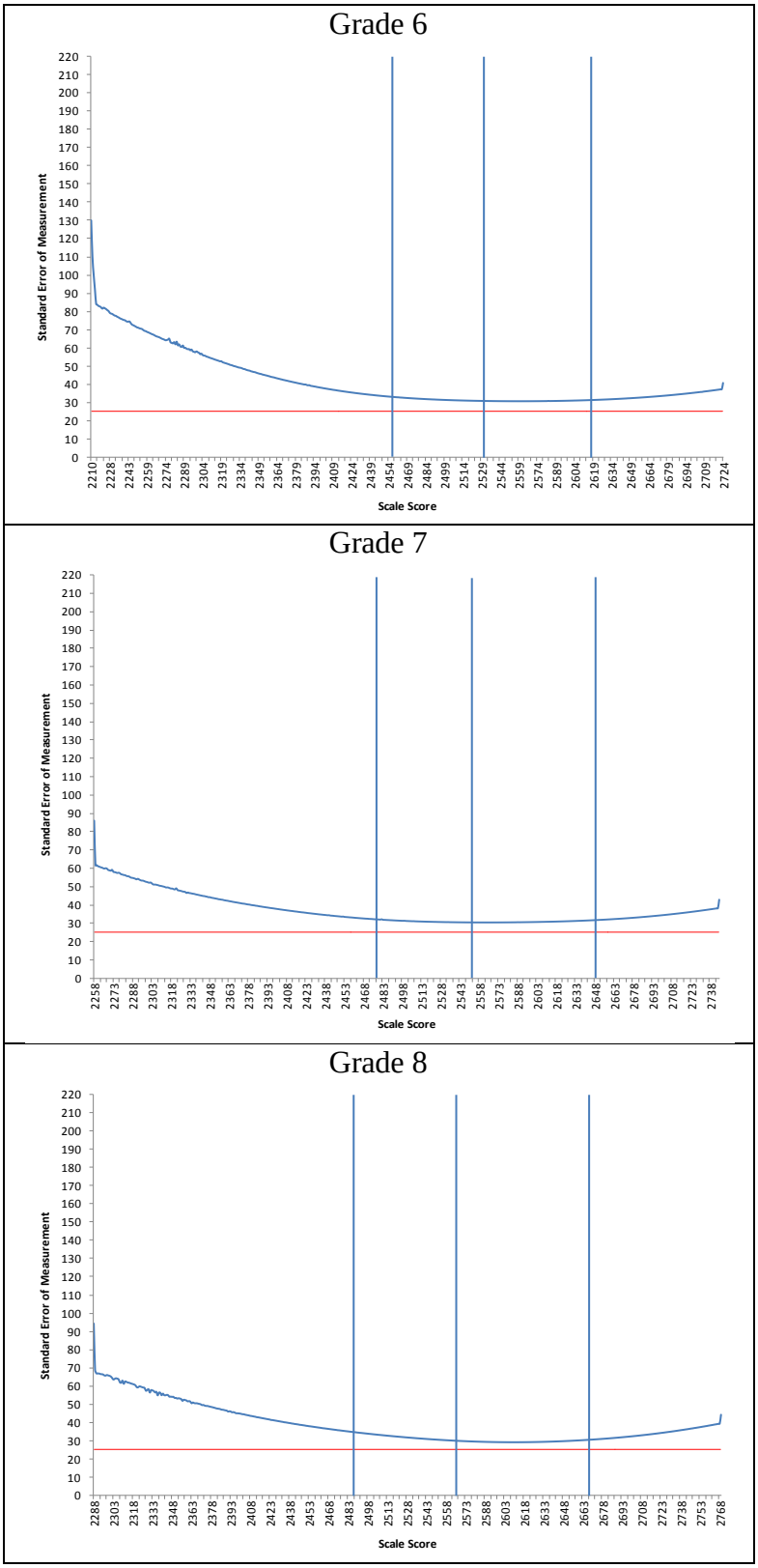


Figure 7.2: Conditional Standard Error of Measurement Curves of Scale Score, Mathematics

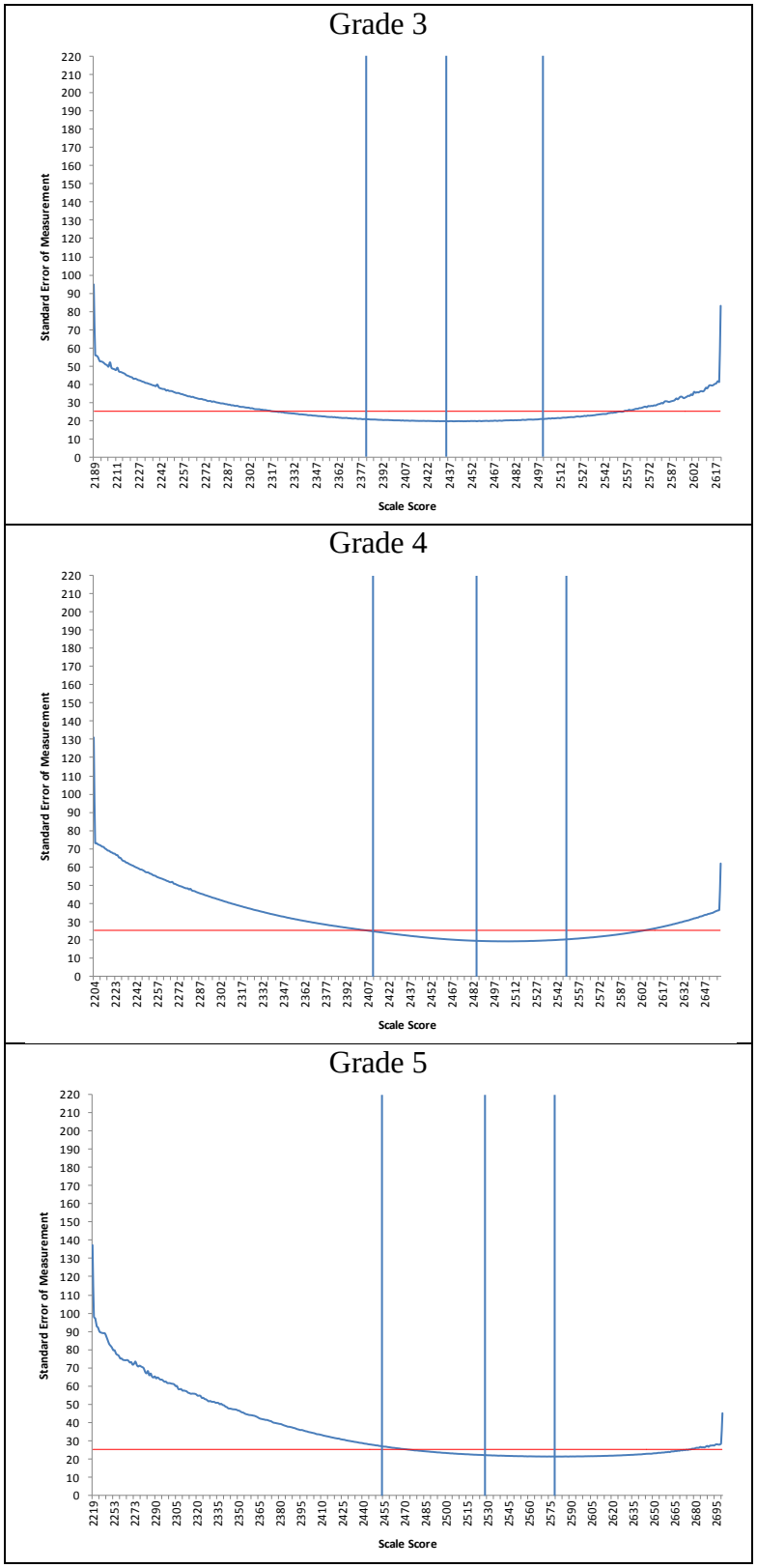
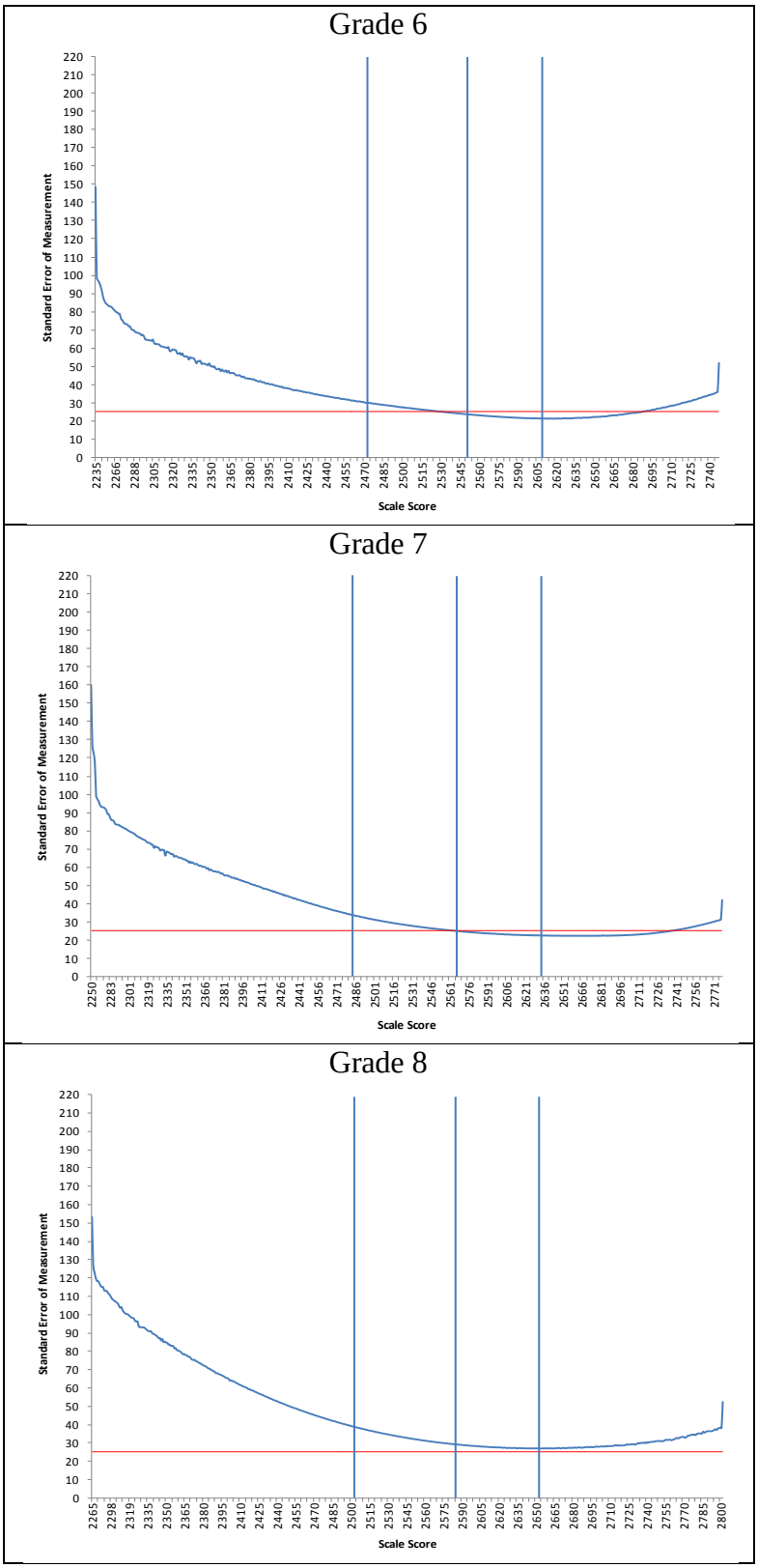


Figure 7.2: Conditional Standard Error of Measurement Curves of Scale Score, Mathematics (Cont'd)



CHAPTER 8: VALIDITY

8.1: Purpose of the Badger Exam

8.2: Evidence Based on Test Content

8.3: Evidence Based on Relations to Other Variables

8.4: Evidence Based on Response Processes

8.5: Evidence Based on Internal Structure

8.6: Evidence Based on Consequences of Testing

|

> Validity refers to 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. The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations (AERA, APA, & NCME, 2014, p.11).

The purpose of test validation is not to validate the test itself but to validate interpretations of the test scores for particular purposes or uses. As an ongoing process, test validation should involve multiple investigations and various kinds of evidence (ETS, 2002; Kane, 2006). Every aspect of an assessment provides evidence in support of its validity, including content specifications, item development, and psychometric quality for scoring and reporting.

In this Chapter, the evidence gathered is presented to support the intended uses and interpretations of scores for the Badger Exam. The description is organized in the manner prescribed by The *Standards* (AERA, APA, & NCME, 2014). These standards require a clear definition of the purpose of the test, which includes a description of the qualities, called constructs, that are to be assessed by a test, the population to be assessed, as well as how the scores are to be interpreted and used.

In addition, the *Standards* identify five kinds of evidence that can provide support for score interpretations and uses:

- evidence based on test content
- evidence based on relations to other variables
- evidence based on response processes
- evidence based on internal structure
- evidence based on the consequences of testing

The next Section defines the purpose of the Badger Exam, followed by a description and discussion of the kinds of validity evidence that have been gathered. For general test validity evidence collected by the Smarter Balanced Assessment Consortium, please refer to Chapter 2 of the Smarter Balanced field test technical report, which is available at http://www.smarterbalanced.org/wordpress/wp-content/uploads/2011/12/Chapter-02-Validity-FINAL-20150504-mm.pdf.

8.1: Purpose of the Badger Exam

The purposes of the Badger Exam are multifold, as outlined in Chapters 1 and 2. The Badger Exam is intended to comply with the federal mandate, No Child Left Behind Act of 2001, which required all states to test all test takers in ELA and Mathematics in grades 3 through 8. The results of the Badger Exam are used by the DPI as an accountability measure for school improvement to:

- meet its statutory requirement of identifying low-performing schools as stipulated by Wisconsin Statutes 115.38(4)

- meet the statutory requirement to assess test takers in writing at grades 4 and 8 as stipulated by Wisconsin Statutes 118.30
- meet the federal Title 1, No Child Left Behind requirement to determine how well children are learning
- determine the extent to which schools and districts across the state are meeting the Wisconsin proficiency standards

Also, since September 2002, scores from state summative tests have been used as one of several criteria for advancing test takers from fourth to fifth grade and from eighth to ninth grade.

The Constructs to Be Measured. The Badger Exam is designed to show how well test takers perform relative to the Smarter Balanced Assessment Consortium content standards, which are aligned to the Common Core State Standards. These standards describe what test takers should know and be able to do at each grade level. Test blueprints define the procedures used to measure the claims and standards. These documents also provide an operational definition of the construct to which each set of standards refers. That is, they define, for each content area, the subject to be assessed, the tasks to be presented, the administration instructions to be given, and the rules used to score test taker responses. They control as many aspects of the measurement procedure as possible so that the testing conditions will remain the same over test administrations (Cronbach, 1971) in order to minimize construct irrelevant score variance (Messick, 1989). The test blueprints for the Badger Exam can be found in Chapter 2 of the current technical manual and also on the DPI webpage at http://dpi.wi.gov/assessment/smarter/resources. The Smarter Balanced Assessment Consortium content specification and summative test blueprints, which are used to create the Badger Exam test blueprints, can be found on the Smarter Balanced webpage at http://www.smarterbalanced.org/smarter-balanced-assessments.

The Interpretations and Uses of the Scores Generated. Total scores expressed as scale scores, performance levels, and strength and weakness levels for each claim are generated for both ELA and Mathematics tests. An inference is drawn about how much knowledge and skill in the subject area the test taker has, based on a test taker's total score. The total score is also used to classify test takers in terms of their level of knowledge and skill in the subject area. These levels are called performance levels and are labeled Below Basic, Basic, Proficient, and Advanced.

The strength and weakness levels are used to draw inferences about a test taker's achievement in each of several specific knowledge or skill areas (i.e., claims) covered by each test. A detailed description of the uses and applications of the Badger Exam test scores is presented in Chapters 1 and 6. The description of individual test taker reports is provided on the DPI webpage at http://dpi.wi.gov/sites/default/files/imce/oea/pdf/Results%20info%20handout.pdf

The Badger Exam provides results or score summaries that are used for different purposes. The four major purposes are:

- communicating with parents and guardians
- informing decisions needed to support test taker achievement
- evaluating school programs
- providing data for state and federal accountability programs for schools

These are the only uses and interpretations of scores for which validity evidence has been gathered. If the user wishes to interpret or use the scores in other ways, the user is cautioned that

the validity of doing so has not been established. The user is advised to gather evidence to support these additional interpretations or uses (AERA, APA, & NCME, 2014, *Standard* 1.4).

Intended Test Population. Wisconsin public school test takers are the intended test population for the Badger Exam. Students in grades 3 through 8 are tested in ELA and Mathematics. In addition, private schools' test takers have the opportunity to take the Badger Exam directly from ETS. Further details regarding test taker participation and accommodations can be found in Chapter 3 and Section 5.4 of the current technical manual.

8.2: Evidence Based on Test Content

According to the *Standards* (AERA, APA, & NCME, 2014), analyses that demonstrate a strong relationship between a test's content and the construct that the test was designed to measure can provide important evidence of validity. In current K–12 testing, state content standards specify what constructs are to be tested. Test blueprints specify what kinds of items will be used to test those constructs. Therefore, validation requires evidence that the specifications in the test blueprints match the content standards and evidence that the items meet the specifications in the test blueprints.

Evidence based on test content may involve logical analyses of test content in which experts judge the adequacy with which the test content conforms to the test specifications and represents the intended domain of content. Such reviews can also be used to determine whether the test contains material that is not relevant to the construct of interest. Analyses of test content may also involve empirical evidence of item quality.

The procedures for test administration and test scoring are also to be considered in evaluating test content. As Kane (2006, p. 29) has noted, although evidence that appropriate administration and scoring procedures have been used that does not necessarily support a particular score interpretation or use, such evidence may prove useful in refuting rival explanations of test results. Evidence based on content includes the following:

Description of the State Standards. As was noted in Chapter 1, Wisconsin adopted the Common Core State Standards in 2010. The Common Core State Standards are K–12 academic standards that are aligned with college and work expectations. Wisconsin adopted these standards to guide instruction and learning for all test takers in the state and to bring Wisconsin test takers to world–class levels of achievement. The detailed description of Wisconsin state content standards can be found on the DPI webpage at http://dpi.wi.gov/sites/default/files/imce/common-core/pdf/ela-stds-app-a-revision.pdf for ELA and http://dpi.wi.gov/sites/default/files/imce/common-core/pdf/common-core-math-standards.pdf for Mathematics.

Item Specifications and Test Blueprints. Because Wisconsin elected to join the Smarter Balanced Assessment Consortium in 2010, Smarter Balanced maintains item development specifications for each ELA and Mathematics assessment of the Badger Exam. The item specifications describe the characteristics of the items that should be written to measure each content standard. A thorough description of the specifications for the Badger Exam can be found in Chapter 2, and the Smarter Balanced content and item specifications are available at http://www.smarterbalanced.org/smarter-balanced-assessments/. Once the item pool was developed by Smarter Balanced, ETS selected ELA and Mathematics items to conform to the Wisconsin content standards and test blueprints for the Badger Exam. Test blueprints for the

components of the Badger Exam were proposed by ETS and reviewed and approved by the DPI. The test blueprints for the Badger Exam also can be found in Chapter 2 and on the DPI webpage at http://dpi.wi.gov/assessment/smarter/resources.

Item Development and Review Processes. A detailed description of the content and psychometric criteria applicable to the construction of the Smarter Balanced item pool is included in the Smarter Balanced field test technical report, available at the following web addresses:

- for overall content validity, http://www.smarterbalanced.org/wordpress/wp-content/uploads/2011/12/Chapter-04-Test-Design-FINAL-20150504-mm.pdf
- for item development, http://www.smarterbalanced.org/wordpress/wp-content/uploads/2015/05/Chapter-03-Item_Development.pdf
- for item writing and review, http://www.smarterbalanced.org/smarter-balanced-assessments/item-writing-and-review/

Alignment Study. Strong alignment between standards and assessments is fundamental to meaningful measurement of test taker achievement and instructional effectiveness. Alignment results should demonstrate that the assessments represent the full range of the content standards and that these assessments measure test taker knowledge in the same manner and at the same level of complexity as expected in the content standards. The alignment study for the Smarter Balanced assessments was completed in 2014. For details, please refer to the Smarter Balanced Assessment Consortium alignment study report, which is available at http://www.smarterapp.org/documents/AlignmentStudyReport.pdf.

8.3: Evidence Based on Relations to Other Variables

Empirical results concerning the relationships between scores on a test and measures of other variables external to the test can also provide evidence of validity when these relationships are found to be consistent with the definition of the construct that the test is intended to measure. Evidence based on relations to other variables includes the traditional forms of criterion-related validity evidence: concurrent and predictive validity. The variables investigated can include other tests that measure the same construct and different constructs, criterion measures that scores on the test are expected to predict, and demographic characteristics of test takers that are expected to be related and unrelated to test performance.

An important area of research is the relationship of a Smarter Balanced assessment with other important national and international large-scale assessment programs, such as the National Assessment of Educational Progress (NAEP) and the Program for International Student Assessment (PISA). During the 2014 Smarter Balanced field test administration, a subset of test takers who took NAEP and PISA items also took Smarter Balanced items and performance tasks. Inferences concerning relative performance on these items relied on assumptions concerning factors like test delivery mode effects, item context effects, the impact of different testing windows and years, and the impact of different test purposes. The NAEP Mathematics, reading or writing, and PISA literacy content and skills frameworks are also quite different from a Smarter Balanced assessment. A brief summary of the resulting item performance for NAEP, PISA, and all Smarter Balanced field test items was described in Chapters 7 and 8 of the Smarter

Balanced field test technical report, which is available at http://www.smarterbalanced.org/smarter-balanced-assessments/.

In addition, several schools/districts in each state also administered the NAEP tests in the spring of 2015. The NAEP average performance by various groups (including states and demographic groups) and achievement level distributions can be compared to those of Smarter Balanced assessments. This is the approach essentially taken by Phillips (2012) in comparing NAEP results to state achievement test results.

8.4: Evidence Based on Response Processes

As noted in the *Standards* (AERA, APA, & NCME, 2014), additional support for a particular score interpretation or use can be provided by theoretical and empirical evidence indicating that test takers are using the intended response processes when responding to the items in a test. Finally, evidence may also be derived from feedback provided by raters (or scorers) involved in the scoring of test taker responses.

Evidence of Interrater Agreement. Rater consistency for the human-scored items is critical for the Badger Exam test scores and their interpretations. These findings provide evidence of the degree to which raters agree in their observations about the qualities evident in test takers' responses. In order to evaluate the reliability of the test taker scores on the human-scored items, two raters scored approximately 10% of the test taker responses. The data collected were used to evaluate the interrater agreement.

The findings from weighted kappa statistics provide evidence of the degree to which test takers' scores may vary from rater to rater. Without explicit criteria to guide the rating process, two independent raters might not assign the same score to a given response. The results showed moderate levels of agreement between raters' scores on test takers' responses to the human-scored items administered in ELA and Mathematics tests. The Chapter 7 (Reliability) of the current technical manual contains the intraclass correlation and weighted kappas as well as exact and adjacent agreement rates.

Data Forensic Analysis Using Item and Test Response Time Data. After the 2015 Badger Exam administration, ETS conducted data forensic analyses to discover the trend of test performance by auxiliary information such as total testing time, the total number of times test takers visited each item, and so on. The relationship between the total testing time response and total test score from each test taker was further investigated to identify any highly implausible test performances. There were a number of test takers (approximately 100) who got high test scores with very short testing times in both subject areas, but there was no suspicious group-level behavior among schools and districts by testing time.

In addition, item-level data forensic analyses were conducted for the items in Section 2 (i.e., performance task) to detect any evidence of potential item breach. Two item-level statistics (*p*-value and item-total correlation) for each performance task item were checked with two auxiliary variables (item response time and the number of times test takers visited each item) to determine if there were any irregular patterns between the item statistics and the auxiliary variables (e.g., very high *p*-value in short item response time). There was no suspicious item-level breach for any of the 59 performance task items.

8.5: Evidence Based on Internal Structure

As suggested by the *Standards* (AERA, APA, & NCME, 2014), evidence of validity can also be obtained from studies of the properties of the reported test scores and the relationship between these scores and scores on components of the test. To the extent that the score properties and relationships found are consistent with the definition of the construct measured by a test, support is gained for interpreting these scores as measures of the construct.

For the Badger Exam, it is assumed that a single construct underlies the total scores obtained on each test. Evidence to support this assumption can be gathered from the results of item analyses, evaluations of internal consistency, and studies of model-data fit, dimensionality, and reliability. With respect to the strength and weakness levels (derived from claim scores) that are reported, claim scores are intended to reflect test takers' knowledge and/or skill in an area that is part of the construct underlying the total test.

Classical Item Statistics. Point-biserial correlations calculated for the items in a test show the degree to which the items discriminate between test takers with low and high scores on a test. High correlations provide evidence that the items assess the same construct. Showing that levels of item difficulties span the range of test taker ability provides evidence that test takers at all levels of ability are adequately measured by the items. The point-biserial item-total correlations and *p*-values (or item difficulties) for the items in the Badger Exam are presented in Chapter 6 of the current technical manual.

Reliability. Reliability is a prerequisite for validity. The finding of reliability in test taker scores supports the validity of the inference that the scores reflect a consistent construct. Results of the reliability analyses (including internal consistency, classification consistency, and accuracy) for each grade level and subject area of the Badger Exam are presented in Chapter 7 of the current technical manual. The results for internal consistency indicate that the reliabilities for all Badger Exams were moderate to high, ranging from 0.87 to 0.91. The findings from the reliability of classification decisions are also described in Chapter 7 of the current technical manual. When the decisions are collapsed to below proficient (i.e., "Below Basic" and "Basic") versus at or above proficient (i.e., "Proficient" and "Advanced"), which are the critical categories for school accountability analyses, the proportion of test takers that were estimated to be classified accurately ranged from 0.88 to 0.90 across all grade levels and subject areas of the Badger Exam. Similarly, the proportion of test takers that were estimated to be classified consistently ranged from 0.84 to 0.88 for test takers classified into below proficient versus at or above proficient.

Correlations Within and Between Content Areas. The distinctiveness and reliability of the claim scores in each subject area are important because the Badger Exam reports strength and weakness levels based on claim scores. The interrelationships of claim scores should be shown to be consistent with the construct being assessed. Tables 8.1 through 8.6 provide the intercorrelations between claim scores within two subject areas (i.e., ELA and Mathematics). The summary of results are below:

- The intercorrelations among ELA claim scores ranged from 0.42 to 0.71.
- The intercorrelations among Mathematics claim scores ranged from 0.47 to 0.70.
- The correlations between ELA and Mathematics claim scores ranged from 0.32 to 0.68.

There was an interesting finding that some results between ELA and Mathematics claim scores showed higher correlation than the results within each subject area. For example, in Table 8.1,

the correlation between ELA Claim 2 and Mathematics Claim 1 (i.e., 0.63) was higher than the correlation between ELA Claims 2 and 3 (i.e., 0.55) or the correlation between ELA Claims 2 and 4 (i.e., 0.58). The number of items in Claims 2, 3, and 4 (6 to 10 items in each of ELA and Mathematics claims) was smaller than Claim 1 (18 to 21 items from both subject areas). The smaller number of items results in lower reliability and that can affect the correlations among Claims 2, 3, and 4. As would be expected because the Mathematics Claim 1 score had more items than other claim scores, the Mathematics Claim 1 score was more highly correlated with the other claim scores, even with different subject area claim scores (e.g., ELA Claim 2 score).

Item-Level Confirmatory Factor Analysis. Confirmatory factor analysis is a useful statistical methodology as it can evaluate whether performance on the reporting categories of each of the Badger Exams reflects a single underlying construct. Evaluating dimensionality is very important in order to ensure a test is measuring a single construct as purported. Findings from this type of analysis are also helpful as they can provide data supporting the assumption of unidimensionality for IRT models used to calibrate the Badger Exam items.

It is assumed that each Badger Exam is unidimensional, measuring a specific subject area (e.g., ELA or Mathematics). Each Badger Exam is also designed to measure three or four claims within a specific subject area. These claims represent different knowledge and skills but are correlated to some degree. The strength and weakness levels derived from claim scores are reported to provide teachers, parents, and test takers more detailed information about test takers' learning and performance on the test. Two hypothetical constructs are statistically tested and compared to examine the structure of the Badger Exams:

- Each test is strictly unidimensional where all items in a test measure a single knowledge and skill. This is a single-factor structural model in which all items load on a general factor. This model presumes all modeled items contribute to the estimation of a general ability factor.
- The second structural model supports claim score reporting and hypothesizes that each test measures several distinct but correlated areas of knowledge and skills. This is a multifactor model in which an item loads on the claim to which it corresponds.

MPLUS 7.11 (Muthén & Muthén, 2012) was used to fit specified factor models to the data. For each subject area, two models were fit to the data: a one-factor model and a multifactor model corresponding to the items in each area. Parameter estimation for item-level data was performed using a robust weighted least squares estimator with the diagonal weight matrix (WLSMV) method. The first indicator of each factor was set to one for scale identification.

The hypothesized models were evaluated based on goodness-of-fit indices, model parsimony, and reasonableness of individual parameter estimates (i.e., statistical significance, residuals, and modification indices). In addition, the estimated correlations among the latent factors were considered (i.e. correlations higher than 0.90 were flagged extreme). The following goodness-of-fit indices were used:

- Root mean squared error of approximation (RMSEA) and its 90% confidence interval (90% CI). RMSEA values of 0.05 or below indicate good fit (Browne & Cudeck, 1993).
- Comparative fit index (CFI) and Tucker-Lewis index (TLI). A cut-off value close to 0.95 is recommended for both indices (Hu & Bentler, 1999).

- Satorra-Bentler scaled chi-square statistic (SBχ^2; Satorra & Bentler, 1994). However, model chi-square statistics are sensitive to sample size, and trivial departures from fit will yield a significant chi-square (Fabriger, Wegener, MacCallum, & Strahan, 1999). In this analysis, the sample size was large and, as a result, this statistic should be interpreted with caution.

Due to the large number of cases analyzed, the chi-square statistics were all artificially inflated (see Tables 8.7 and 8.8). However, the other fit statistics (CFI, TLI, and RMSEA) are within acceptable ranges for good model fit. In general, the improvement in the fit indices was small when multifactor models (i.e., models that conform to the strength and weakness level reporting) were fit to the data. Since the single-factor model fits reasonably well to the data for all of the grade level tests of each subject area, the unidimensionality assumption and the IRT-based ability estimation are both supported. In addition, the Smarter Balanced Assessment Consortium conducted dimensionality studies based on IRT procedures during pilot test administration. Results are available at http://www.smarterbalanced.org/wordpress/wp-content/uploads/2011/12/Chapter-6.pdf.

8.6: Evidence Based on Consequences of Testing

As observed in the *Standards* (AERA, APA, & NCME, 2014), tests are usually administered "with the expectation that some benefit will be realized from the intended use of the scores" (p. 19). When this is the case, expected-benefits evidence will provide support for intended use of the scores. DPI and ETS will consider what kinds of information can be gathered to assess the consequences of administration of the Badger Exam.

Table 8.1: Claim Score Intercorrelations between ELA and Mathematics, Grade 3

Test	Strength &Weakness	ELA				Mathematics		
		Claim 1	Claim 2	Claim 3	Claim 4	Claim 1	Claim 2&4	Claim 3
ELA	Claim 1	--						
	Claim 2	0.70	--					
	Claim 3	0.63	0.55	--				
	Claim 4	0.66	0.58	0.56	--			
Mathematics	Claim 1	0.62	0.63	0.51	0.53	--		
	Claim 2&4	0.59	0.58	0.50	0.51	0.70	--	
	Claim 3	0.61	0.58	0.51	0.52	0.67	0.62	--

Table 8.2: Claim Score Intercorrelations between ELA and Mathematics, Grade 4

Test	Strength &Weakness	ELA				Mathematics		
		Claim 1	Claim 2	Claim 3	Claim 4	Claim 1	Claim 2&4	Claim 3
ELA	Claim 1	--						
	Claim 2	0.63	--					
	Claim 3	0.65	0.55	--				
	Claim 4	0.55	0.47	0.53	--			
Mathematics	Claim 1	0.62	0.56	0.54	0.46	--		
	Claim 2&4	0.57	0.51	0.50	0.42	0.68	--	
	Claim 3	0.58	0.51	0.52	0.44	0.67	0.62	--

Table 8.3: Claim Score Intercorrelations between ELA and Mathematics, Grade 5

Test	Strength &Weakness	ELA				Mathematics		
		Claim 1	Claim 2	Claim 3	Claim 4	Claim 1	Claim 2&4	Claim 3
ELA	Claim 1	--						
	Claim 2	0.71	--					
	Claim 3	0.67	0.59	--				
	Claim 4	0.69	0.62	0.60	--			
Mathematics	Claim 1	0.64	0.59	0.56	0.58	--		
	Claim 2&4	0.64	0.58	0.56	0.57	0.70	--	
	Claim 3	0.63	0.58	0.56	0.57	0.69	0.66	--

Table 8.4: Claim Score Intercorrelations between ELA and Mathematics, Grade 6

Test	Strength &Weakness	ELA				Mathematics		
		Claim 1	Claim 2	Claim 3	Claim 4	Claim 1	Claim 2&4	Claim 3
ELA	Claim 1	--						
	Claim 2	0.59	--					
	Claim 3	0.62	0.47	--				
	Claim 4	0.65	0.49	0.51	--			
Mathematics	Claim 1	0.68	0.55	0.57	0.56	--		
	Claim 2&4	0.58	0.47	0.50	0.47	0.70	--	
	Claim 3	0.56	0.45	0.47	0.46	0.64	0.58	--

Table 8.5: Claim Score Intercorrelations between ELA and Mathematics, Grade 7

Test	Strength &Weakness	ELA				Mathematics		
		Claim 1	Claim 2	Claim 3	Claim 4	Claim 1	Claim 2&4	Claim 3
ELA	Claim 1	--						
	Claim 2	0.67	--					
	Claim 3	0.64	0.57	--				
	Claim 4	0.62	0.55	0.55	--			
Mathematics	Claim 1	0.65	0.62	0.57	0.57	--		
	Claim 2&4	0.53	0.49	0.47	0.46	0.65	--	
	Claim 3	0.57	0.53	0.50	0.50	0.68	0.54	--

Table 8.6: Claim Score Intercorrelations between ELA and Mathematics, Grade 8

Test	Strength &Weakness	ELA				Mathematics		
		Claim 1	Claim 2	Claim 3	Claim 4	Claim 1	Claim 2&4	Claim 3
ELA	Claim 1	--						
	Claim 2	0.66	--					
	Claim 3	0.53	0.42	--				
	Claim 4	0.64	0.53	0.42	--			
Mathematics	Claim 1	0.62	0.56	0.48	0.49	--		
	Claim 2&4	0.59	0.53	0.44	0.47	0.69	--	
	Claim 3	0.45	0.40	0.32	0.36	0.52	0.47	--

Table 8.7: Confirmatory Factor Analysis for ELA

Grade	Model	Chi-square			*RMSEA*	*CFI*	*TLI*	Range of correlations among factors
		df	SBχ^2	*p*				
3	Single-Factor	945	25,244	.000	.020	.980	.979	
	Four-Factor	939	22,262	.000	.019	.982	.981	.88 – .96
4	Single-Factor	1,034	16,422	.000	.016	.984	.984	
	Four-Factor	1,028	12,347	.000	.013	.989	.988	.85 – .95
5	Single-Factor	945	19,674	.000	.018	.986	.985	
	Four-Factor	939	17,438	.000	.017	.988	.987	.90 – .96
6	Single-Factor	1,080	13,939	.000	.014	.986	.985	
	Four-Factor	1,074	12,173	.000	.013	.988	.987	.87 – .96
7	Single-Factor	1,034	7,292	.000	.014	.986	.985	
	Four-Factor	1,028	6,187	.000	.013	.988	.988	.87 – .96
8	Single-Factor	1,034	19,807	.000	.017	.978	.977	
	Four-Factor	1,028	15,551	.000	.015	.983	.982	.74 – .97

Table 8.8: Confirmatory Factor Analysis for Mathematics

Grade	PT set	Model	Chi-square			*RMSEA*	*CFI*	*TLI*	Range of correlations among factors
			df	SBχ^2	*p*				
3	PT1	Single-Factor	594	33,969	.000	.043	.945	.942	
		Three-Factor	591	30,454	.000	.041	.951	.948	.90 – .95
	PT2	Single-Factor	629	30,978	.000	.039	.956	.954	
		Three-Factor	626	28,610	.000	.038	.960	.957	.91 – .98
4	PT1	Single-Factor	594	36,670	.000	.045	.952	.949	
		Three-Factor	591	32,528	.000	.042	.958	.955	.91 – 1.00*
	PT2	Single-Factor	629	43,435	.000	.047	.945	.941	
		Three-Factor	626	37,897	.000	.044	.952	.949	.90 – 1.00*
5	PT1	Single-Factor	629	29,748	.000	.038	.960	.958	
		Three-Factor	626	27,633	.000	.037	.963	.961	.94 – 1.00*
	PT2	Single-Factor	629	37,982	.000	.044	.951	.948	
		Three-Factor	626	34,048	.000	.042	.956	.953	.92 – .1.00*
6	PT1	Single-Factor	560	16,007	.000	.030	.977	.975	
		Three-Factor	557	15,603	.000	.030	.977	.976	.96 – 1.00*
	PT2	Single-Factor	594	17,017	.000	.030	.976	.975	
		Three-Factor	591	16,404	.000	.030	.977	.975	.96 – .99
7	PT1	Single-Factor	594	23,701	.000	.036	.977	.976	
		Three-Factor	591	22,280	.000	.035	.978	.977	.94 – 1.00*
	PT2	Single-Factor	560	23,463	.000	.037	.977	.976	
		Three-Factor	557	22,094	.000	.036	.979	.977	.94 – .99
8	PT1	Single-Factor	560	10,679	.000	.025	.977	.976	
		Three-Factor	557	10,484	.000	.025	.977	.976	.96 – .99
	PT2	Single-Factor	594	13,429	.000	.026	.974	.972	
		Three-Factor	591	13,321	.000	.026	.974	.972	.97 – .98

*The latent variable covariance matrix (psi) was not positive definite which was caused by a correlation equal to one between two latent variables.

CHAPTER 9: QUALITY CONTROL PROCEDURES

9.1: Overview of Quality Control Process

9.2: Quality Control of Test Materials

9.3: Quality Control of System Functionality

9.4: Quality Control of Psychometric Analyses

9.5: Quality Control of Scoring and Reporting

9.1: Overview of Quality Control Process

DPI, ETS, and CA&L implemented rigorous quality control procedures throughout the test development, administration, scoring, and analyses processes. As part of this effort, ETS staff consulted with the Office of Professional Standards. The Office of Professional Standards publishes and maintains the *ETS Standards for Quality and Fairness* (ETS, 2002), with the purposes of a) helping design, develop, and deliver technically sound, fair, and useful products and services, and b) helping the public and auditors evaluate those products and services. In the following Sections, quality control procedures during the test administrations are outlined.

9.2: Quality Control of Test Materials

ETS meticulously followed a set of internal quality standards to ensure high-quality online published products for all testing-related materials. Quality control in test administration requires that the contents of all test materials (including digital information, ad hoc documents, and test administration manuals) align with one another and present accurate information because contradicting information creates frustration to the test users and may impact the validity of test use.

To help ensure consistency in test materials used for the Badger Exam, the manuals (i.e., test administration manuals, training materials, and technical manual) and the digital information (i.e., internal memo posted in TOMS) were reviewed by subject matter experts at ETS and DPI. Documents were developed through multiple iterations such as content review cycles and then underwent an editorial review by ETS internal editors. Digital information was first drafted by ETS program management and then reviewed by multiple teams whose compositions depended on the content of the information. All test material reviews were built into the planned test administration schedule and all materials were released to appropriate test users after the DPI's final approval.

9.3: Quality Control of System Functionality

For the Badger Exam, the ETS quality assurance team and CA&L conducted testing procedures on the following aspects of the end-to-end system in both the user acceptance and production environments: test delivery, item content rendering, and TOMS. These activities adhere to the software development life cycle process as follows:

- ***Software Testing.*** ETS developed user acceptance test plans and test scripts. A number of testing activities took place with these plans and scripts, including the testing of software components, security testing, integration testing, hardware and network capacity testing, data conversion testing, and load testing.
- ***Data Conversion Testing.*** ETS performed testing for data conversions in the system. These data conversions include but are not limited to: test taker data import, test taker scoring, raw score test taker assignment, and raw score to individual test taker report conversion. Quality assurance professionals compared samples of data in order to verify that the source data matches the converted data in the destination systems.

- ***Hardware and Network Capacity Testing.*** ETS and CA&L provided readiness tools to help districts and schools prepare their hardware and networks for the testing windows.
- ***System and Integration Testing.*** The ETS software quality assurance staff performed system-level testing. The staff validated the system against all requirements. This process included verifying system accessibility, links, scoring, reporting, security, and performance. During this phase, staff could detect and correct issues before the final release.
- ***Operational Trial.*** Prior to the release of each project, the ETS software quality assurance staff performed full system-level tests in an independent test environment that mirrors the production configuration. Staff members also tested the system on all supported computer platforms and browsers. These system-level tests included comprehensive assessments on functionality, usability, reliability, security, and overall performance. The staff verified that each webpage, link, item, and image displayed properly through the graphical user interface standards. During this process, the staff members also validated system content for accuracy.
- ***Load Testing.*** ETS regularly performed extensive load testing to determine system capacity and to provide quality delivery of online assessments. Load testing consisted of employing machines across the Internet to simulate the test taker testing environment. During this testing period, ETS staff obtained data that enables long-term scalability planning.
- ***Security Testing.*** In order to establish the integrity, confidentiality, and availability of DPI's data, ETS and CA&L used industry-standard tools to regularly run automated security scans against production and development networks and systems. Real-time vulnerability updates protected ETS and CA&L systems against the very latest known threats.
- ***User Acceptance Testing.*** ETS worked with DPI to develop and review the user acceptance tests to confirm the system meets the requirements of the contract. ETS staff participated in the User Acceptance Testing. All participants were highly proficient and knowledgeable on TOMS, the Badger Exam, and the Smarter Balanced Assessment Consortium requirements. The testing simulated the test experience for schools including loading test taker information into TOMS, generating test tickets, logging into the assessment, evaluating the functionality of all universal tools, designated supports, and test accommodations, and the navigation features of the assessment. In addition to the quality assurance checks on functionality, the system's consistency in capturing responses and transferring of test taker data for scoring was also evaluated.

For each of these quality checks, ETS and CA&L staff members were required to evaluate specific functioning outlined on a quality assurance checklist.

9.4: Quality Control of Psychometric Analyses

ETS and CA&L took various necessary measures to ascertain that the scoring keys were applied to the test taker responses as expected and the test taker scores were computed accurately. Scoring keys, provided in the answer key files, were produced by Smarter Balanced Assessment Consortium and were verified thoroughly by performing various quality-control checks. ETS created form planners based on the Smarter Balanced scoring key information. The form planners contain the information about an assembled test form including the test name,

administration year, claim identification, and scoring key. The form planners also have IRT statistics associated with each item.

Various checks are performed for the Smarter Balanced answer key files and the Badger Exam form planners. The answer key files and form planners were checked for accuracy against the specification and layout documents before the keys were loaded into the score key management system at ETS. ETS independently generated scores and verified CA&L's scoring of the early return data by comparing the two independent results. Any discrepancies were then resolved. The entire scoring system was tested using a test deck that includes typical and atypical response vectors. Preliminary classical item analyses were conducted on an early return sample to provide an additional check of the keys.

The psychometric analyses conducted at ETS underwent comprehensive quality checks by a team of psychometricians and data analysts. Detailed checklists were consulted by lead psychometricians to systematically review the statistical procedures performed on each Badger Exam. The results of preliminary classical item analyses that provided a check on scoring keys were reviewed by two psychometricians before pre-equated test scores were released onto TOMS. In addition to operational scoring and quality-control processes by ETS internal teams (i.e., scoring key management and data quality services), the psychometric team verified the results with technical and information processing service (TIPS) to parallel process all psychometric analyses.

9.5: Quality Control of Scoring and Reporting

ETS's scoring and reporting systems have quality control procedures integrated throughout, including both automated and manual inspections, to ensure data accuracy. ETS Assessment Development, Research and Statistical Analysis, Performance Assessment Scoring Service, and Information Technology groups all participated in certifying the scoring and reporting system to ensure operational readiness and scoring integrity. All teams followed established procedures required by the International Organization for Standardization (ISO) 9000 family of standards. The combined efforts of each of these groups provided multiple layers of quality assurance and control.

ETS built and reviewed the scoring system models based on the Smarter Balanced Assessment Consortium scoring specifications and DPI requirements. Machine-scored item responses and demographic information were sent into a master test taker file. Human-scored item responses were also sent electronically to the ETS Online Network for Evaluation (ONE) scoring centers for scoring by trained, qualified raters. Record counts were verified against the counts obtained during security check-in from the document processing staff to ensure all test takers were accounted for in the file.

Once the record counts were reviewed, the machine-scored item responses were scored against the appropriate approved answer key provided by the Smarter Balanced Assessment Consortium. In addition, the test taker's original response string was stored for data verification and auditing purposes. ETS worked with DPI to determine and refine the documentation of specifications for the scoring of answer documents well in advance of the receipt of test materials. These specifications contained detailed scoring procedures, along with the procedures for determining whether a test taker has attempted a test and whether that test taker should be included in

statistics and calculations for computing summary data. Standard quality inspections were performed on all data files, including the evaluation of each test taker data record for correctness and completeness. Student results were kept confidential and secure at all times.

Upon the completion of the thorough data verification process, quality checks were performed on the data placement and report file formatting for each data element displayed on the reports. All reporting data elements were verified by comparing back to the production data file and the reporting processing rules. Additional quality crosschecks were performed to ensure accuracy and consistency across all reporting media for the assessment. Similar quality checks were also used to validate data at the school, district, and state levels. Senior software quality assurance analysts conducted a second review of each report to ensure methodology, processes, and procedures were followed, and they verified the reports that were approved for production. An additional post-print review was conducted before any electronic reports were produced and distributed.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick (Eds.), *Statistical Theories of Mental Test Scores* (pp. 395–479). Reading: Addison-Wesley.

Browne, M.W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. Bollen & J.S. Long (Eds.), *Testing Structural Equation Models* (pp. 136–162). Newbury Park, CA: Sage Publications.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Cronbach, L.J. (1971). Test validation. In R.L. Thorndike (Ed.), *Educational Measurement* (2nd ed.). Washington, DC: American Council on Education.

Cronbach, L.J., Schönemann, P., & McKie, D. (1965). Alpha coefficients for stratified parallel tests. *Educational and Psychological Measurement, 25*, 291–312.

Darling-Hammond, L., & Pecheone, R. (2010). *Developing an internationally comparable balanced assessment system that supports high-quality learning*. Retrieved from http://www.k12center.org/publications.html.

Educational Testing Service (2002). *ETS Standards for Quality and Fairness.* Princeton, NJ: Educational Testing Service.

Fabrigar, L.R., Wegener, D.T., MacCallum, R.C., & Strahan, E.J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*, 272–299.

Haertel, E.H. (2006). Reliability. In R. L. Brennan (Ed.), Educational Measurement (4th ed., pp. 65–110). Washington, DC: American Council on Education and National Council on Measurement in Education.

Hu, L.-T., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55.

Human Resources Research Organization. (2014). *Smarter Balanced Assessment Consortium: Alignment Study Report.* Retrieved from http://www.smarterapp.org/documents/AlignmentStudyReport.pdf

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17–64). Washington, DC: American Council on Education and National Council on Measurement in Education.

Kolen, M.J. (2011). *Issues associated with vertical scales for PARCC assessments.* PARCC. Retrieved from http://parcconline.org/files/40/Technical%20Advisory%20Committee/43/Vertical-Scales-Kolen.pdf

Kolen, M.J., & Brennan, R.L. (2014). *Test Equating, Scaling, and Linking: Methods and Practices* (3rd ed.). New York, NY: Springer.

Livingston, S.A., and Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32,* 179–197.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18,* 5–11.

Mislevy, R.J., Almond, R.G., & Lukas, J.F. (2003). *A brief introduction to evidence-centered design*. Retrieved from http://www.education.umd.edu/EDMS/mislevy/papers/BriefIntroECD.pdf

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159–176.

Muthén, L., & Muthén, B. (1998-2012). *Mplus user's guide.* (7th ed.). Los Angeles, CA.

Phillips, G. (2012). The benchmark method of standard setting. In G.J. Cizek (Ed.), *Setting Performance Standards: Foundations, Methods, and Innovations.* New York: Routledge.

Satorra, A., & Bentler, P.M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C.C. Clogg (Eds.), *Latent Variables Analysis: Applications for Developmental Research* (pp. 399–419). Thousand Oaks, CA: Sage.

Stocking, M.L. (1996). An alternative method for scoring adaptive tests. *Journal of Educational and Behavioral Statistics, 21,* 365–389.

Webb, N.L. (2005). *Web Alignment Tool (WAT) training manual*, Retrieved from http://wat.wceruw.org/index.aspx

Webb, N.L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education, 20,* 7–25.

APPENDIX A: OPERATIONAL ITEM STATISTICS

Table A.1: Operational Item Statistics, ELA Grade 3

Item Number	Item Statistic Fields				Flag			
	Item Type	*p*-value	Item-Total Correlation	Omit Rate	*p*-value	Item-Total Correlation	Distractor Correlation	Omit Rate
1	HTQ	0.31	0.48	0.3				
2	HTQ	0.66	0.43	0.1				
3	MC	0.85	0.45	0.1				
4	MC	0.36	0.23	0.1				
5	MS	0.10	0.31	0.2	X			
6	MC	0.55	0.40	0.1				
7	MC	0.78	0.49	0.2				
8	EBSR	0.41	0.53	0.1				
9	MS	0.36	0.38	0.1				
10	MS	0.17	0.23	0.2	X			
11	SA	0.44	0.59	0.3				
12	MC	0.56	0.46	0.2				
13	MC	0.59	0.52	0.3				
14	HTQ	0.78	0.41	0.2				
15	MS	0.33	0.48	0.2				
16	HTQ	0.32	0.53	0.2				
17	MS	0.31	0.45	0.2				
18	MC	0.39	0.31	0.2				
19	MC	0.65	0.43	0.2				
20	MC	0.30	0.22	0.2				
21	MC	0.57	0.43	0.2				
22	MS	0.15	0.44	0.2	X			
23	EBSR	0.14	0.31	0.2	X			
24	HTQ	0.27	0.45	0.2	X			
25	MC	0.38	0.22	0.2				

Note. MC=Multiple-choice single-select; MS=Multiple-choice multi-select; HTQ=Hot text QTI version; EBSR=Evidence-based selected-response; SA=Short answer.

Table A.2: Operational Item Statistics, ELA Grade 3 (Cont'd)

Item Number	Item Statistic Fields				Flag			
	Item Type	*p*-value	Item-Total Correlation	Omit Rate	*p*-value	Item-Total Correlation	Distractor Correlation	Omit Rate
26	MC	0.64	0.55	0.2				
27	MC	0.37	0.31	0.2				
28	MC	0.51	0.40	0.3				
29	EBSR	0.23	0.50	0.2	X			
30	MC	0.77	0.48	0.3				
31	EBSR	0.46	0.64	0.3				
32	MC	0.55	0.52	0.3				
33	MC	0.72	0.50	0.4				
34	MC	0.49	0.35	0.3				
35	MS	0.23	0.42	0.3	X			
36	MC	0.61	0.51	0.5				
37	EBSR	0.32	0.56	0.3				
38	MC	0.60	0.40	0.4				
39	MS	0.19	0.39	0.5	X			
40	MC	0.27	0.27	0.3	X			
41	MI	0.44	0.31	0.3				
42	MS	0.11	0.33	0.4	X			
43	EBSR	0.32	0.54	0.2				
44	MC	0.40	0.40	0.2				
45	MC	0.46	0.47	0.3				

Note. MC=Multiple-choice single-select; MS=Multiple-choice multi-select; HTQ=Hot text QTI version; EBSR=Evidence-based selected-response; SA=Short answer.

Table A.2: Operational Item Statistics, ELA Grade 4

Item Number	Item Statistic Fields				Flag			
	Item Type	*p*-value	Item-Total Correlation	Omit Rate	*p*-value	Item-Total Correlation	Distractor Correlation	Omit Rate
1	MC	0.88	0.36	0.1				
2	MC	0.47	0.28	0.1				
3	MC	0.51	0.20	0.1		X		
4	HTQ	0.39	0.31	0.1				
5	HTQ	0.27	0.16	0.2	X	X		
6	MS	0.22	0.15	0.1	X	X		
7	HTQ	0.20	0.38	0.1	X			
8	EBSR	0.39	0.60	0.0				
9	EBSR	0.15	0.19	0.1	X	X		
10	HTQ	0.51	0.31	0.1				
11	MS	0.46	0.58	0.1				
12	MC	0.34	0.26	0.1				
13	EBSR	0.31	0.37	0.1				
14	SA	0.44	0.52	0.2				
15	MS	0.22	0.45	0.1	X			
16	MC	0.49	0.37	0.1				
17	HTQ	0.66	0.52	0.2				
18	MC	0.42	0.31	0.2				
19	HTQ	0.61	0.60	0.1				
20	MC	0.75	0.52	0.1				
21	MS	0.25	0.30	0.1	X			
22	MC	0.38	0.25	0.1				
23	EBSR	0.49	0.39	0.1				
24	HTQ	0.16	0.32	0.4	X			
25	MS	0.21	0.27	0.2	X			

Note. MC=Multiple-choice single-select; MS=Multiple-choice multi-select; HTQ=Hot text QTI version; EBSR=Evidence-based selected-response; SA=Short answer; MI=Matching item; TI=Fill-in table item; EQ=Equation editor.

Table A.2: Operational Item Statistics, ELA Grade 4 (Cont'd)

Item Number	Item Statistic Fields				Flag			
	Item Type	*p*-value	Item-Total Correlation	Omit Rate	*p*-value	Item-Total Correlation	Distractor Correlation	Omit Rate
26	HTQ	0.21	0.48	0.2	X			
27	MC	0.45	0.15	0.1		X		
28	MC	0.41	0.19	0.2		X		
29	MS	0.14	0.35	0.2	X			
30	MC	0.33	0.32	0.4				
31	HTQ	0.60	0.61	0.3				
32	MS	0.14	0.26	0.3	X			
33	MC	0.51	0.30	0.3				
34	MS	0.16	0.30	0.3	X			
35	SA	0.20	0.42	0.5	X			
36	MS	0.15	0.30	0.2	X			
37	MC	0.45	0.53	0.3				
38	MS	0.14	0.36	0.4	X			
39	EBSR	0.26	0.45	0.2	X			
40	MC	0.58	0.45	0.3				
41	MC	0.44	0.46	0.5				
42	EBSR	0.34	0.39	0.2				
43	MC	0.65	0.52	0.3				
44	MC	0.80	0.53	0.4				
45	EBSR	0.56	0.59	0.3				
46	MC	0.45	0.40	0.2				
47	MS	0.19	0.39	0.2	X			

Note. MC=Multiple-choice single-select; MS=Multiple-choice multi-select; HTQ=Hot text QTI version; EBSR=Evidence-based selected-response; SA=Short answer; MI=Matching item; TI=Fill-in table item; EQ=Equation editor.

Table A.3: Operational Item Statistics, ELA Grade 5

Item Number	Item Statistic Fields				Flag			
	Item Type	*p*-value	Item-Total Correlation	Omit Rate	*p*-value	Item-Total Correlation	Distractor Correlation	Omit Rate
1	MC	0.70	0.35	0.1				
2	MC	0.49	0.41	0.1				
3	HTQ	0.41	0.37	0.1				
4	HTQ	0.26	0.39	0.0	X			
5	MS	0.31	0.31	0.1				
6	EBSR	0.38	0.49	0.1				
7	HTQ	0.37	0.49	0.1				
8	MC	0.63	0.43	0.1				
9	MC	0.87	0.45	0.1				
10	MC	0.50	0.46	0.1				
11	MC	0.71	0.36	0.1				
12	MC	0.75	0.59	0.1				
13	MC	0.58	0.44	0.1				
14	MC	0.67	0.49	0.1				
15	MS	0.39	0.46	0.1				
16	HTQ	0.65	0.52	0.2				
17	MS	0.36	0.49	0.1				
18	EBSR	0.25	0.42	0.1	X			
19	EBSR	0.48	0.54	0.1				
20	MC	0.71	0.50	0.2				
21	MC	0.53	0.31	0.1				
22	MS	0.20	0.42	0.1	X			
23	MC	0.68	0.46	0.1				
24	MC	0.44	0.31	0.1				
25	HTQ	0.41	0.60	0.2				

Note. MC=Multiple-choice single-select; MS=Multiple-choice multi-select; HTQ=Hot text QTI version; EBSR=Evidence-based selected-response; SA=Short answer; MI=Matching item; TI=Fill-in table item; EQ=Equation editor.

Table A.3: Operational Item Statistics, ELA Grade 5 (Cont'd)

Item Number	Item Statistic Fields				Flag			
	Item Type	*p*-value	Item-Total Correlation	Omit Rate	*p*-value	Item-Total Correlation	Distractor Correlation	Omit Rate
26	HTQ	0.44	0.50	0.2				
27	SA	0.49	0.51	0.3				
28	MC	0.80	0.50	0.2				
29	MC	0.86	0.51	0.2				
30	HTQ	0.32	0.41	0.2				
31	SA	0.37	0.48	0.4				
32	HTQ	0.41	0.49	0.2				
33	MC	0.69	0.49	0.3				
34	MS	0.28	0.47	0.2	X			
35	MC	0.82	0.50	0.2				
36	MS	0.15	0.37	0.3	X			
37	EBSR	0.27	0.48	0.2	X			
38	MC	0.53	0.29	0.3				
39	MS	0.11	0.24	0.3	X			
40	EBSR	0.20	0.40	0.2	X			
41	MC	0.61	0.52	0.3				
42	MS	0.40	0.42	0.4				
43	EBSR	0.31	0.26	0.2				
44	MC	0.54	0.37	0.2				
45	MS	0.40	0.60	0.2				

Note. MC=Multiple-choice single-select; MS=Multiple-choice multi-select; HTQ=Hot text QTI version; EBSR=Evidence-based selected-response; SA=Short answer; MI=Matching item; TI=Fill-in table item; EQ=Equation editor.

Table A.4: Operational Item Statistics, ELA Grade 6

Item Number	Item Statistic Fields				Flag			
	Item Type	*p*-value	Item-Total Correlation	Omit Rate	*p*-value	Item-Total Correlation	Distractor Correlation	Omit Rate
1	MC	0.49	0.30	0.1				
2	HTQ	0.37	0.44	0.1				
3	MC	0.41	0.16	0.1		X		
4	HTQ	0.16	0.40	0.1	X			
5	MC	0.32	0.17	0.1		X		
6	MC	0.60	0.51	0.1				
7	HTQ	0.12	0.31	0.1	X			
8	MC	0.43	0.25	0.1				
9	MC	0.69	0.40	0.1				
10	HTQ	0.29	0.26	0.1	X			
11	EBSR	0.39	0.37	0.1				
12	MC	0.50	0.21	0.1				
13	SA	0.28	0.48	0.8	X			
14	MC	0.26	0.13	0.2	X	X		
15	MS	0.27	0.35	0.1	X			
16	SA	0.29	0.47	1.0	X			
17	HTQ	0.47	0.41	0.3				
18	MS	0.20	0.36	0.2	X			
19	MC	0.47	0.46	0.2				
20	MS	0.09	0.26	0.2	X			
21	MC	0.69	0.41	0.2				
22	EBSR	0.26	0.54	0.3	X			
23	MC	0.54	0.36	0.2				
24	MC	0.65	0.46	0.2				
25	MC	0.38	0.29	0.2				

Note. MC=Multiple-choice single-select; MS=Multiple-choice multi-select; HTQ=Hot text QTI version; EBSR=Evidence-based selected-response; SA=Short answer; MI=Matching item; TI=Fill-in table item; EQ=Equation editor.

Table A.4: Operational Item Statistics, ELA Grade 6 (Cont'd)

Item Number	Item Statistic Fields				Flag			
	Item Type	*p*-value	Item-Total Correlation	Omit Rate	*p*-value	Item-Total Correlation	Distractor Correlation	Omit Rate
26	MS	0.50	0.55	0.2				
27	MC	0.47	0.28	0.3				
28	MS	0.45	0.49	0.3				
29	HTQ	0.43	0.28	0.3				
30	MC	0.79	0.56	0.3				
31	EBSR	0.28	0.43	0.3	X			
32	HTQ	0.54	0.56	0.3				
33	MC	0.48	0.48	0.3				
34	MS	0.09	0.30	0.3	X			
35	MC	0.57	0.44	0.4				
36	MS	0.23	0.44	0.4	X			
37	MC	0.44	0.21	0.4				
38	MC	0.51	0.53	0.3				
39	MC	0.46	0.24	0.5				
40	EBSR	0.34	0.49	0.3				
41	MC	0.28	0.28	0.4	X			
42	MS	0.15	0.20	0.5	X	X		
43	EBSR	0.38	0.43	0.3				
44	MC	0.69	0.50	0.4				
45	MI	0.58	0.32	0.3				
46	EBSR	0.17	0.26	0.3	X			
47	MC	0.60	0.44	0.3				
48	MC	0.37	0.21	0.3				

Note. MC=Multiple-choice single-select; MS=Multiple-choice multi-select; HTQ=Hot text QTI version; EBSR=Evidence-based selected-response; SA=Short answer; MI=Matching item; TI=Fill-in table item; EQ=Equation editor.

Table A.5: Operational Item Statistics, ELA Grade 7

Item Number	Item Statistic Fields				Flag			
	Item Type	*p*-value	Item-Total Correlation	Omit Rate	*p*-value	Item-Total Correlation	Distractor Correlation	Omit Rate
1	MC	0.88	0.36	0.1				
2	MC	0.67	0.56	0.1				
3	HTQ	0.58	0.43	0.1				
4	MC	0.61	0.29	0.1				
5	SA	0.38	0.49	0.7				
6	EBSR	0.54	0.47	0.1				
7	MS	0.41	0.37	0.2				
8	HTQ	0.58	0.50	0.2				
9	MC	0.51	0.25	0.2				
10	MC	0.53	0.20	0.3				
11	MC	0.77	0.35	0.2				
12	SA	0.32	0.46	1.1				
13	MS	0.42	0.30	0.1				
14	HTQ	0.14	0.23	0.2	X			
15	HTQ	0.47	0.43	0.2				
16	MC	0.61	0.45	0.2				
17	MS	0.24	0.49	0.2	X			
18	MC	0.48	0.37	0.2				
19	HTQ	0.78	0.53	0.3				
20	MC	0.92	0.32	0.2				
21	SA	0.26	0.58	1.4	X			
22	MC	0.45	0.35	0.2				
23	MC	0.42	0.24	0.3				
24	MC	0.70	0.35	0.4				
25	HTQ	0.33	0.27	0.5				

Note. MC=Multiple-choice single-select; MS=Multiple-choice multi-select; HTQ=Hot text QTI version; EBSR=Evidence-based selected-response; SA=Short answer; MI=Matching item; TI=Fill-in table item; EQ=Equation editor.

Table A.5: Operational Item Statistics, ELA Grade 7 (Cont'd)

Item Number	Item Statistic Fields				Flag			
	Item Type	*p*-value	Item-Total Correlation	Omit Rate	*p*-value	Item-Total Correlation	Distractor Correlation	Omit Rate
26	MC	0.69	0.43	0.2				
27	MC	0.43	0.40	0.3				
28	MS	0.18	0.36	0.6	X			
29	MS	0.23	0.49	0.8	X			
30	MS	0.20	0.34	0.3	X			
31	HTQ	0.14	0.23	0.4	X			
32	EBSR	0.27	0.43	0.4	X			
33	MS	0.12	0.20	0.4	X	X		
34	MC	0.54	0.36	0.4				
35	HTQ	0.43	0.41	0.5				
36	HTQ	0.60	0.38	0.4				
37	MC	0.57	0.32	0.4				
38	MS	0.21	0.39	0.5	X			
39	EBSR	0.31	0.44	0.4				
40	MC	0.60	0.38	0.4				
41	MC	0.46	0.27	0.5				
42	EBSR	0.55	0.45	0.3				
43	MC	0.47	0.29	0.4				
44	MC	0.55	0.23	0.5				
45	EBSR	0.53	0.56	0.3				
46	MC	0.76	0.52	0.3				
47	MS	0.32	0.38	0.3				

Note. MC=Multiple-choice single-select; MS=Multiple-choice multi-select; HTQ=Hot text QTI version; EBSR=Evidence-based selected-response; SA=Short answer; MI=Matching item; TI=Fill-in table item; EQ=Equation editor.

Table A.6: Operational Item Statistics, ELA Grade 8

Item Number	Item Statistic Fields				Flag			
	Item Type	*p*-value	Item-Total Correlation	Omit Rate	*p*-value	Item-Total Correlation	Distractor	Omit
1	MC	0.63	0.13	0.1		X		
2	MC	0.74	0.39	0.1				
3	MC	0.66	0.34	0.1				
4	HTQ	0.35	0.37	0.1				
5	SA	0.38	0.53	1.1				
6	HTQ	0.31	0.27	0.1				
7	MC	0.73	0.53	0.1				
8	MC	0.87	0.43	0.1				
9	MS	0.49	0.36	0.1				
10	EBSR	0.36	0.51	0.1				
11	SA	0.36	0.69	1.4				
12	MC	0.62	0.44	0.2				
13	MC	0.48	0.29	0.3				
14	MC	0.44	0.27	0.2				
15	MC	0.60	0.34	0.2				
16	MC	0.64	0.55	0.2				
17	MS	0.24	0.36	0.7	X			
18	MC	0.32	0.27	0.3				
19	EBSR	0.31	0.35	0.3				
20	HTQ	0.39	0.44	0.3				
21	HTQ	0.35	0.41	0.3				
22	MS	0.20	0.41	0.3	X			
23	EBSR	0.44	0.57	0.7				
24	MC	0.61	0.47	0.4				
25	MS	0.27	0.37	0.2	X			

Note. MC=Multiple-choice single-select; MS=Multiple-choice multi-select; HTQ=Hot text QTI version; EBSR=Evidence-based selected-response; SA=Short answer; MI=Matching item; TI=Fill-in table item; EQ=Equation editor.

Table A.6: Operational Item Statistics, ELA Grade 8 (Cont'd)

Item Number	Item Statistic Fields				Flag			
	Item Type	*p*-value	Item-Total Correlation	Omit Rate	*p*-value	Item-Total Correlation	Distractor	Omit
26	SA	0.41	0.52	2.0				
27	MS	0.12	0.25	0.2	X			
28	HTQ	0.34	0.39	0.2				
29	MC	0.33	0.21	0.4				
30	MS	0.31	0.30	0.3				
31	EBSR	0.13	0.12	0.3	X	X		
32	EBSR	0.35	0.56	0.3				
33	MC	0.51	0.40	0.3				
34	MS	0.25	0.39	0.4	X			
35	MC	0.42	0.33	0.5				
36	MC	0.66	0.35	0.4				
37	MC	0.43	0.31	0.4				
38	MC	0.53	0.34	0.5				
39	MC	0.70	0.32	0.4				
40	MC	0.33	0.24	0.5				
41	MC	0.47	0.33	0.5				
42	EBSR	0.53	0.41	0.3				
43	MC	0.61	0.32	0.4				
44	MS	0.22	0.22	0.4	X			
45	MC	0.39	0.18	0.3		X		
46	MS	0.12	0.13	0.2	X	X		
47	MS	0.17	0.36	0.2	X			

Note. MC=Multiple-choice single-select; MS=Multiple-choice multi-select; HTQ=Hot text QTI version; EBSR=Evidence-based selected-response; SA=Short answer; MI=Matching item; TI=Fill-in table item; EQ=Equation editor.

Table A.7: Operational Item Statistics, Mathematics Grade 3

Item Number	Item Statistic Fields					Flag			
	Test Type	Item Type	*p*-value	Item-Total Correlation	Omit Rate	*p*-value	Item-Total Correlation	Distractor	Omit
1	Section 1	EQ	0.94	0.32	0.0				
2	Section 1	MC	0.87	0.37	0.0				
3	Section 1	MC	0.79	0.39	0.0				
4	Section 1	EQ	0.79	0.50	0.0				
5	Section 1	EQ	0.74	0.55	0.1				
6	Section 1	EQ	0.70	0.54	0.1				
7	Section 1	EQ	0.69	0.56	0.1				
8	Section 1	EQ	0.52	0.47	0.1				
9	Section 1	EQ	0.38	0.60	0.1				
10	Section 1	EQ	0.21	0.53	0.2	X			
11	Section 1	MC	0.73	0.49	0.0				
12	Section 1	MC	0.43	0.42	0.1				
13	Section 1	EQ	0.11	0.46	0.1	X			
14	Section 1	EQ	0.37	0.55	0.1				
15	Section 1	MC	0.63	0.40	0.1				
16	Section 1	MI	0.55	0.56	0.0				
17	Section 1	MI	0.43	0.52	0.0				
18	Section 1	MS	0.31	0.45	0.1				
19	Section 1	MS	0.18	0.49	0.1	X			
20	Section 1	MC	0.39	0.44	0.2				
21	Section 1	MC	0.49	0.47	0.1				
22	Section 1	EQ	0.07	0.36	0.1	X			
23	Section 1	EQ	0.26	0.49	0.1	X			
24	Section 1	EQ	0.31	0.59	0.2				
25	Section 1	EQ	0.60	0.54	0.1				

Note. MC=Multiple-choice single-select; MS=Multiple-choice multi-select; HTQ=Hot text QTI version; EBSR=Evidence-based selected-response; SA=Short answer; MI=Matching item; TI=Fill-in table item; EQ=Equation editor; PT=Performance task.

Table A.7: Operational Item Statistics, Mathematics Grade 3 (Cont'd)

Item Number	Item Statistic Fields					Flag			
	Test Type	Item Type	*p*-value	Item-Total Correlation	Omit Rate	*p*-value	Item-Total Correlation	Distractor	Omit
26	Section 1	EQ	0.67	0.59	0.1				
27	Section 1	EQ	0.54	0.56	0.1				
28	Section 1	EQ	0.75	0.48	0.1				
29	Section 1	EQ	0.70	0.49	0.1				
30	Section 1	EQ	0.79	0.54	0.1				
31	Section 1	MC	0.90	0.33	0.1				
1	PT1	EQ	0.66	0.38	0.1				
2	PT1	EQ	0.54	0.54	0.0				
3	PT1	EQ	0.44	0.54	0.1				
4	PT1	SA	0.29	0.50	0.1	X			
5	PT1	MC	0.43	0.07	0.1		X		
1	PT2	EQ	0.56	0.60	0.1				
2	PT2	EQ	0.56	0.59	0.1				
3	PT2	SA	1.00[a]		0.2				
4	PT2	TI	0.67	0.61	0.4				
5	PT2	SA	0.54	0.58	0.1				
6	PT2	SA	0.25	0.60	0.3	X			

Note. MC=Multiple-choice single-select; MS=Multiple-choice multi-select; HTQ=Hot text QTI version; EBSR=Evidence-based selected-response; SA=Short answer; MI=Matching item; TI=Fill-in table item; EQ=Equation editor; PT=Performance task.
[a] All responses that were not blank received full credit due to a Spanish translation issue.

Table A.8: Operational Item Statistics, Mathematics Grade 4

Item Number	Item Statistic Fields					Flag			
	Test Type	Item Type	*p*-value	Item-Total Correlation	Omit Rate	*p*-value	Item-Total Correlation	Distractor	Omit
1	Section 1	MI	0.88	0.27	0.0				
2	Section 1	EQ	0.90	0.27	0.0				
3	Section 1	EQ	0.86	0.57	0.0				
4	Section 1	EQ	0.80	0.47	0.0				
5	Section 1	EQ	0.39	0.33	0.1				
6	Section 1	EQ	0.55	0.68	0.0				
7	Section 1	EQ	0.51	0.65	0.0				
8	Section 1	MI	0.40	0.23	0.0				
9	Section 1	EQ	0.38	0.60	0.1				
10	Section 1	EQ	0.24	0.51	0.1	X			
11	Section 1	MS	0.32	0.53	0.0				
12	Section 1	EQ	0.17	0.33	0.1	X			
13	Section 1	EQ	0.10	0.44	0.1	X			
14	Section 1	MS	0.19	0.39	0.1	X			
15	Section 1	EQ	0.28	0.44	0.1	X			
16	Section 1	MC	0.63	0.39	0.0				
17	Section 1	MI	0.37	0.38	0.1				
18	Section 1	MS	0.36	0.50	0.1				
19	Section 1	MS	0.28	0.40	0.1	X			
20	Section 1	MS	0.38	0.54	0.0				
21	Section 1	MS	0.32	0.51	0.1				
22	Section 1	MI	0.11	0.47	0.1	X			
23	Section 1	MI	0.22	0.21	0.1	X			
24	Section 1	MI	0.29	0.58	0.1	X			
25	Section 1	EQ	0.56	0.55	0.1				

Note. MC=Multiple-choice single-select; MS=Multiple-choice multi-select; HTQ=Hot text QTI version; EBSR=Evidence-based selected-response; SA=Short answer; MI=Matching item; TI=Fill-in table item; EQ=Equation editor; PT=Performance task.

Table A.8: Operational Item Statistics, Mathematics Grade 4 (Cont'd)

Item Number	Item Statistic Fields					Flag			
	Test Type	Item Type	*p*-value	Item-Total Correlation	Omit Rate	*p*-value	Item-Total Correlation	Distractor	Omit
26	Section 1	EQ	0.57	0.63	0.1				
27	Section 1	EQ	0.58	0.64	0.1				
28	Section 1	EQ	0.54	0.56	0.1				
29	Section 1	EQ	0.78	0.34	0.1				
30	Section 1	MC	0.84	0.46	0.1				
31	Section 1	MC	0.85	0.46	0.1				
1	PT1	EQ	0.67	0.50	0.0				
2	PT1	EQ	0.17	0.47	0.0	X			
3	PT1	SA	0.30	0.48	0.0				
4	PT1	SA	0.23	0.47	0.1	X			
5	PT1	SA	0.33	0.59	0.2				
1	PT2	EQ	0.54	0.48	0.0				
2	PT2	MC	0.42	0.44	0.0				
3	PT2	SA	0.34	0.48	0.1				
4	PT2	SA	0.33	0.60	0.1				
5	PT2	SA	0.49	0.55	0.1				
6	PT2	SA	0.25	0.55	0.2	X			

Note. MC=Multiple-choice single-select; MS=Multiple-choice multi-select; HTQ=Hot text QTI version; EBSR=Evidence-based selected-response; SA=Short answer; MI=Matching item; TI=Fill-in table item; EQ=Equation editor; PT=Performance task.

Table A.9: Operational Item Statistics, Mathematics Grade 5

Item Number	Item Statistic Fields					Flag			
	Test Type	Item Type	*p*-value	Item-Total Correlation	Omit Rate	*p*-value	Item-Total Correlation	Distractor	Omit
1	Section 1	EQ	0.84	0.42	0.0				
2	Section 1	MC	0.62	0.45	0.0				
3	Section 1	MC	0.54	0.43	0.0				
4	Section 1	EQ	0.50	0.66	0.1				
5	Section 1	MI	0.45	0.42	0.0				
6	Section 1	EQ	0.42	0.56	0.1				
7	Section 1	EQ	0.29	0.61	0.2	X			
8	Section 1	MI	0.26	0.37	0.0	X			
9	Section 1	EQ	0.12	0.47	0.2	X			
10	Section 1	EQ	0.08	0.41	0.5	X			
11	Section 1	EQ	0.69	0.62	0.1				
12	Section 1	MC	0.50	0.44	0.1				
13	Section 1	EQ	0.37	0.57	0.1				
14	Section 1	EQ	0.23	0.56	0.2	X			
15	Section 1	MC	0.57	0.53	0.1				
16	Section 1	MC	0.55	0.23	0.1				
17	Section 1	EQ	0.48	0.61	0.1				
18	Section 1	EQ	0.27	0.56	0.2	X			
19	Section 1	MC	0.41	0.41	0.1				
20	Section 1	MC	0.40	0.22	0.1				
21	Section 1	MC	0.67	0.55	0.1				
22	Section 1	EQ	0.07	0.34	0.3	X			
23	Section 1	EQ	0.10	0.44	0.5	X			
24	Section 1	EQ	0.11	0.47	0.2	X			
25	Section 1	EQ	0.25	0.43	0.2	X			

Note. MC=Multiple-choice single-select; MS=Multiple-choice multi-select; HTQ=Hot text QTI version; EBSR=Evidence-based selected-response; SA=Short answer; MI=Matching item; TI=Fill-in table item; EQ=Equation editor; PT=Performance task.

Table A.9: Operational Item Statistics, Mathematics Grade 5 (Cont'd)

Item Number	Item Statistic Fields					Flag			
	Test Type	Item Type	*p*-value	Item-Total Correlation	Omit Rate	*p*-value	Item-Total Correlation	Distractor	Omit
26	Section 1	EQ	0.24	0.43	0.3	X			
27	Section 1	MC	0.44	0.38	0.2				
28	Section 1	EQ	0.51	0.68	0.1				
29	Section 1	EQ	0.50	0.53	0.2				
30	Section 1	MC	0.63	0.31	0.1				
31	Section 1	MC	0.80	0.41	0.1				
1	PT1	EQ	0.23	0.53	0.1	X			
2	PT1	EQ	0.25	0.54	0.1	X			
3	PT1	TI	0.28	0.53	0.3	X			
4	PT1	SA	0.13	0.48	0.3	X			
5	PT1	TI	0.23	0.50	0.6	X			
6	PT1	SA	0.37	0.53	0.4				
1	PT2	EQ	0.36	0.56	0.1				
2	PT2	EQ	0.19	0.31	0.0	X			
3	PT2	SA	0.33	0.57	0.1				
4	PT2	SA	0.53	0.60	0.1				
5	PT2	SA	0.22	0.55	0.2	X			
6	PT2	SA	0.08	0.45	0.4	X			

Note. MC=Multiple-choice single-select; MS=Multiple-choice multi-select; HTQ=Hot text QTI version; EBSR=Evidence-based selected-response; SA=Short answer; MI=Matching item; TI=Fill-in table item; EQ=Equation editor; PT=Performance task.

Table A.10: Operational Item Statistics, Mathematics Grade 6

Item Number	Item Statistic Fields					Flag			
	Test Type	Item Type	*p*-value	Item-Total Correlation	Omit Rate	*p*-value	Item-Total Correlation	Distractor	Omit
1	Section 1	EQ	0.87	0.40	0.1				
2	Section 1	EQ	0.75	0.57	0.1				
3	Section 1	EQ	0.59	0.61	0.2				
4	Section 1	MS	0.52	0.50	0.1				
5	Section 1	MI	0.50	0.41	0.1				
6	Section 1	MS	0.46	0.48	0.1				
7	Section 1	MS	0.39	0.55	0.1				
8	Section 1	EQ	0.15	0.52	0.2	X			
9	Section 1	EQ	0.22	0.52	0.3	X			
10	Section 1	EQ	0.31	0.53	0.3				
11	Section 1	MS	0.49	0.63	0.1				
12	Section 1	MC	0.52	0.43	0.1				
13	Section 1	EQ	0.53	0.54	0.5				
14	Section 1	EQ	0.63	0.65	0.2				
15	Section 1	EQ	0.81	0.53	0.2				
16	Section 1	EQ	0.27	0.55	0.1	X			
17	Section 1	MI	0.24	0.47	0.2	X			
18	Section 1	EQ	0.07	0.37	0.2	X			
19	Section 1	MS	0.14	0.30	0.2	X			
20	Section 1	EQ	0.28	0.58	0.4	X			
21	Section 1	EQ	0.35	0.49	0.2				
22	Section 1	EQ	0.25	0.58	0.2	X			
23	Section 1	MS	0.13	0.36	0.1	X			
24	Section 1	MI	0.09	0.37	0.2	X			
25	Section 1	MS	0.23	0.36	0.1	X			

Note. MC=Multiple-choice single-select; MS=Multiple-choice multi-select; HTQ=Hot text QTI version; EBSR=Evidence-based selected-response; SA=Short answer; MI=Matching item; TI=Fill-in table item; EQ=Equation editor; PT=Performance task.

Table A.10: Operational Item Statistics, Mathematics Grade 6 (Cont'd)

Item Number	Item Statistic Fields					Flag			
	Test Type	Item Type	*p*-value	Item-Total Correlation	Omit Rate	*p*-value	Item-Total Correlation	Distractor	Omit
26	Section 1	MC	0.33	0.24	0.2				
27	Section 1	EQ	0.21	0.50	0.3	X			
28	Section 1	EQ	0.26	0.59	0.3	X			
29	Section 1	MC	0.45	0.35	0.2				
30	Section 1	MC	0.59	0.45	0.2				
31	Section 1	MC	0.77	0.20	0.2				
1	PT1	SA	0.26	0.61	0.2	X			
2	PT1	SA	0.27	0.60	0.3	X			
3	PT1	EQ	0.22	0.51	0.2	X			
4	PT1	SA	0.13	0.49	0.3	X			
1	PT2	EQ	0.22	0.45	0.2	X			
2	PT2	SA	0.41	0.60	0.2				
3	PT2	TI	0.29	0.52	0.5	X			
4	PT2	SA	0.62	0.54	0.2				
5	PT2	SA	0.17	0.50	0.7	X			

Note. MC=Multiple-choice single-select; MS=Multiple-choice multi-select; HTQ=Hot text QTI version; EBSR=Evidence-based selected-response; SA=Short answer; MI=Matching item; TI=Fill-in table item; EQ=Equation editor; PT=Performance task.

Table A.11: Operational Item Statistics, Mathematics Grade 7

Item Number	Item Statistic Fields					Flag			
	Test Type	Item Type	*p*-value	Item-Total Correlation	Omit Rate	*p*-value	Item-Total Correlation	Distractor	Omit
1	Section 1	MC	0.68	0.35	0.1				
2	Section 1	EQ	0.72	0.56	0.4				
3	Section 1	MI	0.34	0.46	0.1				
4	Section 1	EQ	0.35	0.62	0.4				
5	Section 1	EQ	0.08	0.41	0.7	X			
6	Section 1	EQ	0.11	0.42	1.0	X			
7	Section 1	EQ	0.28	0.48	0.5	X			
8	Section 1	EQ	0.52	0.68	0.5				
9	Section 1	EQ	0.57	0.68	0.3				
10	Section 1	MC	0.85	0.41	0.1				
11	Section 1	MI	0.75	0.47	0.0				
12	Section 1	MS	0.40	0.43	0.1				
13	Section 1	MS	0.34	0.36	0.1				
14	Section 1	EQ	0.20	0.52	0.3	X			
15	Section 1	MI	0.12	0.22	0.1	X			
16	Section 1	MC	0.48	0.35	0.1				
17	Section 1	EQ	0.16	0.45	0.3	X			
18	Section 1	EQ	0.05	0.35	0.6	X			
19	Section 1	EQ	0.05	0.28	0.4	X			
20	Section 1	EQ	0.19	0.44	0.5	X			
21	Section 1	SA	0.19	0.47	0.6	X			
22	Section 1	MS	0.27	0.53	0.4	X			
23	Section 1	MS	0.04	0.25	0.3	X			
24	Section 1	EQ	0.21	0.53	0.4	X			
25	Section 1	MS	0.34	0.55	0.4				

Note. MC=Multiple-choice single-select; MS=Multiple-choice multi-select; HTQ=Hot text QTI version; EBSR=Evidence-based selected-response; SA=Short answer; MI=Matching item; TI=Fill-in table item; EQ=Equation editor; PT=Performance task.

Table A.11: Operational Item Statistics, Mathematics Grade 7 (Cont'd)

Item Number	Item Statistic Fields					Flag			
	Test Type	Item Type	*p*-value	Item-Total Correlation	Omit Rate	*p*-value	Item-Total Correlation	Distractor	Omit
26	Section 1	EQ	0.13	0.44	0.6	X			
27	Section 1	MS	0.21	0.49	0.3	X			
28	Section 1	EQ	0.31	0.66	0.4				
29	Section 1	EQ	0.34	0.67	0.4				
30	Section 1	MI	0.44	0.28	0.2				
31	Section 1	MC	0.78	0.50	0.2				
1	PT1	EQ	0.32	0.61	0.1				
2	PT1	EQ	0.25	0.55	0.2	X			
3	PT1	SA	0.07	0.40	1.2	X			
4	PT1	SA	0.37	0.64	0.8				
5	PT1	MC	0.33	0.18	0.3		X		
1	PT2	EQ	0.60	0.60	0.2				
2	PT2	EQ	0.34	0.56	0.1				
3	PT2	SA	0.12	0.40	0.4	X			
4	PT2	EQ	0.23	0.48	0.3	X			

Note. MC=Multiple-choice single-select; MS=Multiple-choice multi-select; HTQ=Hot text QTI version; EBSR=Evidence-based selected-response; SA=Short answer; MI=Matching item; TI=Fill-in table item; EQ=Equation editor; PT=Performance task.

Table A.12: Operational Item Statistics, Mathematics Grade 8

Item Number	Item Statistic Fields					Flag			
	Test Type	Item Type	*p*-value	Item-Total Correlation	Omit Rate	*p*-value	Item-Total Correlation	Distractor	Omit
1	Section 1	MC	0.73	0.30	0.1				
2	Section 1	MC	0.49	0.37	0.1				
3	Section 1	MC	0.46	0.42	0.1				
4	Section 1	MI	0.20	0.20	0.1	X			
5	Section 1	EQ	0.18	0.50	0.8	X			
6	Section 1	MC	0.24	0.39	0.1	X			
7	Section 1	MC	0.46	0.38	0.1				
8	Section 1	EQ	0.50	0.56	0.5				
9	Section 1	MC	0.47	0.40	0.1				
10	Section 1	MI	0.48	0.53	0.1				
11	Section 1	MC	0.34	0.12	0.1		X	X	
12	Section 1	MI	0.41	0.48	0.0				
13	Section 1	EQ	0.21	0.50	0.3	X			
14	Section 1	EQ	0.20	0.45	0.7	X			
15	Section 1	MC	0.62	0.56	0.1				
16	Section 1	EQ	0.46	0.61	0.2				
17	Section 1	MS	0.04	0.22	0.3	X			
18	Section 1	EQ	0.15	0.48	0.6	X			
19	Section 1	MS	0.56	0.58	0.5				
20	Section 1	EQ	0.07	0.40	0.7	X			
21	Section 1	MC	0.32	0.26	0.3				
22	Section 1	MC	0.29	0.16	0.1	X	X		
23	Section 1	MS	0.23	0.39	0.1	X			
24	Section 1	MS	0.09	0.41	0.2	X			
25	Section 1	MS	0.25	0.39	0.1	X			

Note. MC=Multiple-choice single-select; MS=Multiple-choice multi-select; HTQ=Hot text QTI version; EBSR=Evidence-based selected-response; SA=Short answer; MI=Matching item; TI=Fill-in table item; EQ=Equation editor; PT=Performance task.

Table A.12: Operational Item Statistics, Mathematics Grade 8 (Cont'd)

Item Number	Item Statistic Fields					Flag			
	Test Type	Item Type	*p*-value	Item-Total Correlation	Omit Rate	*p*-value	Item-Total Correlation	Distractor	Omit
26	Section 1	MC	0.35	0.33	0.1				
27	Section 1	EQ	0.16	0.33	0.5	X			
28	Section 1	MS	0.23	0.48	0.2	X			
29	Section 1	MS	0.27	0.55	0.1	X			
30	Section 1	EQ	0.30	0.62	0.6				
31	Section 1	EQ	0.41	0.49	1.0				
32	Section 1	EQ	0.35	0.62	0.7				
1	PT1	EQ	0.16	0.49	0.4	X			
2	PT1	SA	0.16	0.62	0.6	X			
3	PT1	SA	0.33	0.49	0.6				
1	PT2	EQ	0.33	0.42	0.3				
2	PT2	EQ	0.20	0.44	0.2	X			
3	PT2	SA	0.14	0.41	2.0	X			
4	PT2	EQ	0.31	0.56	0.5				

Note. MC=Multiple-choice single-select; MS=Multiple-choice multi-select; HTQ=Hot text QTI version; EBSR=Evidence-based selected-response; SA=Short answer; MI=Matching item; TI=Fill-in table item; EQ=Equation editor; PT=Performance task.